TOTh 2010, Terminology & Ontology: Theories and applications

Christophe Roche

Terminologie & Ontologie : Théories et Applications

Actes de la conférence

TOTh 2010

Annecy – 3 & 4 juin 2010

Publications précédentes

TOTh 2007

Actes de la première conférence TOTh - Annecy - 1er juin 2007

TOTh 2008

Actes de la deuxième conférence TOTh - Annecy – 5 et 6 juin 2008

TOTh 2009

Actes de la troisième conférence TOTh - Annecy – 4 et 5 juin 2009

Commandes à adresser à : toth@porphyre.org

Titre : TOTh 2010. *Actes de la quatrième conférence TOTh - Annecy – 3 & 4 juin 2010*
Editeur : Institut Porphyre, *Savoir et Connaissance*
http://www.porphyre.org
Annecy, 2010
ISBN 978-2-9536168-1-1
EAN 9782953616811

Terminologie & Ontologie : Théories et applications

Actes de la conférence

TOTh 2010

Annecy – 3 & 4 juin 2010

avec le soutien de :

- Ministère de la Culture et de la Communication, Délégation Générale à la Langue Française et aux Langues de France
- Association Européenne de Terminologie
- Société française de terminologie
- Ecole d'ingénieurs Polytech'Savoie – Université de Savoie
- Université de Sorbonne nouvelle
- Association EGC (Extraction et Gestion des Connaissances)
- ISKO (International Society for Knowledge Organization) France

Institut Porphyre
Savoir et Connaissance

`http://www.porphyre.org`

Comité scientifique

Président du Comité Scientifique : Christophe Roche

Comité de pilotage

Loïc Depecker	Professeur, Université de Sorbonne nouvelle
André Manificat	Directeur, GRETh
Christophe Roche	Professeur, Université de Savoie
Philippe Thoiron	Professeur émérite, Université de Lyon II

Comité de programme

Bruno de Bessé	Professeur, Université de Genève
Franco Bertaccini	Professeur, Université de Bologne
Gerhard Budin	Professeur, Université de Vienne
Marc van Campenhoudt	Professeur, Termisti, ISTI, Bruxelles
Danielle Candel	CNRS, Université Paris Diderot
Stéphane Chaudiron	Professeur, Université de Lille 3
Rute Costa	Professeur, Universidade Nova de Lisboa
Luc Damas	MCF, Université de Savoie
Sylvie Després	Professeur, Université Paris 13
François Gaudin	Professeur, Université de Rouen
Anne-Marie Gendron	Chancellerie fédérale suisse, Section terminologie
Jean-Yves Gresser	Ancien Directeur à la Banque de France
Ollivier Haemmerlé	Professeur, Université de Toulouse
Michèle Hudon	Professeur, Université de Montréal
John Humbley	Professeur, Université Paris 7
Michel Ida	Directeur MINATEC, CEA
Hendrik Kockaert	Professeur, Lessius Hogeschool (Anvers)
Michel Léonard	Professeur, Université de Genève
Pierre Lerat	Professeur honoraire, Equipe Condillac
Widad Mustafa	Professeur, Université de Lille 3
Fidelma Ní Ghallchobhair	Foras na Gaeilge (The Irish-Language Body)
Henrik Nilsson	Terminologicentrum TNC, Suède
Jean Quirion	Professeur, Université d'Ottawa
Renato Reinau	Suva, Lucerne
François Rousselot	MCF, Université de Strasbourg
Gérard Sabah	CNRS, Orsay
Michel Simonet	CNRS, Grenoble
Marcus Spies	Professeur, Université de Munich
Dardo de Vecchi	Professeur associé, Euromed-Management

Comité d'organisation :

Responsable : Luc Damas

Samia Chouder, Joëlle Pellet

Avant propos

Cette année la conférence a été précédée d'une journée de formation consacrée à la terminologie et l'ontologie, à leurs liens et leurs apports mutuels. L'intérêt qu'a suscité cette journée nous amènera certainement à réitérer l'opération les années suivantes.

Le succès de la conférence d'ouverture de notre collègue Frédéric Nef, portant sur l'ontologie prise dans sa dimension philosophique, a montré, s'il en était encore besoin, la richesse d'une approche pluridisciplinaire.

Animées par différents présidents, les sessions ont alterné présentations théoriques et démonstrations de systèmes, offrant ainsi l'opportunité à plusieurs industriels de nous parler de leurs projets. L'éventail des sujets abordés, à travers les quatorze présentations retenues (incluant la conférence d'ouverture) réparties sur deux jours, illustre la richesse mais aussi la vitalité de notre communauté : aide à la traduction, thésaurus multilingue, phraséologie, entité nommé, recherche d'information, etc. L' « actualité » n'a pas été oubliée à travers une ontologie des risques financiers.

Enfin, les Conférences TOTh sont devenues internationales à partir de cette année avec le français et l'anglais comme langues officielles. Le comité de programme s'est ouvert à de nouveaux membres portant à dix le nombre de pays représentés et à plus de 40% le nombre de personnalités étrangères. Gageons que cette ouverture sera prometteuse.

Christophe Roche

Président du Comité Scientifique

Table des matières

CONFERENCE INVITEE

ARTICLES

DEMONSTRATIONS

CONFERENCE INVITEE

L'Ontologie au miroir de la Terminologie

Frederic Nef
EHESS/Institut Jean-Nicod

On peut distinguer trois grandes étapes de l'histoire de l'ontologie : la première liée à la métaphysique classique (cf. l'ONTOLOGIE A plus bas) ou '*ontologia*' qui remonte au XVII[è] siècle et irrigue le XVIII[e] siècle (Clauberg, puis Leibniz, Wolff…) jusqu'à Hume et Kant, et une seconde, datant des années 70, qui coïncide avec la renaissance des logiques modales et de l'épistémologie générale (D. Lewis, Armstrong) et enfin une troisième très récente – qui date des années 1990 – (Gruber, puis R. Poli, Guarino, B. Smith…) et que l'on peut caractériser comme un renouveau de l'ontologie classique lié à la sémantique de la logique modale et catalysé par le développement du paradigme informationnel (IA etc.), (cf. ONTOLOGIE B plus bas).

Naturellement il y des exceptions à l'énorme vide ontologique entre Kant et Armstrong-Lewis : certains philosophes ont certes maintenu la tradition ontologique entre 1800 et 1950, par exemple Nicolaï Hartmann, Edmund Husserl (au début de sa carrière), Alexius Meinong etc. et une certaine tradition d'ontologie néo-scolastique s'est maintenue (cf. par exemple J. Maritain et E. Gilson en France).

Trois hypothèses peuvent être avancées au sujet des relations entre l'ontologie A et l'ontologie B :

- **H1**. (hypothèse discontinuiste) : la différence entre A et B renvoie à une différence à l'intérieur de A entre l'ontologie comme SCIENCE DE CE QUI EST (A1) et ontologie comme SYSTEME DES CONCEPTS les plus généraux. (A2).

- **H2**.(hypothèse continuiste) : cette différence entre A1 et A2 existe aussi dans B entre ontologie formelle REALISTE (B1) et B2: ontologie formelle REPRESENTATIONNELLE

- **H3**. (hypothèse consensuelle) l'examen critique de ces définitions peut éclairer les rapports entre terminologie et ontologie

L'ETAPE A DE L'ONTOLOGIE

Le développement de l'ontologie classique va de Suarez à Kant et connaît deux moments cruciaux: a) introduction du terme 'ontologia' vers 1640, ce que l'on peut appeler le 'moment lexical' de l'ontologie classique, moment qui intéresse ceux qui veulent asseoir la terminologie sur une ontologie rationnelle, b) systématisation de l'ontologie par la *Schulphilosophie* sp. Christian Wolff vers 1720-1740

Le moment lexical

Le terme '*ontologia*' (en compétition avec le terme '*ontosophia*' qui constitue un calque de '*philosophia*') apparaît dans les années 1600-1620 (rappelons que le *Discours de la Méthode* date de 1637). Les premiers textes avec le terme '*ontologia*' sont les suivants :

- Lorhardt de 1597 à 1606 œuvres diverses dont l'*Ogdoas* Scholastica (Ogdoade *Scolastique*) où figurent des tableaux dichotomiques extrêmement développés (rééditée en 1613 sous le titre : *Theatrum Philosophicum)*

- Hojer, 1613 (*Disputatione ontologica de bono et malo*) (L'ontologie apparaît là sous une forme adjective)

- Goclenius (ou Göckel), *Lexicon Philosophicum*, 1613 (sous l'intitulé « abstraction », voir *infra*).

- Alsted 1620 *Encyclopedia.* Le fait pour le terme '*ontologia*' de figurer dans cette encyclopédie qui fit date marque une consécration du terme.

Parmi ces auteurs, Lorhadt a fait l'objet d'une redécouverte récente (P. Ohrstrohm, S. Uckelman, H. Schärfe, 2008a 2008b) ; elle lui a fixé une place assez importante dans la constitution d'une ontologie diagrammatique. Comme l'ont montré des travaux récents, Lorhardt a très probablement eu des relations intellectuelles avec Göckel quand ils ont habité la même ville, Marbourg, en Allemagne. Les tableaux de Lohardt, en quoi consiste son Ogdoade, comme on a pu le remarquer à notre époque (ibid.) opèrent beaucoup plus des divisions conceptuelles qu'ils ne proposent des définitions. Naturellement, les deux choses

sont liées : diviser le concept A en les concepts B et C contribue à sa définition, mais il ne s'agit pas à proprement parler d'une définition : on se contente d'énumérer par paires opposées les marques (*Merkmale*) ou les notes (*notae*) du concept. Par exemple dans son Ogdoade Lorhardt fait partir du concept de 'temps' une accolade qui oppose 'successif' et 'momentané'. Cela revient à diviser le temps en successif (le temps comme durée) et le temps comme non successif (le temps comme instant), mais à proprement parler il ne s'agit pas d'une définition, comme par exemple 'ordre des événements'.

Une remarque. Les travaux récents cités ici à deux reprises n'ont pas insisté sur une filiation qui me semble évidente : celle entre les représentations diagrammatiques des connaissances dans les méthodes mnémotechniques de la Renaissance. F. Yates a jadis consacré un livre entier à ces arts de la mémoire (*The Art of Memory* 1966*)* qui la représentent sous la forme d'un théâtre (cf. le titre de la seconde édition du livre de Lorhardt : *Theatrum Philosophicum*) ou de palais, ou encore de magasin[1], où les connaissances sont rangées sur des gradins ou dans des tiroirs, pour faciliter leur visualisation et donc leur mémorisation. Cela dit, il ne semble pas que l'*Ogdoade Scolastique* vise à une manipulation des images mentales de l'ontologie. La mise en œuvre est similaire, mais le but différent : l'ontologie prétend livrer la structure conceptuelle de tout ce qui existe, tandis que la mnémotechnique vise plus modestement à faciliter la restitution des articulations de cette structure, à des fins rhétoriques d'élocution ou d'écriture.

Lorhardt considérait la métaphysique comme l'étude de la structure conceptuelle du monde ; il la définit par exemple ainsi en 1597 dans son *Liber de adeptione* :

Metaphysica, quae res omnes communiter considerat, quatenus sunt ὀντα*, quatenus summa genera & principia, nullis sensibilibus hypothesibus subnixa.* [1597: 75]

(traduction : la métaphysique qui considère toutes les choses en général, dans la mesure où elles existent et dans la mesure où elles sont du genre le plus élevé et des principes qui ne sont pas des hypothèses supportées par les sens). L'abstraction, si importante dans la définition de l'*ontologia* par Göckel/Goclenius est mentionnée ici de deux manières : généralité (*summa genera*) et détachement du sensible (*nullis sensibilis hypothesibus subnixa*).

[1] Ces images du palais ou du magasin trouvent leur origine dans les *Confessions* de St Augustin, au livre X, consacré à la mémoire et au temps.

Une remarque générale à propos des tableaux d'ontologie de l'âge classique s'impose ici. On a en effet remarqué que la mise en tableaux de l'ontologie s'inscrit à l'intérieur d'un plus vaste mouvement, qui englobe, outre la mnémotechnique, que l'on vient de mentionner, la logique, avec l'efflorescence incroyable des manuels ramistes à la fin du XVIè siècle[2]. Ramus (1515-1572) a publié en 1543 une *Dialectique* qui initia un mouvement de publication de dialectiques dites ramistes dans toute l'Europe. Ramus est un philosophe et logicien (?) qui s'oppose violemment à la logique aristotélicienne, y compris dans ses versions médiévales, pour des raisons qui ne nous concernent pas ici. Les dialectiques ramistes se présentent en général comme des classements de concepts, au détriment de l'étude (syllogistique) des inférences. Leibniz prendra une position opposée à celle de Ramus et des ramistes, revalorisant la syllogistique aristotélicienne, et estimant la logique médiévale, malgré ses propres penchants nominalistes, plus proche d'un autre courant hétérodoxe, celui des lullistes, qui tentèrent de mécaniser l'art de penser et aboutirent eux aussi à privilégier les aspects diagrammatiques, mais dans une optique mécanique.

On peut résumer ce moment ainsi : L'ontologie de Lorhardt (ca 1600) à Baumagarten inclus (ca 1750) se présente souvent sous forme de tableaux, de dichotomies. On peut parler d'ontologie diagrammatique (comme on parle de logique diagrammatique avec Venn, Peirce…). Les structures conceptuelles sont introduites sous formes de relations graphiques, essentiellement d'arbres de Porphyre ou de systèmes d'accolades. Les structures ontologiques sont essentiellement des paires de concepts.

Ce moment lexical se prolonge sur une grande partie du XVIIè siècle. On peut citer : Segers, *De Ontologia Generali* 1639; Clauberg, *Ontosophia nova …,* 1660; Micraelius, *Lexicon Philosophicum*, 1662, Du Hamel, 1678… Même si des termes concurrents apparaissent, comme *ontosophia* (calqué sur *philosophia*), noologia (variante épistémique ?), l'usage du terme '*ontologia*' se répand et élimine les autres.

On peut remarquer que le terme 'ontologie' est beaucoup plus tardif que l'ontologie ! On peut parler d'ontologie dès le chapitre 2 des *Catégories* d'Aristote,

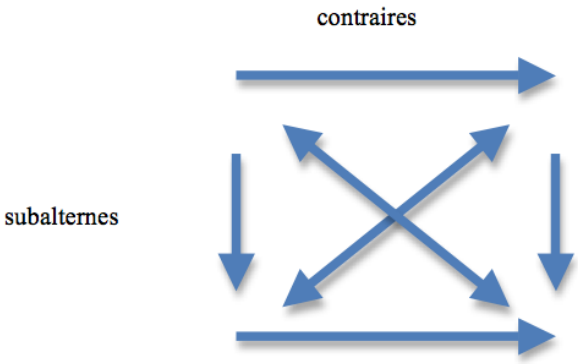
[2] André Robinet en compte 20 000 sur une soixantaine d'années ! Mais je ne sais pas quel est le sens de cette évaluation : si les dialectiques se recopiaient les unes les autres, ce chiffre ne veut rien dire, en tout les cas rien de plus que le signe d'un succès de librairie. À ce compte les 600 000 exemplaires du livre de Luc Ferry auraient un sens philosophique, alors que probablement le succès en ce cas ne dit rien sur le contenu philosophique de l'ouvrage (mais plutôt sur l'adéquation entre un produit et une demande, ce qui a pu être le pas aussi pour les dialectiques ramistes).

2000 ans avant l'émergence lexicale du terme ! Aristote utilise le terme 'philosophie première' qui est le terme le plus proche de notre 'ontologie'. Dans les *Catégories,* II, on a en effet l'ontologie suivante. Aristote distingue des relations primitives et des entités primitives. Les deux relations primitives sont: 'être dit de'/ 'être dans' (avec les négations cela donne quatre possibilités de relations). Les entités fondamentales sont les suivantes: substances premières (individus), substances secondes (genres), accidents individuels (moments), accidents universels (universaux). Exemples: un moment d'un individu se dit d'un individu et il est dans l'individu (ex. la pâleur de Socrate à un temps t). Un individu ne se dit pas d'un individu, Pierre ne peut être prédiqué de Paul[3] et il n'est pas dans un individu, il ne peut constituer en aucun cas une partie d'un individu (il peut être une partie d'un tout, par exemple d'une société, d'une famille…, mais pas d'un individu, même si une partie d'un individu peut devenir une partie d'un autre individu, par exemple dans la prédation ou la greffe).

Les structures ontologiques dans la tradition post aristotélicienne sont en général des carrés (cf. Nef 2010) qui se substituent aux arbres platoniciens (composante catégorielle plus que genres/espèces). Ces carrés dérivent de l'ontologie quadricatégorielle du chapitre II du livre des *Catégories* d'Aristote : il suffit de mettre aux quatre coins d'un carré les quatre catégories fondamentales que je viens d'évoquer et d'interpréter les lignes qui les relient :

Les lignes s'interprètent ainsi :

[3] L'énoncé 'Pierre est Paul' est un énoncé d'identité, pas un énoncé attributif du type 'Pierre est gros'. 'Gros' est un universel (ou un trope) et 'Paul' est un individu.



Les diagonales sont des contradictoires. C'est ce que j'appelle 'carréification' et que certains auteurs appellent 'quadrature'. Ces carrés ontologiques sont calqués sur les carrés logiques (carré des énoncés quantifiés ou carré des modalités; ces carrés remontent à Boèce, commentateur d'Aristote, philosophe platonicien, fonctionnaire romain et martyr chrétien). Ces carrés ontologiques ont été récemment remis au goût du jour notamment par J. Lowe (voir aussi Luc Schneider).

Revenons au moment lexical dans la constitution de l'ontologie. Classiquement, on insiste sur la caractérisation de Göckel (1613), en termes d'abstraction et de transcendantaux. Dans l'article « ontologie », cette dernière est en effet définie à partir de ces concepts. Les sciences opèrent par abstraction. Par exemple la physique abstrait les qualités sensibles, les mathématiques la matière concrète. On dira que les diverses sciences ont pour objet non des *concreta* mais des *abstracta* et que ces entités abstraites ont des degrés variables et croissants d'abstraction (la théologie est plus abstraite que la zoologie), l'ontologie concernant les abstraits suprêmes. En effet les sciences connaissent des degrés d'abstraction : un abstrait peut être doté d'un paramètre temporel et il peut être hors du temps (sans être pour autant éternel). En ce qui concerne la théologie par exemple, qu'il ne faut pas confondre avec l'ontologie, et que les bons auteurs n'ont jamais confondu avec elle, contrairement à ce qu'affirment certains heideggériens, elle est non matérielle et non spatiotemporelle, de par la nature même de son objet. Mais en ce qui concerne l'ontologie, à l'abstraction relativement à la matière, au temps et à l'espace, s'ajoute l'abstraction par rapport à tout contenu particulier : l'ontologie n'a pas de contenu particulier. Elle traite de l'être en tant qu'être, et non en tant que mobile, séparé, abstrait, divin… Göckel définit donc l'ontologie comme philosophie de l'être, mais pas au sens où l'être serait un objet particulier, fut-il abstrait et général : non l'être très exactement

n'est rien ou plus exactement l'être qu'il n'est rien au regard des déterminations des autres objets. C'est en ce sens que l'ontologie est une philosophie des transcendantaux (identifiés aux abstraits suprêmes). Ces transcendantaux dans la philosophie médiévale sont des aspects de l'être : l'être à proprement parler n'a pas de contenu, mais il a des aspects : le beau, le bien, le vrai, l'un. L'être n'est pas beau ou bien, il est ce qui rend possible qu'il y ait de la beauté et de la bonté et c'est sous ce rendre possible que l'on a, à défaut d'un contenu, des aspects de l'être[4].

L'abstraction croissante est d'essence aristotélicienne ; les étapes sont les suivantes (cf. *Métaphysique*, A) :

- objets concrets: perception
- objets abstraits: science
- catégories: sémantique
- transcendantaux: ontologie comme métaphysique générale

QUELS LIENS, QUELLES DIFFERENCE ENTRE A1 ET A2 ?

Rappelons que l'ontologie A1 est l'ontologie définie comme une représentation des relations (d'opposition, ou touts/parties) entre concepts primitifs; elle se présente comme un tableau, comme une logique tabulaire (cf. plus haut Ramus et les Ramistes).

L'ontologie A2 est une science de l'être en tant qu'être, des abstraits ultimes en tant qu'ils sont entendus de manière réaliste comme des Transcendantaux au sens d'un découpage de l'être, une classification des étants (plus tard Kant les interprétera de manière anti-réaliste comme des conditions de possibilités du jugement). La division de la métaphysique au milieu du XVII^{e} siècle marqua un tournant dans la mesure où l'ontologie se voit au sein d'une architecture des sciences attribuer une place éminente. En effet la métaphysique se scinda en métaphysique générale (*metaphysica generalis*) et en métaphysique spéciale (*metaphysica specialis*) ce qui consacrait la conception de l'ontologie comme abstraite et transcendantale, mais qui en même temps en faisait une partie de la métaphysique, même si c'était la plus haute. La science de l'être en tant qu'être se

[4] Le nombre et la nature des transcendantaux sont des questions débattues. On peut aussi se demander si des catégories encore plus générales que celles-ci, comme le 'quelque chose' (le *ti* des Stoïciens), ou l' 'objet en général' de Meinong ou Husserl sont ou non des transcendantaux (les bons auteurs les nomment parfois des hypertranscendantaux).

distingue des sciences de l'être en tant que divin, mental ou physique, qui sont l'objet de la théologie, de la psychologie (transcendantale) et de la cosmologie. Le « en tant que » est un opérateur sémantique dont la fonction est de focaliser sur un aspect au détriment des autres : « en tant qu'homme d'affaires Paul est absolument nul » implique (ou présuppose ?) qu'il y a des domaines où Paul n'est pas absolument nul (ou en tout cas la possibilité reste ouverte d'un contraste comme dans : « en tant que professeur il a de l'autorité, en tant que père il a démissionné ».) Le « en tant que » (*qua, quatenus)* est un réduplicatif, il rabat la prédication sur elle-même en la focalisant. Il existe une interprétation méréologique des réduplicatifs qui interprète leur fonction ainsi : ils focalisent la prédication sur une partie du sujet (d'où l'exemple canonique dans la littérature classique : en tant qu'il a des dents l'Éthiopien est blanc !). Ces parties de l'objet (comme par exemple moi en tant que professeur) constituent les '*qua* objets' de Barry Smith, qui jouent un rôle énorme en ontologie sociale : nous sommes socialement des conglomérats de *qua* objets.

L'architecture ontologique est donc la suivante : la métaphysique générale se confond avec l'ontologie générale ; c'est la science de l'être en tant qu'être, des premiers principes, des Transcendantaux (Unité, Bonté, Vérité).

La métaphysique spéciale est diffractée en ontologies spéciales (ou régionales, cf. Husserl, voir plus bas pour ontologie formelle vs. ontologies régionales ou matérielles) ; elle est en effet sous divisée en

- théologie rationnelle ou naturelle (être= Dieu)
- cosmologie rationnelle (être= monde)
- psychologie rationnelle (être = âme)[5]

On retrouvera chez Kant avec les Idées de la Raison ces trois objets de la métaphysique ou de l'ontologie spéciale : Dieu, le Monde et le Moi, dont le caractère commun est d'être non des objets, mais des idées transcendantales,

[5] La théologie rationnelle se distingue de la théologie révélée (elle a pour but avec le seule raison de démontrer l'existence de Dieu et de son gouvernement), la cosmologie rationnelle se distingue de la cosmologie physique (elle raisonne *a priori*) et la psychologie rationnelle se distingue de la psychologie empirique (cf. Brentano) ou expérimentale : elle se borne à une connaissance par la seule raison de la nature de l'âme et des facultés. Naturellement, on doit constater que ces ontologies spéciales ont disparu quand la science s'est emparée de leur terrain (ce qui explique la survivance de la théologie rationnelle, car la science ne s'est pas emparée de Dieu comme objet d'enquête, ou alors très récemment dans un certain type de cosmologie spéculative).

c'est-à-dire des idées ultimes qui rendent possible de penser un secteur de la réalité, suprasensible, physique ou mentale. Ces Idées rendent possible à l'entendement de dégager des lois et de former des concepts en fournissant un idéal d'unification que ce soit dans le domaine du sensible, de l'intelligible ou de l'esprit : pour dégager par exemple des lois physiques pour la nature (causalité notamment) il faut un idéal d'unification des phénomènes physiques dans 'le monde' qui n'est jamais parfaitement totalisé, et donc ne peut être un concept, mais qui est la norme idéale d'un horizon de totalisation. Ainsi les transcendantaux sous une guise épistémique se maintiennent dans la métaphysique kantienne, mais sans donner lieu à une ontologie : il ne s'agit que de fournir l'envers transcendantal de la démarche logique d'application des principes et non de dégager la structure ontologique de la réalité. Il est frappant de constater que les transcendantaux se sont transformés autant que les catégories : les catégories aristotéliciennes deviennent des catégories du jugement chez Kant et de la même manière les transcendantaux de l'être deviennent des idées transcendantales. La dernière étape de ces mutations du transcendantale est fournie par Peirce qui réinterprète les Idées kantiennes comme des aspects de l'être, la priméité, la secondarité et la tercéité.

À côté de la définition extrêmement répandue de l'ontologie comme sciences des transcendantaux et de l'abstraction maximale, Clauberg substitue à la définition classique une définition où l'être devient le quelque chose en général, la chose (*res),* l'Ontologie devenant la science de la chose en général, identifiée au pensable en général, à la pensabilité, à la cogitabilité. L'ontologie est alors définie : science de la pensabilité en général. C'est cette définition que retiennent souvent les heidggeriens français (Marion, Courtine…) qui interprètent ce moment comme la constitution de la métaphysique objectale et psychologisante. Ils ajoutent une critique de l'ontothéologie (cf. transcendantaux et Kant) et combinent trois critiques : de l'ontothéologie, de la conception de l'être comme *res*, de la psychologisation (entendue comme la dégradation de l'être en simple pensable). Reprenons donc la distinction des transcendantaux pour éclaircir ce point d'interprétation de l'ontologie.

Les transcendantaux sont les genres suprêmes de l'être ; ils sont interconvertibles (exemple: la vérité est une) ; ce sont les objets de la métaphysique générale ou de l'ontologie dans la ligne Duns Scot-Kant-Peirce; chacun pris en part est l'objet d'une science spéciale. Les hypertranscendantaux: (quelque chose en général, objet quelconque, *res*, X) sont des objets de l'ontologie dans la lignée Clauberg-Wolff-Meinong ; ils ne sont pas interconvertibles et il n'y a pas de science spéciale qui corresponde au quelque chose en général ou à l'objet quelconque. L'ontologie devient tout simplement la science de l'objet quelconque, ce qui est très différent.

Est-ce que Clauberg relève de l'ontologie de l'être (A1) ou de l'ontologie des concepts (A2) ? Est-ce que la définition de l'ontologie comme science de la chose en général ou du pensable en général est une définition représentationnelle ou contentuelle ? Dans la mesure où la *res* est déterminée comme cogitabilité, on penche vers B : le cogitable est une sous-espèce du représentable. On retrouvera cette définition, modifiée, chez Meinong: l'ontologie devient une théorie de l'objet, éventuellement non existant (alors que dans la définition B de l'ontologie comme science de l'être en tant qu'être, l'être est implicitement existant)

Christian Wolff qui a systématisé l'ontologie (*Ontologia*, 1730) a repris la notion de *res* et dans la lignée de Leibniz il fait du possible l'objet même de l'ontologie. En effet son ontologie a les caractéristiques suivantes : l'ontologie est la science des objets en général, qu'ils existent ou pas ; *Gegenstand* (objet) est synonyme de *aliquid* ou *etwas* (quelque chose ; le possible, ou non contradictoire est un objet de l'ontologie. Meinong, à la fin du XIX[e] siècle dans sa théorie des objets reprendra complètement cette orientation vers l'objet et en partie cette orientation vers le possible (la conception purement logique du possible comme non contradictoire semblant à juste titre insuffisante à Meinong).

L'ETAPE B DE L'ONTOLOGIE

La critique kantienne et heideggérienne[6] a semblé détruire l'ontologie qui a subi une éclipse des années 1800 aux années 1960-1970, à un certain nombre d'exceptions près, la plus importante étant celle de Meinong (mais qui ne reprend pas le terme et revendique plutôt, explicitement une théorie des objets qui rompt radicalement avec la limitation métaphysique de l'ontologie à l'existant et au présent). Si l'ontologie est la science de l'être en tant qu'être Meinong est celui qui a cherché à dépasser cette définition : pour lui une ontologie devrait être la science de l'être en tant que possible et en tant qu'objet. La résurrection des années 60 et 70 a été causée en partie justement par le développement de la sémantique des logiques modales, d'une part en sémantique formelle, d'autre part en sémantique des langues naturelles. Un livre fondamental dans l'émergence d'une ontologie des mondes possibles, un outil de la sémantique des logiques modales, *Counterfactuals* de D. Lewis, en 1972, porte par exemple sur un phénomène sémantique celui des conditionnels irréels (« si les kangourous n'avaient pas de queue, ils sauteraient moins loin ») qui réclame pour son

[6] La critique heideggérienne a été conduite à partir des années 1920-1930 ; elle a visé explicitement la destruction de la métaphysique et donc de sa partie essentielle, l'ontologie. De formation catholique, Heidegger a été éduqué philosophiquement dans l'ontologie néo-scolastique qu'il a toujours violemment rejeté.

interprétation la construction d'un modèle de mondes possibles comparés par des relations de ressemblance.

L'ontologie a resurgi sous deux formes: l'ontologie des systèmes de représentation (notamment en IA), l'ontologie analytique (Quine, S. Kripke, B. Smith, P. Simons…) – on retrouve grosso modo la distinction A1 vs. A2 avec B1 vs B2.

Voici deux longues citation d'acteurs typiques de l'ontologie B2 à l'époque du développement des systèmes d'information et d'intelligence artificielle :

> Ontology, on the other side, can be seen as the study of the organisation and the nature of the *world* independently of the form of our knowledge about it.
>
> *Formal* ontology has been recently defined as "the systematic, formal, axiomatic development of the logic of all forms and modes of being" [Cocchiarella 1991]. Although the genuine interpretation of the term "formal ontology" is still a matter of debate, this definition is in our opinion particularly pregnant, as it takes into account *both* the meanings of the adjective "formal": on one side, this is synonymous of "rigorous", while on the other side it means « related to the *forms* of being ».
>
> Therefore, what formal ontology is concerned in is not so much the bare existency of certain individuals, but rather the rigorous description of their forms. In practice, formal ontology can be intended as the theory of a priori distinctions: among the entities of the world (physical objects, events, regions, quantities of matter...); among the meta-level categories used to model the world (concepts, properties, qualities, states, roles, parts...). (Nicolas Guarino : « Formal Ontology, conceptual Analysis and Knowledge Representation »)

On remarque dans ce texte une transformation de la notion d'ontologie formelle. Cette dernière (une invention de Husserl) a pour but de dégager la forme générale de toutes les ontologies locales, en entendant par là les ontologies de régions de savoir, c'est-à-dire le relevé de toutes les connexions objectives à l'intérieur de ces régions et entre ces régions. Pour Husserl dans les *Ideen* l'ontologie formelle a finalement pour objet l'être en général, l'objet quelconque, dont on a parlé plus haut, à propos de la filiation Clauberg-Wolff-Meinong. D'autre part on peut remarquer que Guarino donne une interprétation qui peut apparaître contradictoire de l'ontologie : elle est à la fois indépendante de l'esprit et identifiée à des distinctions conceptuelles (*a priori*). La seule manière de sortir

de cette contradiction serait de faire de ces distinctions conceptuelles *a priori* des distinctions immanentes aux formations et relations objectives elles-mêmes, ce qui poserait de sérieux problèmes. Mais s'il s'agit là d'une difficulté de ce genre de position, elle confirme bien notre interprétation des ontologies B2 comme des systèmes de concepts. Un auteur comme Guarino semble s'appuyer sur la définition husserlienne de l'ontologie formelle pour donner une interprétation purement épistémique de la forme des objets, qui finalement se confondra avec les concepts formels des objets. C'est cette tendance qui est encore plus nette dans la seconde citation, de Gruber

> An ontology is an explicit specification of a conceptualization. The term is borrowed from philosophy, where an Ontology is a systematic account of Existence. For AI systems, what "exists" is that which can be represented. When the knowledge of a domain is represented in a declarative formalism, the set of objects that can be represented is called the universe of discourse. This set of objects, and the describable relationships among them, are reflected in the representational vocabulary with which a knowledge-based program represents knowledge. Thus, in the context of AI, we can describe the ontology of a program by defining a set of representational terms. In such an ontology, definitions associate the names of entities in the universe of discourse (e.g., classes, relations, functions, or other objects) with human-readable text describing what the names mean, and formal axioms that constrain the interpretation and well-formed use of these terms. Formally, an ontology is the statement of a logical theory. (Gruber What is an Ontology ?)

On peut difficilement être plus clair : ce qui existe c'est ce qui peut être représenté (cela résonne comme un écho du esse est percipi de Berkeley 7). Ce qui est très intéressant dans ce texte c'est l'assimilation du domaine de l'être, de ce qui est, (normalement le domaine de l'ontologie) à l'univers de discours. On sait qu'en sémantique des langages logiques, l'univers de discours est l'ensemble des entités à notre disposition pour interpréter (au sens d'assigner des valeurs de vérité aux formules) le langage formel en question et on sait que cet univers de discours est soit fixe (Frege et la tradition frégéenne) soit variable (Boole-Peirce-Schröder et toute cette tradition dont Putnam a montré l'importance). Gruber reprend le schéma de l'interprétation des langages formels pour son ontologie : on interprète des textes à l'aide d'un univers de discours, qui contient des programmes. Il s'agit d'une ontologie des programmes. C'est aussi peu une

[7] « Ce qui est c'est ce qui est perçu » ou « l'être c'est le perçu ». On considère en général ce principe comme la source d'un idéalisme objectif.

ontologie, au regard de l'ontologie classique ou A, que la valse des prix est une danse.

Il continue :

> We use common ontologies to describe ontological commitments for a set of agents so that they can communicate about a domain of discourse without necessarily operating on a globally shared theory. We say that an agent commits
>
> to an ontology if its observable actions are consistent with the definitions in the ontology. The idea of ontological commitments is based on the Knowledge-Level perspective (Newell, 1982). The Knowledge Level is a level of description of the knowledge of an agent that is independent of the symbol-level representation used internally by the agent.

Il y a ici une grande différence dans la conception de l'engagement ontologique avec le sens de Quine: un engagement ontologique via une quantification envers une entité x introduit x dans l'ontologie (exemple: individu, événement). Ici l'engagement est celui des agents dont les actions sont compatibles avec les définitions de l'ontologie. Le niveau de connaissance de l'agent permet de déterminer la nature et l'ampleur de cet engagement. D'un point de vue ontologique, cela revient à relativiser l'engagement ontologique à la connaissance des agents particuliers, ce qui renforce le caractère radical du virage épistémique. L'ontologie n'est plus une théorie de l'objet, ou même du possible, mais une simple astuce qui relie les connaissances des agents à des univers de discours conceptuels.

Une autre différence avec Quine (qui est l'inventeur du critère d'engagement ontologique) est que pour ce dernier l'ontologie est relative à une théorie (scientifique en général), et donc qu'il n'y a pas d'ontologie de la langue naturelle pour Quine et que les systèmes de concepts des agents relèvent de l'idéologie, non de l'ontologie – rappelons qu'une idéologie d'une théorie T est l'ensemble des concepts primitifs de T, tandis qu'une ontologie relative à T est l'ensemble des entités à l'existence desquelles s'engage T dans ses énoncés normaux, c'est-à-dire quantifiés (puisque tout doit être exprimé dans la logique du premier ordre avec quantification). L'ontologie concerne les entités primitives d'une théorie (Exemple: les nombres [en mathématiques], la cellule [en biologie], alors que l'Idéologie concerne les concepts fondamentaux d'une théorie (Exemple: 'axiome' [en mathématiques], 'reproduction sexuée' [en biologie]

Cependant ce qui est commun à ces différentes versions est un engagement ontologique qui passe par le langage formel de représentation. L'ontologie d'une théorie (scientifique) a été définie par Quine comme l'ensemble des entités à l'existence desquelles engagent les énoncés quantifiés présents dans cette théorie.

Pour Gruber l'ontologie est relative à un programme, non à une théorie. Dans les deux cas une formalisation, une symbolisation et, peut-être, une axiomatisation (?) sont nécessaires avant de construire l'ontologie, sauf que la nature procédurale des programmes est encore plus exigeante que la nature purement déclarative des théories scientifiques. Se pose donc différemment le problème de l'ontologie naïve (cf. 'physique naïve', 'sémantique naïve') dans l'ontologie B1 et dans l'ontologie B2. Dans l'ontologie B1 il s'agit d'avoir des procédures de traduction ou de réduction de l'ontologie naïve (celle des agents rationnels aux prises avec les apparences et les régularités du monde), bref des règles de passage du naïf au savant, du matériel au formel, de la perception individuelle à la formalisation collective, alors que dans l'ontologie B2 on formalise directement cette ontologie naïve en la réduisant non à des bouts de science, mais à des programmes de manipulation de contenus sensibles par des agents.

Cependant, il ne faut pas durcir le propos, forcer le trait et opposer trop radicalement les deux types d'ontologies, B1 et B2: il existe des pistes qui ont été explorées de l'ontologie opérationnelle ou informatique à l'ontologie spéculative analytique. En voilà par exemple un témoignage, par l'un des auteurs qui a fait le plus pour les rapprocher :

> The methods of philosophical ontology are the methods of philosophy in general. They include the development of theories of wider or narrower scope and the testing and refinement of such theories by measuring them up, either against difficult counterexamples or against the results of science. (…) Some philosophical ontologists conceived ontology as being based on a special a priori insight into the essence of being or reality. (…) Seen from this perspective ontology is like physics or chemistry; it is part of a piecemeal, on-going process of exploration, hypothesis-formation, testing and revision. Ontological claims advanced as true today may well be rejected tomorrow in light of further discoveries or new and better arguments. (Barry Smith)

Barry Smith donne ici une vision de l'ontologie B1 qui est B2-compatible. L'ontologie philosophique apparaît en effet comme empirique, révisable et donc rein n'empêche a priori de la modifier pour la rapprocher de l'ontologie B, et c'est ce qui est concrètement le cas dans l'utilisation de l'ontologie philosophique,

notre ontologie A, dans des programmes de constitution de banques de données bio-médicales par exemple.

L'ontologie philosophique pour Barry Smith est une ontologie réaliste dans le sens suivant :

> Philosophical ontology as I shall conceive it here is what is standardly called descriptive or realist ontology. It seeks not explanation but rather a description of reality in terms of a classification of entities that is exhaustive in the sense that it can serve as an answer to such questions as: What classes of entities are needed for a complete description and explanation of all the goings-on in the universe? Or: What classes of entities are needed to give an account of what makes true all truths? Or: What classes of entities are needed to facilitate the making of predictions about the future? Sometimes a division is made – as for example in the case of Husserl and Ingarden – between formal and material (or regional) ontology. Formal ontology is domain-neutral; it deals with those aspects of reality (for example parthood and identity) which are shared in common by all material regions. Material ontology deals with those features (for example mind or causality) which are specific to given domains. (idem)

On remarque que la définition et la fonction de l'ontologie formelle sont fort différentes de ce que déclarait plus haut Guarino. Pour Barry Smith l'ontologie formelle traite d'aspects de la réalité, transversaux à tous les domaines d'icelle (cf. la fin de la citation (je souligne les passages importants pour notre propos) : Therefore, what formal ontology is concerned in is not so much the bare existency of certain individuals, but rather the rigorous description of their forms. In practice, formal ontology can be intended as the theory of a priori distinctions: among the entities of the world (physical objects, events, regions, quantities of matter...); among the meta-level categories used to model the world (concepts, properties, qualities, states, roles, parts...). Pour Guarino en termes techniques l'ontologie formelle concerne la base catégorielle du monde, tandis que pour Barry Smith elle concerne les connexions formelles a priori comme les connexions méréologiques ou les concepts transcendantaux comme l'identité.

Quant à la pulsion réaliste de Barry Smith, notons qu'il existe au moins quatre versions réalistes de B1, Barry Smith occupant une place dans cette combinatoire:

- ontologie néo-aristotélicienne (Barry Smith, Lowe, Pouivet, Schneider…). Cette ontologie formalise l'esquisse d'ontologie

catégorielle dont j'ai parlé plus haut (celle qui correspond au 'carré ontologique')

- ontologie tropiste. Cette ontologie n'admet comme primitifs ontologiques que des propriétés particulières, et les objets sont considérés comme des faisceaux (en un sens non technique) de propriétés particulières ou 'tropes'. .
- ontologie néo-meinongienne (Dale Jaquette, Reinhardt, …). Cette ontologie reprend le programme meinongien qui a été interrompu pendant une quarantaine d'années, en partie à cause des critiques (injustes) de Russell.
- ontologie modale (Lewis…). Ce type d'ontologie analytique développe le second volet de l'ontologie wolffienne : une réflexion sur le possible, mais à partir d'une définition non strictement logique du possible (comme non contradictoire), mais d'une définition ontologique, le possible devenant une quantification, et donc un engagement ontologique, sur des mondes.

À première vue le type de réalisme qui correspond à une ontologie en vue d'une terminologie scientifique appartient soit à la première, soit à la troisième des possibilités énoncées ci-dessus. L'existence d'une base catégorielle dans la première facilite le passage du vocabulaire à la liste des entités, en distinguant bien ce qui est type et occurrence et en donnant éventuellement un composant procédural qui permet de dériver des catégories dérivées de catégories primitives. De plus il semble que l'investissement dans la catégorie de la substance ne pose pas de problème à une enquête qui aborde la réalité à partir d'une description de type lexical ou grammatical. Quant à la troisième (l'ontologie néo-meinongienne) elle a ceci d'attirant pour l'enquête terminologique qu'elle s'ouvre au non existant, qui peut lui aussi être l'objet d'une expression dans la terminologie par exemple des fictions et des possibilités, ce qui recouvre tout le vocabulaire explicatif, interprétatif dans les sciences humaines. Une ontologie terminologique par exemple du discours historique gagnerait à cette ouverture sur le non existant, car les situations contrefactuelles dans un premier temps sont dites ne pas exister, et l'on comprendrait que l'investissement réaliste dans toutes les situations contrefactuelles serait pour la terminologie plus un obstacle qu'un avantage.

Les types d'entités à l'existence de laquelle s'engagent ces ontologies varient suivant les quatre choix distingués sommairement :

- Substances, accidents, relations, universaux, modes… pour l'ontologie néo-aritotélicienne.
- Particuliers abstraits, relations individuelles, structures ontologiques… pour l'ontologie tropiste (mais les particuliers abstraits, qui correspondant aux propriétés particulières sont des primitifs)
- Objets, propriétés, objets d'ordre supérieur, objets non existants, relations…dans l'ontologie néo-meinongienne.
- Mondes possibles, propriétés, relations…pour l'ontologie du réalisme modal.

On peut noter qu'il y a un noyau commun: le triplet objet/propriété/relation[8], et que ce qui ce qui divise le plus est la question de l'universel et du particulier

Mais, aussi important et peut-être moins facile à déceler, les modes d'existence varient également d'ontologie B1 en ontologie B1 : L'ontologie A distinguait être et exister, essence et existence, attribuait aux accidents un être moindre, pouvait nier que les relations existassent et leur conférait alors un statut d'entité mentale. L'ontologie néo-meinongienne distingue exister et subsister. La thèse quinienne généralement admise est qu'il y a un seul sens de l'existence – mais il y a aussi des fictions.

Il existe cependant une forme générale d'une ontologie qui traverse la variété des modes d'existence. Une ontologie structurale contient des structures ontologiques qui sont définies comme suit: <E,R>, i.e. un ensemble d'entités et un ensemble de relations entre ces entités. Parmi ces relations hiérarchiquement supérieures (ou transcendantales ?): la compresence entre particuliers abstraits, l'exemplification des universaux, l'instanciation, dépendance, survenance, etc.

Revenons aux hypothèses de départ

Notons que le réalisme philosophique se déploie sous une forme quadruple dans l'ontologie analytique contemporaine :

[8] C'est la raison pour laquelle j'a consacré un premier livre à l'objet (*L'objet quelconque*, Vrin, Paris 1998), un second aux propriétés (*Les propriétés des choses*, Vrin, Paris, 2006) et que je prépare un livre sur la connexion. Entre 1998 et 2006 je suis passé d'un néo-meinongien actualiste à un tropisme possibiliste modéré.

- H1. La différence entre A et B renvoie à une différence à l'intérieur de A entre A1 et A2:

- H1 est vérifiée en partie seulement : A2 annonce l'évolution contemporaine de l'ontologie vers la représentation des structures de concept dans des arbres et déjà dans le conflit entre les deux conceptions classiques de l'ontologie se dessine l'évolution vers une conception plus modeste de l'ontologie où celle-ci est en fait équivalente à un langage de représentation des structures conceptuelles. Mais en fait il y a une tendance B1 plus réaliste qui dans un certain sens accomplit A1: on a pu montrer par exemple qu'il y a une filiation Wolff/Meinong. Il faut donc modifier H1 et on aboutit à H2

- H2. La différence entre A1 et A2 étant comparable ou analogue à celle dans B entre B1 et B2, on peut vérifier H2: il y a un changement important de A à B, car notamment le contexte métaphysique et scientifique n'est pas le même, mais l'opposition entre réalisme et anti-réalisme traverse ce changement de contexte.

S'il existe des arguments forts pour la viabilité d'une ontologie structurale, alors il existe des raisons de distinguer nettement entre terminologie et ontologie (au sens de deux disciplines) : la première établit des liens structuraux entre des réalités mi physiques mi intentionnelles que sont les significations des termes lexicaux, tandis que l'ontologie établit des liens entre des réalités de divers types et ressortissant de différents domaines. En ce sens la terminologie est une ontologie structurale plongée dans une factualité restreinte: les terminologues sont comme Monsieur Jourdain, ils font de l'ontologie sans le savoir… et les ontologues s'ils négligent la terminologie sont prompts à se prendre les pieds dans le tapis du lexique, risquant de prendre les vessies du langage pour les lanternes de la réalité – que l'on pense ici à certains développements sur les termes de masse.

Posons pour finir une dernière hypothèse : la terminologie est neutre relativement à l'opposition ontologie *vs* idéologie (au sens quinien).

Références

Guarino N. *(1998) Formal ontology in information systems.* In Formal Ontology in Information Systems, N. Guarino (Ed.), IOS Press, Amsterdam, p. 3-15.

Guarino, N., Oberle, D., and Staab, S. (2009*) What is an Ontology*? In S. Staab and R. Studer (eds.), Handbook on Ontologies, Second Edition. International handbooks on information systems. Springer Verlag: 1-17.

Munn K. Smith B. (2008) *Applied Ontology. An introduction* Metaphysical Research vol. IX

Nef, F. (2009) *Les catégories aristotéliciennes et la division de l'être : types de divisions et types d'ontologies,* Les diviseurs de l'être, Cahiers Philosophie de Caen, Presses Universitaires de Caen, V. Carraud & S. Chauvier éds.

Ohrstrohm, P, Schärfe, H., Uckelman S. (2008) *« Jacob Lorhard's Ontology:A 17th Century Hypertext on the Reality and Temporality of the World of Intelligibles »* Proceedings ICCS, 16th International Conference on Conceptual Structures

Ohrstrohm, P, Schärfe, H., Uckelman S. (2007) *« Historical and Conceptual Foundation of Diagrammatical Ontology »,* Proceedings ICCS, 15th International Conference on Conceptual Structures

Smith, B. *(2003) Ontology,* in L. Floridi (ed.), Blackwell Guide to the Philosophy of Computing and Information, Oxford: Blackwell, 155–166.

Wolf, C. *(1962) Philosophia Prima sive Ontologia,* 1730, rééd. J. Ecole éd, G. Olms, (Gesammelte Werke section II, vol.3)

Yates F (1966) *The Art of Memory*, Routledge and Kegan (traduction française : L'Art de la Mémoire, Gallimard, 1975)

ARTICLES

Le travail sur la représentation (visuelle) des connaissances en terminologie : un retour d'expérience

Dardo de Vecchi

Résumé : La représentation visuelle des connaissances peut se faire de multiples manières. Depuis la cartographie en passant par les arborescences terminologiques ou les ontologies, le travail sur les notions recourt à l'image en mettant en relation les notions et en les exploitant de manières différentes. Les relations entre les notions sont explicitées dans chaque cas de manière différente : à partir de l'expérience, faisant appel à la sémantique ou à la logique. Cependant le travail sur les notions n'implique pas un travail sur le visuel, mais lorsque le mode image n'est pas disponible, le travail exige une transposition (cas des aveugles et malvoyants). Les arborescences terminologiques, des images à part entière, sont un exemple riche en renseignements. En tant qu'outils, elles fournissent une autre manière d'approcher les connaissances d'un domaine qui est exportable à des situations. Le travail sur les termes se distingue du travail avec des termes. Le premier mène à des terminologies. Le second à un autre type de réflexion qui peut s'orienter vers l'organisation des connaissances ou encore vers la sauvegarde des savoir-faire dans les métiers où les moyens - notamment financiers - ne permettrait pas d'avoir une démarche en faveur d'un travail terminologique ou d'ontologies.

Mots-clés : terminologie, représentation visuelle, arborescences, cartographie, connaissances, métiers, gestion des connaissances

1. Introduction

Le rapport entre l'image et les connaissances mérite l'attention du terminologue et des ontologues. L'étymologie est à ce titre fort utile. Pour nombre de dictionnaires actuels, représenter est *mettre devant les yeux, rendre sensible par l'image*. Il nous intéresse ici d'étudier cet aspect en partant de la terminologie pour explorer des nouvelles applications. Cependant, toute représentation des connaissances exige des questions liminaires dans la mesure où leurs réponses impacteront le résultat visuel.

Pourquoi représenter ? Parce que selon le désir ou le besoin, les objectifs et les moyens mis en œuvre seront différents. Que représente-t-on ? La réponse est multiple : ce que l'on veut ou doit savoir dire ou faire, un modèle, un point de vue, la vérité, voire une manipulation[1]… Qui représente : un individu ou un groupe ? Cela suppose un accord préalable avec soi-même ou avec les « constructeurs » de la représentation pour ce qui est de l'utilisation des signes. Comment représenter ? Le type d'image est à préciser parce qu'elles ne s'équivalent pas toutes (carte, schéma, arbre, image, texte…) Sur quel support représente-t-on ? L'appel à une surface, à un volume, aux deux, de manière réelle ou virtuelle affecte le résultat. Pour qui représente-t-on ? L'expert, le néophyte ou le curieux ont des besoins bien différents.

Voici quelques réponses. Les notions à représenter doivent pouvoir être interprétées. Les rapports entre ces notions doivent aussi être interprétées. Le mode de représentation doit être accessible (du point de vue sensoriel et intellectuel). Le destinataire de la représentation doit être précisé. Le résultat doit pouvoir, selon le cas, être partageable. Il faut cependant préciser – en relativisant la portée des informations recueillies - que la représentation n'est qu'un point de vue résultant d'une stratégie d'approche des connaissances représentées, à un instant *t*. Nous entendons par connaissances en accord avec Costa et Silva (2008 : 5) : « l'ensemble de choses qui sont sues et connues par un individu en tant que membre d'une communauté de spécialité ». Par spécialité, nous entendons l'utilisation théorique ou empirique de ces connaissances et par communauté le groupe d'utilisateurs.

Dans l'enseignement de la terminologie en L3 à l'Université Paris Diderot – EILA, les réponses que nous venons de voir d'apporter comme des prérequis au

[1] Pour le rapport entre cartographie et manipulation lire Jean-Pierre Bord in bibliographie.

travail terminologique et terminographique dont l'un des résultats est justement une représentation visuelle partageable des notions abordées. Mais cette image n'est pas « partageable » lorsque l'enseignement s'adresse à des aveugles ou malvoyants ce qui implique de devoir transposer un système de représentation en un autre. L'arborescence terminologique fait découvrir une autre manière de représenter un domaine notionnel. Et c'est cette « image » qui a été exploitée non comme résultat mais comme point de départ dans un cours d'initiation à la gestion des connaissances à Euromed-Management – École de management. Un bref rappel historique faisant écho à nos questions liminaires nous sera utile pour orienter la réflexion autour de l'idée de relation entre notions. Nous présenterons ces enseignements et ce que le travail sur l'image apporte. Avant de conclure, nous ferons une réflexion sur les champs d'application où la gestion des connaissances est un besoin constant : les entreprises.

2. Représentations et transmissions multimodales

La communication et le partage de connaissances font appel à l'utilisation de signes très différents de nature. Si lors de l'utilisation de la langue – orale ou écrite –, le mode texte reste un procédé privilégié voire un réflexe, il n'est pas le seul : le recours au mode image est souvent de rigueur. La cartographie et la réalisation de schémas, de diagrammes ou de plans l'attestent. La transmission et le partage de cette connaissance en dépendent au point que, pour certaines disciplines comme l'architecture, les plans sont indispensables. L'utilisation de langue et de l'image est souvent concomitante voire complémentaire.

Le recours à l'image est à l'œuvre partout pour autant que l'image renvoie à une réalité perceptible et saisissable par ceux qui partagent un univers expérientiel. C'est tout le sens de l'icône de Peirce (Peirce, 1978 : 140 et 148). Lorsque nous sommes devant une icône que nous ne reconnaissons pas et que nous ne pouvons pas interpréter, l'information qu'elle est censée transmettre nous échappe et les connaissances desquelles elle participe se voient en parties occultées. Par exemple : les icônes utilisées dans beaucoup de communautés de travail sont opaques pour ceux qui ne les utilisent pas. C'est donc la capacité de l'image et de sa mise en situation qu'il faut interroger pour pouvoir transmettre des informations et construire des connaissances.

« Chaque cognition est le résultat d'un processus psychique, qui débouche sur la connaissance. Ce processus n'est pas un état mais une activité du sujet. Tout comme la connaissance, la cognition est une chose psychique, propre à l'individu. Il ne peut y avoir de cognition objective détachée » (Felber, 1978 : 87). La citation de Felber est importante ; si nous considérons que la connaissance est

un processus psychique, son extériorisation puis sa transmission se font grâce à l'utilisation de systèmes sémiotiques dont « certains » font appel à l'image. L'image peut apparaître aussi comme un assemblage de symboles dont l'interaction permet de « visualiser » un message et dont la globalité représente *in fine* des connaissances à condition que l'ensemble puisse être perçu, reconnu puis interprété. Tel est le cas notamment des cartes géographiques, symboles du réel, qui n'acquièrent leur valeur que dans la mesure où les signes et consignes de lecture sont partagés par l'émetteur et le récepteur. La recul du cartographe est instructif, Luc Cabrézy écrit : « […] le géographe *découpe* le monde en objets thématiques et spatiaux […] De la part des géographes, ce souci de classification s'accompagne d'un effort tout particulier pour transmettre et faire partager les connaissances accumulées dans un langage certes codé mais supposé être accessible par le plus grand nombre ». (Cabrézy, 1995 : 129-130)

FIG. 1 – Charif Al Idrisi (v. 1100 – v. 1165) : la région de la Méditerranée, dans **Le livre de Roger**.

La carte d'Idrissi (Fig. 1) représente en image les connaissances que l'on possède sur la Méditerranée à son époque. Il y a des conventions, mais aussi des résultats de la perception (mer bleue ou montagnes représentées comme des chaînes). Chaque notion géographique est associée à un symbole graphique. Mer : couleur bleue et filets ondulés blancs. Cours d'eau : couleur verte. Villes : rosettes dorées. Montagnes : chaînes. L'ensemble de la codification permet de reconstituer les connaissances représentées[2], mais les liens entre les notions ne

[2] *Bibliothèque nationale de France,* al-Idrîsî : la Méditerranée au XIIe siècle

http://classes.bnf.fr/idrisi/pres/index.htm

sont pas explicités en tant que tels. Pour signaler la notion de « mer » ou de « terre », il est nécessaire de représenter aussi leur contiguïté pour que l'ensemble puisse représenter une notion d'île, de lac ou de fleuve. Une carte en noir et blanc sans nuances de gris est difficilement interprétable surtout si les formes ne sont pas reconnues. Les anamorphoses fréquentes en géographie brisent ce lien à l'expérience. La figure 2 ci-dessous montre l'éloignement des villes par le réseau ferroviaire en France en 2001. Bien qu'éloignées dans le temps, dans ces deux cartes ce sont des codes de lecture qui se combinent avec des formes pour transmettre ou partager une connaissance.

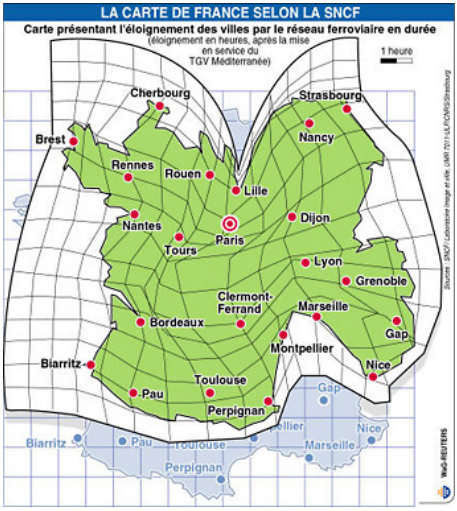


FIG. 2 – Carte de l'éloignement des villes en France en 2001.

Dans les schémas (histogrammes, courbes, nuages de points, etc. selon le support choisi), les notions codifiées se juxtaposent et l'ensemble, si interprétable, véhicule des connaissances. La carte géographique ou l'arborescence terminologique sont visuellement dans la même situation. Contrairement au mode textuel, le mode visuel synthétise et rend très économique la représentation de l'expérience et des connaissances.

3. Le travail à l'Université Paris Diderot - EILA

3.1. Les arborescences terminologiques

L'enseignement de la terminologie à l'Université Paris Diderot, dans l'UFR d'Etudes interculturelles et langues appliquées se fonde sur une mise en parallèle

des acquis théoriques et de leur application à travers la réalisation d'un dictionnaire terminologique sous la forme de base de données et accompagné d'une arborescence terminologique. Il importe que l'étudiant diplômé (L3) puisse exploiter à son niveau et dès la fin du cursus des éléments théoriques et pratiques de manière autonome. L'extraction de termes est manuelle tout comme la représentation notionnelle. Il n'est pas prévu à ce stade de la formation initiale d'utiliser un langage formel tel que UML (*Unified Modeling Language*)[3], le programme et le temps impartis à nos enseignements ne le permettant pas. En conséquence, le recours à toute démarche informatique est réduit à la manipulation des outils de recherche documentaire et à la manipulation des bases de données.

Par groupes, les étudiants travaillent sur un sujet de leur choix, validé par les enseignants. Ils constituent un corpus à l'aide du cours de recherche documentaire à partir duquel ils sélectionnent des termes à traiter dans des fiches terminologiques dans une base de données.

Le premier travail de représentation visuelle est celui de l'arbre de domaine. Il permet de défricher le mode d'approche, condition préalable à la détermination exacte du domaine à traiter. Par exemple : dans le domaine de la kinésithérapie, le travail terminologique fait appel à plusieurs domaines connexes. En effet, une terminologie de la « pratique de la kinésithérapie des insuffisances respiratoires chez l'enfant » fait appel à des domaines connexes comme les pathologies des poumons, leur diagnostic ou leur traitement. Ces domaines ne peuvent être écartés sous peine de rendre le travail final incomplet ou incohérent. Cependant, leurs termes font-ils partie de la kinésithérapie ? La réponse est négative mais il faut en tenir compte pour la compréhension du domaine choisi, du choix des termes et pour enfin réaliser l'arborescence. Il faut donc à chaque fois élaborer une stratégie particulière pour aborder le domaine en fonction notamment du destinataire du produit terminologique.

L'appel à des domaines « lointains » est fréquent. Une *terminologie de la mode* est vite confrontée à une chaîne qui va depuis la production de l'étoffe ou des accessoires, à l'histoire ou au marketing qui doivent être mis en place. De la même manière, une *terminologie du marketing du ballet classique* nécessite l'investigation de domaines très écartés de la danse proprement dite comme la politique culturelle, les ressources humaines ou la finance. La « domaine de

[3] La modélisation proposée par Kockaert et Bassey (2008) n'est pas pertinente à ce stade de l'enseignement.

connaissance » a parfois du mal à trouver ses frontières. Le travail sur le visuel joue dans cette tâche un rôle primordial et permet de mettre en contact des domaines *a priori* non connexes mais dont l'interconnexion est nécessaire pour extraire la réalisation du dictionnaire.

Dans de nombreux cas, les liens entre termes exigent des « astuces visuelles » pour être justifiés. Le support en deux dimensions trouve sa limite et le travail en 3D devient souhaitable. L'arborescence de la *terminologie du VTT* (vélo tout terrain) montre le passage de l'arbre (partiel) du domaine des véhicules (haut de la Fig. 3) à celui du VTT proprement dit[4]. La figure 4 fait un zoom sur des détails de l'arborescence terminologique et montre la complexité visuelle des relations entre termes qui exigent des lignes distinctives pour chaque type de relation.

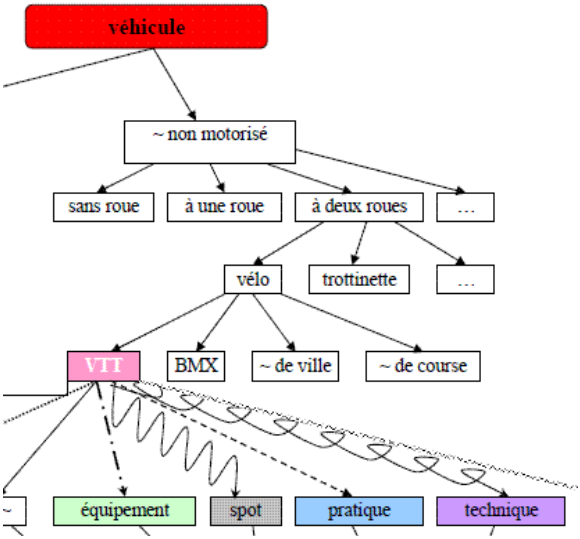


FIG. 3 – Arborescence de la terminologie du VTT (p. 3).

[4] Consultable dans le site de l'Université Paris Diderot - EILA : *http://www.eila.univ-paris-diderot.fr/user/mojca_pecman/terminologie*

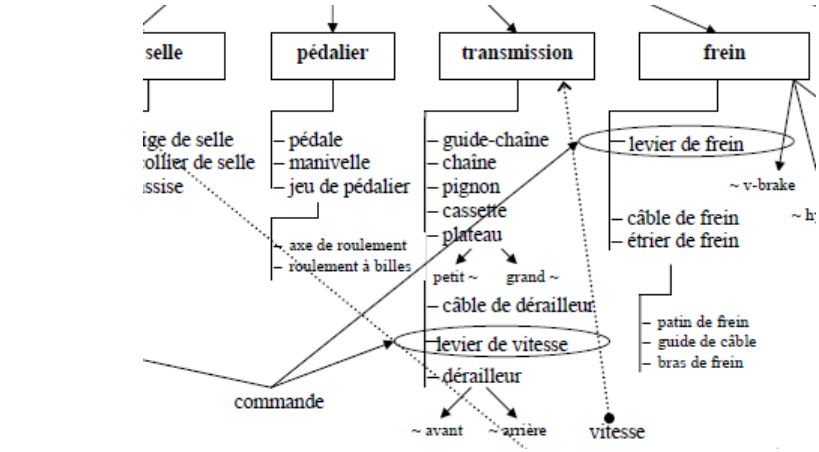


FIG. 4 – Détail de l'arborescence de la terminologie du VTT (p. 2).

Le résultat visuel de la représentation notionnelle est une image dont les codes de lecture permettent comme au § 2 de rendre compte des connaissances, aussi partielles soient-elles, du sujet choisi. L'arborescence permet aussi d'en discourir et d'expliquer le domaine abordé autrement que par la rédaction d'un texte.

Dans le cas de l'arbre de domaine ou de l'arborescence terminologique, l'image résultante constitue un point de départ pour l'exportation de la méthode à d'autres manières de traiter les connaissances comme nous le verrons plus bas (§4).

3.2. Le travail avec des aveugles et des malvoyants : une remise en question du « visuel » en terminologie

Tout ce que nous avons évoqué jusqu'ici est valable tant que le mode visuel est opérationnel. En revanche, que se passe-t-il lorsque le mode visuel ne peut pas être utilisé ? « Tout travail terminologique a pour point de départ les notions » (Felber, 1987 : 82). Il n'y a aucune raison pour que la terminologie soit dépendante de la représentation visuelle. Les modèles représentationnels en arbre, en tableaux, grilles, numériques, etc. (Felber, 1987 : 112 et s.) ne sont pas adaptés en l'état aux déficients visuels. Le travail mentionné dans les paragraphes précédents doit en conséquence être adapté. L'utilisation du tableau classique de la salle de cours est inopérante et les énonciations de l'enseignant doivent être formulées autrement, les déictiques perdant leur utilité car le fait de signaler une image perd tout son objectif.

Outre la mise en place d'ordinateurs adaptés à la lecture en braille, la recherche d'un autre moyen s'est imposée lorsque deux étudiants sont venus au cursus ILTS (Industries de la langue et traduction spécialisée de l'Université Paris Diderot - EILA). Le travail a montré qu'un codage des termes permet de transposer la représentation visuelle en une autre qui peut être suivie par l'étudiant. Bien que l'exercice n'ait pas pu être mené jusqu'à la fin[5], la méthode montrée dans la figure 5 s'est révélée assez efficace. Dans l'adaptation, à chaque terme est attribué un code alphanumérique utile au non-voyant tandis que l'arborescence est maintenue sous son format « classique ». Le partage de la représentation a été ainsi rendu possible. Etant donné qu'il est nécessaire de faire référence à la hiérarchie immédiate supérieure, chaque terme subordonné reprend le code qui correspond au superordonné et ajoute un chiffre.

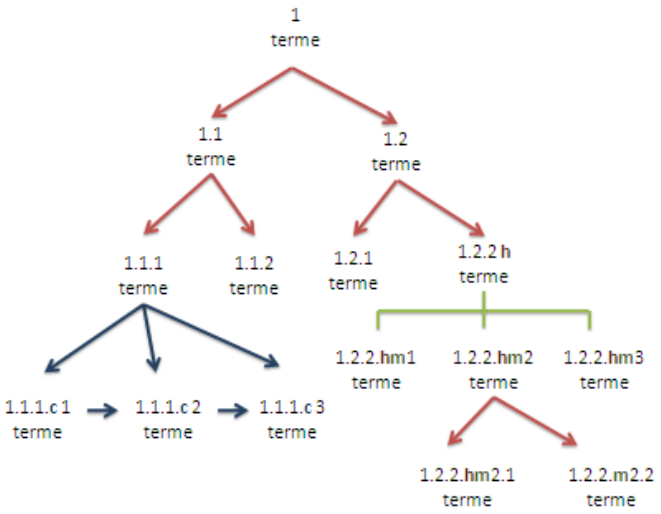


FIG. 5 –Transposition d'interface d'une arborescence.

Par défaut, toute relation est de type hyperonyme-hyponyme. Par exemple : le terme 1.1 possède deux hyponymes, les termes 1.1.1 et le terme 1.1.2. Les isonymes se trouvent clairement distingués. Si la liste devait être prolongée mais la mention d'autres isonymes n'était pas nécessaire, les isonymes possibles non représentés -habituellement et marqués par trois points de suspension - sont

[5] Il fut remplacé par une utilisation adaptée d'un tableur (le logiciel de bases de données étant peu apte à l'utilisation dans ce cas).

remplacés par l'hyperonyme suivi d'un « n ». Par exemple : 1.1.n, étant donné que l'utilisation des « … » prêterait à confusion.

Lorsque la relation est de type holonyme-méronyme, le terme holonyme est suivi d'un « h » signalant son caractère d'holonyme. Ses méronymes reprennent l'holonyme et signalent son caractère de méronyme en ajoutant un « m » et un « numéro d'ordre » après le code attribué à l'holonyme superordonné. Par exemple : 1.2.2.h est un holonyme dont les méronymes sont 1.2.2.hm1, 1.2.2.hm2 et 1.2.2.hm3. Si la liste devait être prolongée mais la représentation concrète non nécessaire, comme pour le cas des isonymes, un « n » suit un méronyme suggéré.

Lorsqu'il s'agit d'une relation chronologique, les étapes appellent la hiérarchie supérieure suivie d'un « c » suivie du numéro de l'étape. Par exemple : 1.1.1.c1, 1.1.1.c2 et 1.1.1.c3.

Il importe de ne pas perdre la suite de la représentation tant sur le plan notionnel que visuel. Bien que la méthode soit au stade expérimental, la réponse semble satisfaisante : les repères notionnels de l'étudiant s'étaient mis en place immédiatement, l'étudiant étant même capable de repérer une erreur dans la représentation visuelle faite au tableau. Bien qu'il soit trop tôt pour tirer des conclusions, il apparaît que la représentation arborescente peut être maintenue et le travail notionnel aussi. La transposition du mode visuel au mode alphanumérique maintient « l'aspect visuel ».

4. Aborder le Knowledge Management[6]

La gestion des connaissances fait partie des enseignements des écoles de management. Les raisons sont claires si l'on suit Ikujiro Nonaka pour qui à un moment de l'histoire de l'économie mondiale où la seule certitude est l'incertitude, la seule source durable d'avantage compétitif est la connaissance (1991). La part langagière des connaissances étant importante, cette « gestion » passe aussi par celle du langage qui sert à les représenter. Que ce soit à titre individuel ou de la communauté de travail, la gestion des connaissances passe

[6] L'expression *knowledge management* est, dans l'usage, beaucoup plus fréquente que celle de gestion des connaissances ou encore - selon les auteurs - de management des savoirs (Tarondeau, 1998). La dénomination *knowledge management* ou *KM* permet d'identifier très vite la thématique.

aussi par celle des termes, que nous considérons comme des noyaux d'information participant à ces connaissances. Il importe aussi de prendre en compte non seulement les discours des individus et des communautés de travail qui les réunissent mais aussi les organisations où ils travaillent.

C'est avec cette optique que nous avons créé un cours spécifique appelée « Aborder le *KM* » et proposé à Euromed-Management au niveau du master. Contrairement à Université Paris Diderot-EILA, la question à laquelle la représentation visuelle doit pouvoir répondre est : quelles connaissances faut-il mobiliser pour traiter d'un sujet donné ?

Trois exercices utilisent la représentation visuelle comme levier des connaissances. Le premier est celui de la représentation de l'organisation voire de l'entreprise. Le deuxième est celui d'une représentation arborescente des expressions (termes) dont la prise en compte est nécessaire pour rendre compte des connaissances à propos du sujet choisi par l'étudiant – et suivie d'une petite base de données. Le troisième est celui d'un travail en groupe sur un sujet imposé où le travail visuel apparaît comme une cartographie du sujet abordé.

Le premier exercice consiste à représenter et à justifier l'organisation d'une entreprise à partir du moment où, dans sa formation, un futur responsable doit connaître les différentes fonctions et leurs rapports. Les représentations en résultant sont très diverses et vont des organigrammes classiques pyramidaux ou plats à d'autres plus complexes voire très créatifs. Il importe de travailler sur le visuel comme moyen de mettre en évidence les secteurs de l'entreprise qui doivent communiquer entre eux ou le cas échéant partager et (co-) construire des connaissances de manière à pouvoir montrer un continuum communicationnel. Les figures 6 à 8 montrent des détails de quelques exemples de ce travail.

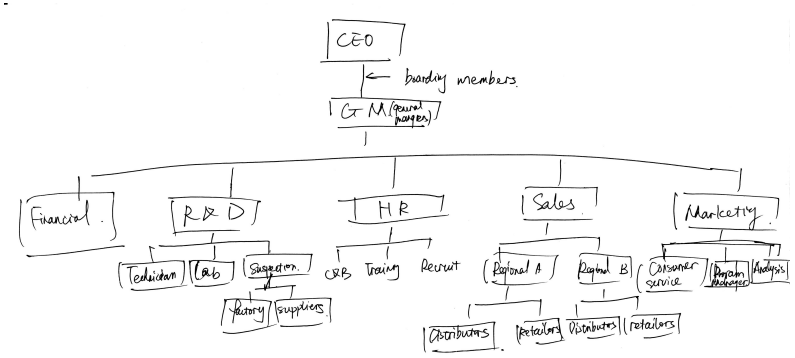


Fig. 6. Exemple d'organigramme « classique »

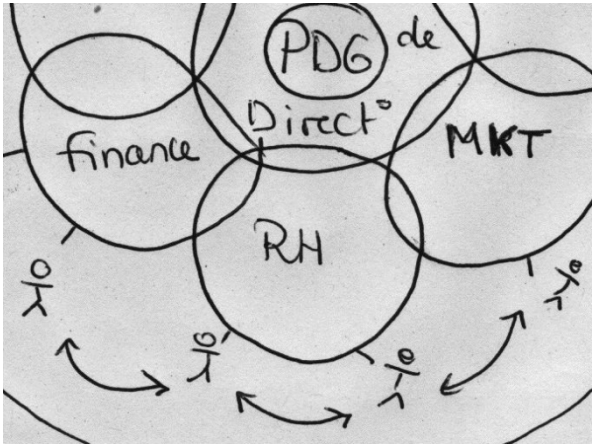


Fig. 6. Détail d'organigramme mettant en évidence les besoins de communication dans l'entreprise

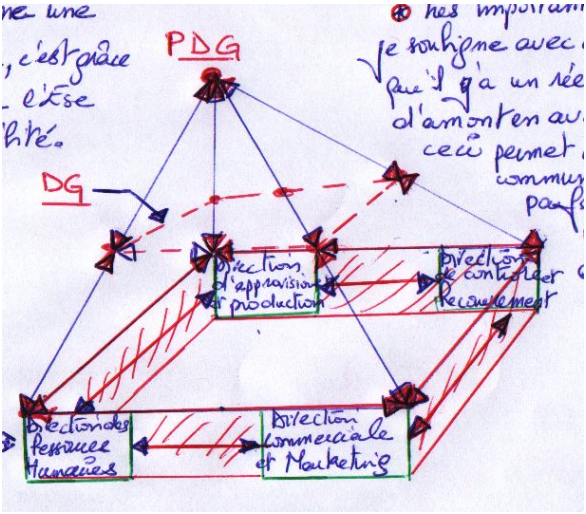


Fig. 7. Détail d'organigramme qui montre différentes « dimensions » nécessaires à la représentation.

Ce travail met en évidence d'une part, les connaissances acquises de manière théorique sur l'organisation et leurs liens, mais aussi telles qu'elles sont vues par l'étudiant en montrant *sa* vision de la réalité entrepreneuriale. Les liens entre secteurs sont formalisés par des lignes qui les relient ; leur dynamique souvent par des flèches. D'autre part, la représentation montre les communautés de travail qui, au moyen de leurs interactions, pourraient être amenées à partager des

connaissances. Il ne s'agit pas tant pour un groupe de savoir au même titre que les autres groupes de l'entreprise mais de posséder ce que Philippe Roqueplo appelle un « savoir décalé » : un savoir sur le savoir de l'autre (Roqueplo, 1990 : 75).

Le deuxième exercice est un travail sur un sujet au choix de l'étudiant ; les sujets sont très divers (le luxe, le football, les ressources humaines, la restauration, etc.). Comme c'est le cas à l'Université Paris Diderot–EILA, c'est d'abord un arbre de domaine qui est effectué et des sous-domaines signalés. Les expressions (termes) considérées comme des noyaux d'information nécessaires à la création des connaissances, apparaissent liées dans l'arborescence. Pour les termes qui sont considérés comme fondamentaux, l'étudiant fera des fiches inspirées des fiches terminologiques où les informations seront consignées (notamment la définition et la traduction dans une autre langue). Les raisons des liens entre termes dans l'arborescence ne sont pas mises en évidence ; dans ce cas, il n'est pas important de connaître la nature d'une relation mais de savoir qu'elle existe et de pouvoir la justifier tout comme les catégorisations qui sont opérées. En parallèle, c'est la lecture des textes d'introduction à la terminologie et aux ontologies qui présentent pour cette population un mode d'approche bien différent de l'acquisition des connaissances en entreprise, puis de la gestion de ces connaissances.

La figure 8 montre un exemple de ce travail sur les ressources humaines et la figure 9 un détail du même travail. Les relations entre les « noyaux porteurs d'information » - qui peuvent ne pas être des termes - sont évidentes (couleurs – rouge et gris dans l'original- ou grosseurs de lignes différentes). La présentation orale faite par ce type d'étudiants explicite ces rapports. C'est la documentation consultée et son propre cheminement intellectuel qui produisent l'arborescence ; sa capacité à en discourir, une preuve de la gestion de ses propres connaissances du sujet. Lorsque deux étudiants ayant travaillé sur le même sujet confrontent leurs représentations des divergences peuvent apparaître sur la représentation choisie mais rarement sur le contenu.

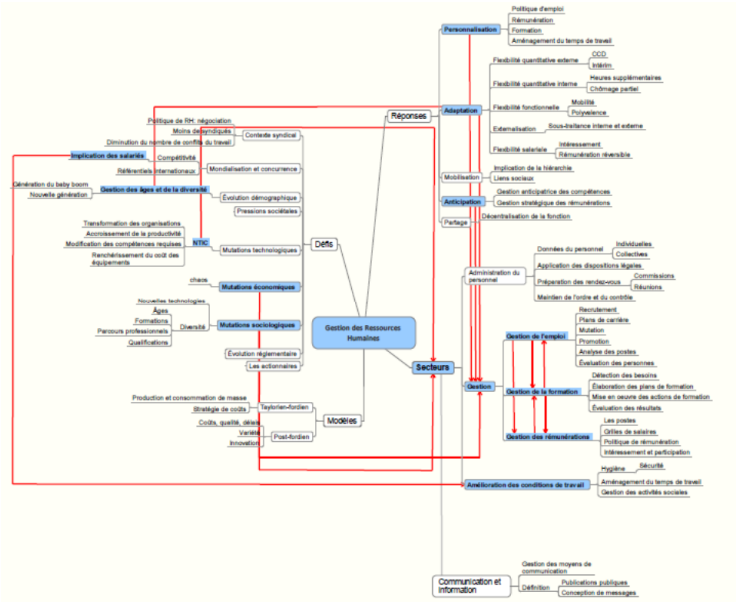


Fig. 7. Vue générale d'un travail sur les ressources humaines.

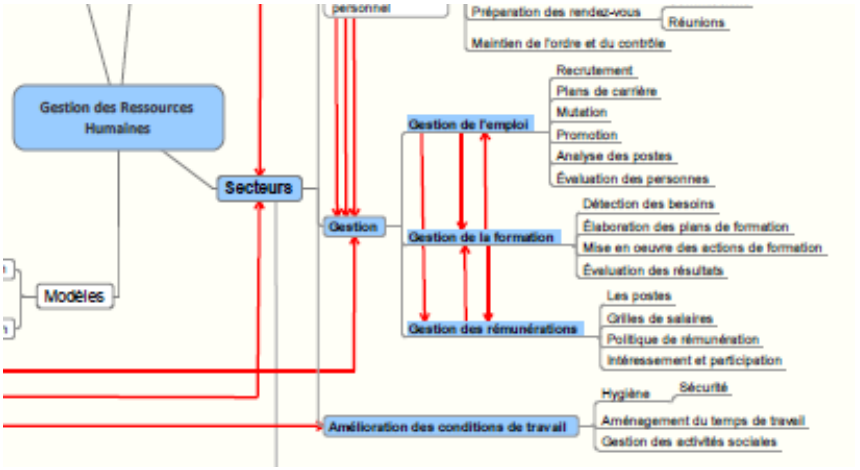


Fig. 8. Détail du même travail.

Le sujet du troisième exercice met les étudiants travaillant en groupe devant l'obligation de rendre compte des connaissances à mobiliser sur un sujet imposé pour permettre de prendre une décision dans le cas d'un investissement en entreprise. L'exercice doit permettre de répondre à la question : pour un

décideur, que faut-il savoir à propos d'un nouveau secteur d'activité en vue d'envisager des investissements ? Par exemple : les prothèses auditives (Fig. 9).

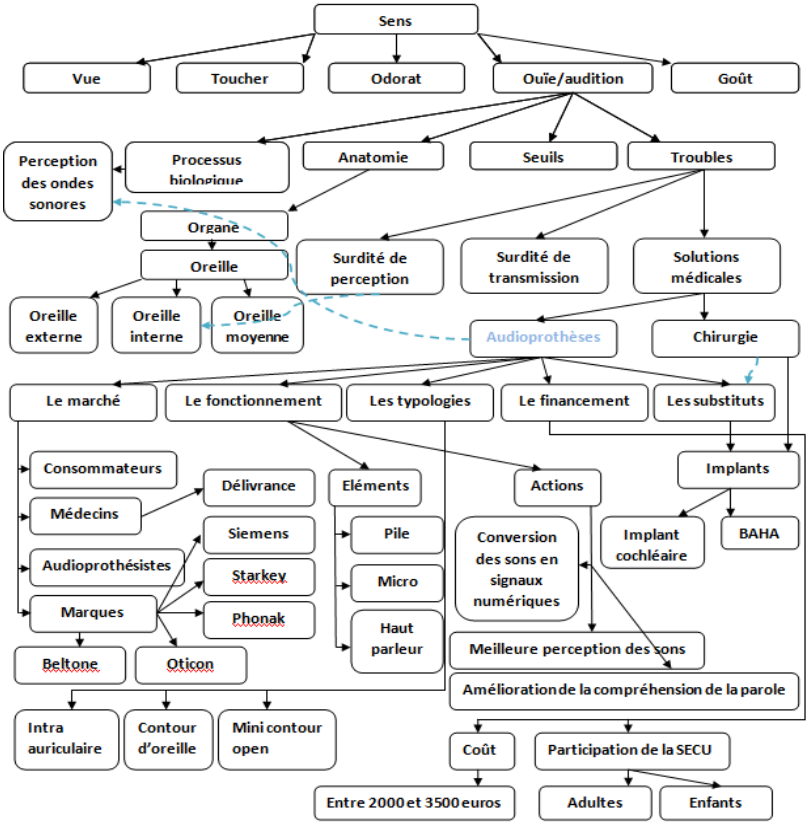


Fig. 9. Connaissances à mobiliser à propos des prothèses auditives.

Le travail présenté part des sens pour explorer, à partir de l'ouïe, différents sous-domaines : troubles, processus, solutions, le marché, les marques, etc. Ces sous-domaines sont reliés entre eux de manière à suivre un « parcours » qui est lui-même une forme de connaissance. Le rendu final comporte une présentation orale, un rapport écrit (pouvant être destiné à un conseil d'administration) et d'un glossaire des termes techniques qui facilite la lecture.

Les trois exercices fondés sur la représentation visuelle permettent de rendre compte d'un cheminement cognitif ; les liens sont fondés sur un mode de raisonnement qui ne part pas d'une démarche terminologique ou de cartographie sémantique (Tricot, 2000). Ils constituent à ce titre une manière concrète de

« gérer des connaissances », comme la « démarche du *KM* » (Bouvard et Storehaye, 2002) le suppose. Tout comme pour les cartes ou les arborescences terminologiques, l'étudiant se trouve en face d'une cartographie de « ce qu'il sait » et dont il a été l'acteur. Il peut donc gérer ses connaissances et les partager ou les transmettre.

5. L'application « économique » du support visuel

La transmission des connaissances et de savoirs-faire [7] est un enjeu majeur des entreprises d'aujourd'hui. C'est un sujet de recherche à part entière (Boutte, 2007 ; Barès, 2007 ; Delbos & Jorion, 1984 ; Bril *et al.*, 2002). Lorsque les personnels quittent leurs organisations ou entreprises, leurs connaissances partent avec eux. La documentation ne peut pas tout faire, à condition qu'elle ait été constituée.

Les terminologies et les ontologies sont des démarches dont l'apport cognitif est inestimable, mais qu'en dit l'utilisateur final ? Est-ce que lire une définition, sa traduction, des remarques techniques voire linguistiques ou les liens à d'autres termes suffisent à reconstruire les connaissances telles qu'elles se construisent chez l'individu ? Les ontologies sont dans le même cas. Dans un cas comme dans l'autre, force est de constater que leur mise en place est longue - et coûteuse - tant la préparation est laborieuse et les moyens mis en œuvre prenants. Le portefeuille de connaissances stratégiques de l'entreprise (grande ou petite, avec ou sans salariés) d'où elle tire sa valeur ajoutée peut se voir concentré dans un groupe restreint de personnes. Les connaissances dont ces entreprises sont dépositaires sont donc fondamentales.

Si l'on croit les fiches de *L'Expansion*, les cent entreprises qui font le plus gros chiffre d'affaires en France en 2008[8], tous secteurs confondus, sont des entreprises avec un grand nombre de salariés. Or, pour la même année, l'INSEE[9] nous apprend que sur plus de 3,5M d'entreprises, seulement 0,049% dépassent les 500 employés (voir fig. 9, graphique en secteurs, un mode représentation des connaissances).

[7] Le pluriel avec s est de plus en plus utilisé.

[8] *http://www.lexpansion.com/economie/classement/atlas.asp?idc=124715*, consulté en février 2010.

[9] http://www.insee.fr/fr/themes/, consulté en février 2010.

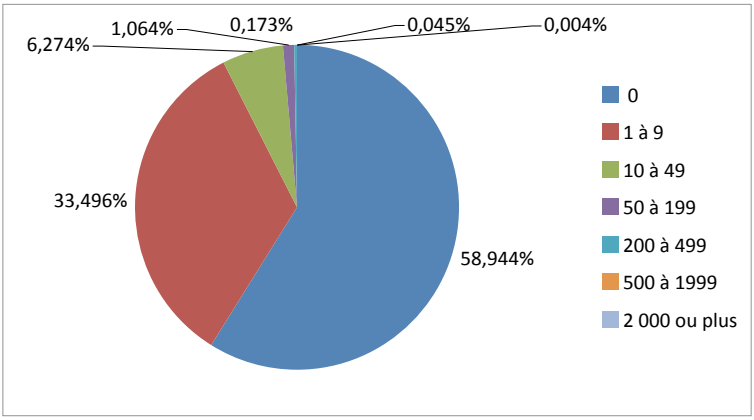


Fig.10. Taille d'établissements, en nombre de salariés en France en 2008.

Les entreprises qui ne comportent aucun salarié ou de un à 9 salariés représentent 92,44% des entreprises en France. Combien parmi elles peuvent financer une ontologie ou une terminologie adaptée à leurs besoins ? Et pourtant, si les connaissances sont une valeur stratégique et font partie des actifs immatériels des entreprises, leur sauvegarde et leur transmission sont fondamentales.

Prenons un exemple. Le métier de paludier était transmis de père en fils jusqu'à 1978 sans autre support que le geste accompagné de la parole. L'activité dispose depuis d'un Brevet professionnel de responsable d'exploitation agricole, option saliculture. Le métier de paludier exprime ses connaissances avec une terminologie particulière[10] indispensable à l'activité. Ce savoir-faire s'exporte et aujourd'hui les chiffres de l'activité économique montrent une progression non négligeable pour un métier si peu connu. En effet, le chiffre d'affaires de la marque Sel de Guérande Le Guérandais est passé de 6 à 31 MF de 1993 à 1997 (Lewi, 2004 : 8) pour atteindre 14,3M € en 2009 (93,8MF). C'est cependant une coopérative à laquelle 185 paludiers adhèrent qui le réalise et non chacun des paludiers en tant qu'entrepreneur indépendant. Combien d'artisans auraient les moyens de faire rédiger un vocabulaire dans les règles de l'art ? Combien d'autres métiers se trouvent dans cette situation ? L'artisanat est au premier rang des activités où les terminologies occupent une place de choix, sans elles il est impossible de transmettre ou partager des connaissances.

[10] http://www.salinesdeguerande.com/index.php?id=80, consulté en février 2010.

6. Conclusions

Les termes sont au centre du travail effectué dans deux institutions aux objectifs différents. Dans un cas comme dans l'autre, le travail est d'abord monolingue ; la traduction étant une étape ultérieure. D'ailleurs, si la terminologie peut être utile à la traduction, elle n'en dépend pas. La traduction peut apporter des informations au travail terminologique, lequel n'en est nullement tributaire. Dans un cas comme dans l'autre, le travail sur des notions permet d'aboutir à une représentation visuelle des connaissances.

Cela posé, deux perspectives semblent se dégager. La première est un travail *sur* les termes c'est-à-dire une terminologie orientée sur la langue spécialisée et le discours produit, que ce soit pour des raisons mono-, bi- ou multilingues. Ici la relation entre les termes est primordiale.

La deuxième perspective est un travail *avec* des termes et orienté sur des notions à mobiliser et indépendant du traitement informatique. L'informatique n'est pas un prérequis au travail terminologique pour représenter « ce que l'on sait » ou « ce que l'on a besoin de savoir ». Le travail *avec* les termes en tant qu'unités indispensables aux connaissances mobilisées n'entraîne pas à notre avis un travail sur les relations mais en revanche sur les catégorisations où se placent ces termes. Pour reprendre l'exemple des prothèses auditives, il importe de *savoir* que deux formes de « surdité » peuvent être catégorisées : « surdité de réception » et « surdité de transmission ». Le rapport hyperonyme-hyponyme n'est important que pour le terminologue ou l'ontologue mais non pour le spécialiste. L'audioprothésiste – et ce n'est qu'un exemple - se trouve au carrefour de plusieurs disciplines dont il a ce savoir décalé à propos duquel écrit Roqueplo. C'est bien pour cette raison que l'arborescence montre autant « le marché » que « le financement » comme catégories à traiter et à connaître. Le travail sur la représentation visuelle montre ce panorama que la terminologie aboutissant à un travail terminographique n'a pas comme objectif primordial, l'arborescence n'étant pas d'ailleurs ni obligatoire ni systématique. Et pourtant, il s'agit *aussi* d'une représentation des connaissances.

Le travail terminologique orienté vers la gestion des connaissances montre aussi qu'il n'est pas possible de dissocier la diversité des acteurs concernés des connaissances qu'ils mettent en commun pour effectuer un travail. La représentation visuelle des connaissances à mobiliser pour rendre compte d'un domaine montre que les terminologies des uns ne sont pas étanches et entrent en rapport avec des terminologies utilisées par les autres. L'interdisciplinarité arrive à point nommé.

La « visualisation des connaissances » permet d'établir des liens entre notions que les modes opératoires de la terminologie ou des ontologies ne considèrent pas parce que ce n'est pas leur objectif. Le mode visuel, y compris adapté aux aveugles et malvoyants, offrirait des perspectives applicatives que la terminologie peut bien enrichir. Il laisse aussi la possibilité d'intégrer dans le mode visuel d'autres aspects que les utilisateurs au travail considèrent indispensables et qui pourraient être oubliés.

Bibliographie

Barès, Michel, (2007) : *Maîtrise du savoir et efficience de l'action*, Paris, L'Harmattan

Bord, Jean-Pierre, (1995) : *« La carte comme manipulation »* in La cartographie en débat, Paris Khratala-Orstom, pp. 57-82

Boutte, Jean-Louis, (2007) : *Transmission du savoir faire*, Paris, L'Harmattan

Bril, Blandine et Roux, Valentine, (2002) : *Le geste technique*, in TIP, Volume XIV, n°2

Bouvard, Patrick et Storehaye, Patrick, (2002) : *Knowledge management*, Colombelles, EMS

Cambrézy, Luc (1995) : *« De l'information géographique à la représentation cartographique »* in La cartographie en débat, Paris Khratala-Orstom, pp. 129-148

Delbos, Geneviève et Jorion, Paul, (1984) : *La transmission des savoirs*, Paris, Éds. MSH

Felber, Helmut (1987): *Manuel de terminologie*, Paris, Unesco

Peirce, Charles Sanders, (1978) : *Ecrits sur le signe*, Paris, Seuil

Kockaert, Hendrik, Bassey Antia, (2008) : *Comment modéliser des concepts en rapprochant un langage orienté objet et deux normes terminologiques orientées concept*, in Terminologie et ontologies : théories et applications. Actes de la conférence TOTh 2008, Roche, Christophe [ed.], Annecy, Institut Porphyre, Savoir et Connaissance, 105-125

Lewi, Georges, (2004) : *La marque*, Paris, Vuibert, 3e éd.

Nonaka, Ikujiro, (1991) : « The Knowledge-Creating Company » in *Harvard Business Review*, Nov-Dec 1991

Roqueplo, Philippe (1990) : *« Le savoir décalé »* in Technologies et symboliques de la Communication, Grenoble, PUG, pp. 75-79

Tarondeau, Jean-Claude, (1998) : *Le management des savoirs*, Paris, PUF, QSJ n°3407

Tricot, Chistophe, (2006) : *Cartographie sémantique. Des connaissances à la carte*, Thèse de doctorat, Université de Savoie

À propos de l'auteur

de Vecchi, Dardo

Professeur associé, Euromed-Management

Membre du laboratoire CLILLAC, Université Paris Diderot - EILA

D. de Vecchi développe une linguistique appliquée au management. Il enseigne la terminologie à Université Paris Diderot - EILA et est professeur invité à l'École Centrale de Paris et à l'Université Paris Descartes. Il est membre de l'équipe Condillac.

Euromed-Management

BP 921 F-13288 Marseille

dardo.devecchi@euromed-management.com

devecchi@eila.univ-paris-diderot.fr

Une « ontoterminologie » pour les interprètes de conférence

Elisa Veronesi, Franco Bertaccini

Résumé : La terminologie et la création d'ontologies s'avèrent de plus en plus utiles pour tous ceux qui travaillent dans les domaines de la traduction et de l'interprétation. Ce projet a donc pour but d'étudier l'usage d'arbres ontologiques en tant que support à l'interprétation simultanée. Nous avons essayé de mettre la terminologie au service des interprètes en concevant une ontologie dynamique et modulaire pouvant être utilisée par les interprètes directement en cabine. Pour ce faire, notre projet s'est basé sur la création d'une conceptualisation formelle d'un domaine spécialisé, développée à partir de ses termes et concepts fondamentaux. La méthodologie de sa structuration repose donc sur la façon dont les termes (et les concepts auxquels ils font référence) sont reliés entre eux d'un point de vue sémantique et logique. Pour cette raison, ce support pourrait être défini comme une « ontoterminologie ». Le domaine conceptualisé est celui du SIDA. En particulier, notre analyse a porté sur l'étude des manifestations cutanéo-muqueuses se présentant au cours de l'infection par le VIH. Nous avons créé deux corpus de langue spécialisée, un en français et un en italien, desquels nous avons extrait les concepts et les termes fondamentaux du domaine et du sous-domaine, ainsi que les relations sémantiques les reliant. Ensuite, nous avons développé une ontoterminologie bilingue italien-français très spécialisée et strictement liée au domaine. Le champ de la biomédecine étant très complexe et articulé, les instances, les relations, ainsi que la structure des diagrammes ont été définies manuellement, en utilisant le logiciel graphique Visio 2007. On a essayé, d'une part, de produire une ontologie flexible, synthétique mais exhaustive ; de l'autre, de faire en sorte qu'elle permette de prévoir (ou de suivre) l'évolution des discours de la conférence, en affichant non seulement les termes et les équivalents qui relèvent du sujet, mais aussi les relations sémantiques les relient. Enfin, nous avons mené un test empirique pour vérifier la « lisibilité » et l'utilité effective du support avec des interprètes professionnels. L'analyse de leurs performances et opinions nous a permis d'évaluer les avantages concrets apportés par l'outil réalisé.

Mots-clés : Terminologie, ontologie, ontoterminologie, interprétation de conférence, modulaire, multi-niveaux, relations sémantiques, corpus, ontologie régionale, ontologie de domaine, SIDA, VIH, pathologies cutanéo-muqueuses, terminologie médicale, médecine.

1. Introduction

Dans un monde de plus en plus globalisé et multiculturel, la terminologie et la structuration du savoir par les ontologies sont deux disciplines essentielles pour le développement de nombreux domaines scientifiques et industriels, ainsi que

pour les recherches relatives aux langues et à la linguistique. En parallèle, la traduction et l'interprétation revêtent une importance majeure pour ce qui est, d'une part, de la préservation du multilinguisme en lui-même, et, de l'autre, de la croissance et de la continuation de toute activité industrielle, commerciale, sociale et de recherche. Interprètes et traducteurs ont donc largement recours à la terminologie et aux ontologies pour développer des bases de données exhaustives, modulaires et interactives, représentant tout domaine de connaissance par des critères sémantiques, afin de relier chaque concept au terme correspondant en toute langue ou bien langage de spécialité. Cependant, le foisonnement multilingue et multiculturel international a atteint un stade où une progression qualitative devient urgente. Étant donné que le catalogage informatique des données multilingues évolue assez rapidement et que l'on peut désormais compter sur des logiciels et des langages de programmation de pointe, l'urgent est de se pencher sur les outils de structuration et référencement des contenus multilingues. Bref, il devient primordial de pouvoir créer des réseaux sémantiques multilingues qui soient mieux articulés et plus flexibles que ceux que l'on peut développer aujourd'hui, afin de produire des structurations du savoir qui soient concrètement utiles pour l'apprentissage et le travail.

Ce projet se situe précisément à ce point de la recherche en interprétation, et a pour but d'étudier l'usage d'arbres ontologiques complexes en tant que support à l'interprétation simultanée. Nous avons essayé de mettre la terminologie au service des interprètes en concevant une ontologie dynamique et modulaire pouvant être utilisée par les interprètes en cabine, durant leur travail. Pour ce faire, notre projet s'est basé sur la création d'une conceptualisation formelle d'un domaine spécialisé (une ontologie), développée à partir de ses termes et concepts fondamentaux. La méthodologie de sa structuration repose par conséquent sur la façon dont les termes (et les concepts auxquels ils font référence) sont reliés entre eux d'un point de vue sémantique et logique. Pour cette raison, ce support pourrait être défini comme une « ontoterminologie ». Nous allons maintenant décrire plus en détail comment notre projet a été structuré et quelles approches ont été suivies pour son développement.

2. Terminologie et ontologies

Aujourd'hui, la terminologie et la création d'ontologies sont devenues deux facteurs clés pour le repérage, la catégorisation et la structuration des savoirs et des connaissances, un sujet très important pour les linguistes, les traducteurs et les interprètes. En particulier, la terminologie joue un rôle primordial, puisqu'elle permet à la fois d'encadrer les connaissances dans des bases de données structurées et plus ou moins « standardisées » et de partager ces connaissances

avec d'autres utilisateurs/experts du domaine. Elle se situe donc au carrefour de plusieurs disciplines. Le terme lui-même est donc un élément « clé », parce qu'il véhicule non seulement les concepts et les idées fondamentales que l'émetteur souhaite transmettre, mais aussi la structure du système notionnel du domaine à l'intérieur duquel il est utilisé. Le système notionnel d'une terminologie est donc défini par un ensemble de concepts qui constitue, pour une communauté donnée, une « description du réel ». Cela nous amène à la notion d'ontologie, pratique strictement liée à celle de la terminologie.

La notion d'ontologie est devenue un élément clé dans toute une gamme d'applications faisant appel à des connaissances. Ces connaissances sont modélisées de manière structurée, et « exprimées à l'aide de langages ayant une sémantique bien définie » (Charlet, 2002 : 3). Plus en détail: « Les ontologies sont des systèmes conceptuels destinés à fournir les notions élémentaires à la formulation des connaissances dont on dispose sur un sujet donné » (Bachimont, 2004 : 128), et exprimés à l'aide de langages opérationnels standards. Une ontologie organise donc les concepts et les relations pertinents et consensuels dans un domaine pour une application donnée. Par conséquent, une ontologie est, en d'autres mots, la conceptualisation des objets reconnus comme existant dans un domaine, de leurs propriétés et des relations les reliant. Comme en terminologie, la signification des termes repose sur une conceptualisation du monde et un vocabulaire de termes pour en parler. L'ontologie vise, *ipso facto*, les mêmes propriétés de consensus, de cohérence, de partage et de possibilité de réutilisation que la terminologie.

2.1. Un support pour l'interprétation simultanée

Dans ce contexte, la terminologie et la construction d'ontologies peuvent être d'une certaine manière très utiles pour la traduction et l'interprétation, aussi. L'énorme contribution que la terminologie apporte au domaine de la traduction spécialisée (en termes de création et consultation de bases de données spécifiques uni- ou multilingues et d'aide au choix des mots/termes les plus adéquats pour un certain type de texte ou registre) est bien connue, tandis que l'utilisation d'ontologies de la part des traducteurs ou de terminologues est en train de se frayer un chemin. Au contraire, l'usage de banques terminologiques et d'ontologies en interprétation de conférence est encore assez négligé ou méconnu.

Pourtant, lors de la préparation en vue d'une conférence, les bases de données terminologiques seraient bien entendu très utiles pour les interprètes, surtout s'ils ne doivent pas les produire eux-mêmes, puisque la création de fiches terminologiques est un procédé lent et complexe et les interprètes n'ont souvent

que quelques jours pour étudier le sujet de la conférence. Mais, une fois en cabine, il est encore très difficile d'utiliser un ordinateur portable au moment de l'interprétation simultanée et de consulter des fiches, puisque, en principe, l'orateur parle trop vite pour que l'interprète puisse rechercher le terme souhaité. Des recherches sont en cours pour comprendre comment fournir aux interprètes de simultanée un support informatique qui puisse les aider dans l'accomplissement de leur tâche. Les glossaires traditionnels sont difficiles à consulter lors de l'interprétation, puisque l'interprète devrait les feuilleter au moment même où il parle et, à cause de l'effort produit pour cette recherche, il risquerait de commettre des erreurs ou de perdre des informations durant la phase d'écoute ou de reformulation. Par contre, l'ordinateur portable en cabine pourrait résoudre le problème, dans la mesure où il permettrait une consultation rapide d'un dictionnaire ou d'une base de données terminologique grâce à un logiciel spécifique. La voie dans la direction de l'utilisation d'outils numériques en tant que support aux interprètes de conférence semble donc ouverte.

2.2. Une « ontoterminologie » pour interprètes

Ce travail a donc eu pour but l'étude d'un outil linguistique gérable avec un ordinateur portable comme support à l'interprétation simultanée. Nous avons analysé la possibilité de mettre la terminologie au service des interprètes à l'aide d'ontologies, qui – tout comme les fiches terminologiques – représentent la conceptualisation d'un domaine, mais en soulignant les relations sémantiques et logiques qui lient les termes à travers des arbres et/ou des graphes. Le type d'ontologie nécessaire étant destiné spécifiquement à des interprètes de conférence et, par conséquent, la création d'arbres modulaires et très spécialisés s'est imposée. Nous avions besoin en fait de terminologies et d'ontologies bilingues – ou même trilingues – liées spécifiquement au sujet de la conférence en préparation. Nous avons donc décidé de modéliser un domaine de connaissance particulier et restreint par le truchement d'une ontologie régionale et structurée *ad hoc* répondant aux besoins des interprètes de conférence, puisque un support linguistique de ce type pourrait constituer une aide précieuse pour le travail en cabine.

Pour la structuration et le développement de notre projet nous nous sommes basés en particulier sur les travaux menés en France par le professeur Christophe Roche et l'Équipe Condillac du Laboratoire Listic, ainsi que sur les projets développés en Italie par le professeur Franco Bertaccini au sein du laboratoire de Terminologie de l'Université pour Interprètes et Traducteurs de Forlì (SSLMIT). Leurs études portent en fait sur la nécessité de fusionner les disciplines de la terminologie et de l'ontologie pour pouvoir développer une approche que l'on

peut qualifier de « ontoterminologique »[1]. Cette approche repose sur l'idée qu'à l'heure actuelle il est désormais nécessaire de replacer le concept et sa dénomination au centre de la terminologie, tout en préservant sa dimension sociolinguistique par la prise en compte des termes d'usage à travers la langue de spécialité. Si employée en tant que critère de construction pour la réalisation d'ontologies (schémas conceptuels) relatives à un domaine donné, cette méthodologie permet de conceptualiser de façon explicite ce champ de connaissance et de rendre formelles les informations collectées et schématisées – qui sont donc manipulables, modifiables et partageables par un ordinateur. Pour cela, nous avons « emprunté » le nom créé par Christophe Roche, et appelé notre prototype une « ontoterminologie ».

De surcroît, on a également pris en compte les principes de la « terminologie conceptuelle »[2], une approche qui préconise l'élaboration d'une terminologie spécifique à un champ de connaissance sur la base de l'analyse des concepts auxquels les termes font référence et qui aboutit à une systématisation de ceux-ci en structures cohérentes telles que des arbres et des graphes conceptuels. De telles structures permettent de représenter graphiquement les domaines et les sous-domaines d'application de certaines connaissances, ainsi que les relations sémantiques qui relient les concepts qui en sont à la base. Ces schémas sont modélisés à partir des requêtes spécifiques des commanditaires et/ou des usagers finaux du catalogage terminologique réalisé. Par conséquent, si la conceptualisation ainsi créée est bien en langue naturelle, elle est également définie dans un langage formel selon des principes épistémologiques. Une conceptualisation terminologique ainsi définie donne, de ce fait, la structure de base pour une ontologie qualifiée de « régionale » et, donc de « domaine », qui ne cherche pas à modéliser le monde en général, comme le font les ontologies de type « *top level* », mais un contexte d'usage bien spécifique dans un domaine ciblé.

En effet, les interprètes requièrent un support qui leur permette non seulement de retrouver des termes et leur équivalent, mais aussi de les placer dans le contexte spécifique de la conférence grâce à la représentation graphique des relations entre les termes et les concepts. Cela signifie que la sélection et la

---

[1] En employant la dénomination créée par C. Roche dans son article *Le terme et le concept : fondements d'une ontoterminologie*, TOTh, 2007.

[2] Cette approche a été en effet récemment suivie également par l'ISO, qui a employé les principes de la terminologie conceptuelle pour les règles, les définitions et la structure de la nouvelle norme ISO 704: 2000. Plus de détails concernant les canons de travail relatifs à la terminologie ainsi dite "conceptuelle" au sein de la norme ISO 704: 2000 peuvent être trouvés dans l'article de Wright "The Once and Future ISO 704", publié dans le magazine de secteur *eDITion* n. 1/2007.

structuration de la terminologie à la base de ce type d'ontologie devraient être étudiées attentivement et préparées *ad hoc* sur la base du sujet de la conférence et de la combinaison linguistique demandée à l'interprète. Et, question qui ne peut être négligée, l'ontologie elle-même devrait permettre à l'interprète de prévoir (ou de suivre) le(s) discour(s) en affichant non seulement tous les termes et les équivalents qui relèvent du sujet abordé, mais aussi les relations sémantiques et logiques qui les relient.

2.3. Les objectifs à atteindre par la création du prototype

Nous nous sommes donc attachés à modéliser une ontologie régionale schématisant la terminologie relative à un domaine scientifique spécifique qui se base sur le primat des concepts et des relations sémantico-logiques les reliant. Nos objectifs étaient les suivants :

- Produire une ontoterminologie bilingue italien–français interactive, dynamique, modulaire, taillée sur le domaine pour lequel elle a été conçue, contextuelle, répondant aux besoins spécifiques des interprètes de conférence.
- Développer le prototype d'un outil concis et exhaustif en même temps, qui soit intuitif et facile à utiliser non seulement lors de la préparation au travail, mais aussi au cours de la conférence-même, en simultanée.
- Produire des schémas ontologiques présentant une hiérarchie des concepts du domaine médical analysé basée sur :
  - les relations hiérarchiques verticales d'identité (*is_a*) et d'appartenance (*part_of*) ;
  - les relations non hiérarchiques causales et temporelles, qui sont d'importance majeure pour la conceptualisation des domaines de médecine et de biomédecine.
- Faire en sorte que les diagrammes ontologiques permettent à l'interprète de « prévoir » ou, au moins, suivre l'évolution du discours, pour que le contexte d'usage des termes fournisse à l'interprète des points de repère.
- En parallèle, permettre à l'usager de créer différents parcours conceptuels de consultation des fiches basés sur différents « points de départ » (selon le concept ou le terme duquel on commence à « lire » les arbres ontologiques).

3. Les phases concrètes de réalisation du projet

Pour réaliser notre projet, nous avons mené une étude empirique portant sur la création d'une ontologie spécifique à un domaine scientifique. Les conférences scientifiques, en effet, se prêtent davantage à être « prévues » (ou analysées, si l'interprète a la chance d'avoir les textes à interpréter en avance) et représentées de façon graphique et schématique. La raison en est que ces types de discours ou de relations utilisent en principe un langage très spécialisé (et qui, donc, ne devrait pas poser de problème d'interprétation au niveau des métaphores ou de la polysémie) et sont d'ordinaire assez rigoureusement structurés. Ainsi, il devrait être possible de représenter la démarche logique et sémantique d'un texte par une ontologie modulaire basée sur un corpus terminologique créé *ad hoc* sur la base de textes scientifiques accrédités (le support d'un expert du domaine s'impose donc).

Le domaine choisi est le champ médical, qui est particulièrement approprié pour notre étude aussi bien pour ses caractéristiques intrinsèques (l'existence d'une langue très spécialisée dont les termes ont, dans la majorité des cas, un équivalent dans les autres langues étrangères ; la facilité de repérer les documents à partir desquels on peut extraire une terminologie bilingue) que pour ses caractéristiques extrinsèques (beaucoup de conférences internationales consacrées à ce secteur sont organisées chaque année ; par conséquent, comme nombre d'interprètes choisissent de se spécialiser en médecine, ce travail pourrait s'adresser à un groupe de professionnels assez vaste). En particulier, l'étude a porté sur le VIH et, plus spécifiquement encore, sur les symptômes dermatologiques liés à l'apparition de cette pathologie. Ce choix a été fait de sorte que notre « simulation » se rapproche le plus possible de la réalité des conférences, où, généralement, les experts et les professionnels se focalisent sur un sujet très spécifique.

Les langues des corpus terminologiques ont été le français et l'italien, soit deux de nos langues de travail, pour lesquelles il est assez facile de repérer du matériel dans les bibliothèques ou sur la toile. L'ontologie a donc été produite pour une interprétation du français en italien.

3.1. L'étude du domaine et la création des corpus de référence

Tout d'abord, nous avons donc décortiqué le sujet du SIDA et du VIH pour mieux nous orienter dans la recherche et comprendre les textes lors de la phase d'extraction terminologique. Ci-après, un exemple des graphes réalisés en italien pour schématiser le domaine à représenter dans notre ontoterminologie.

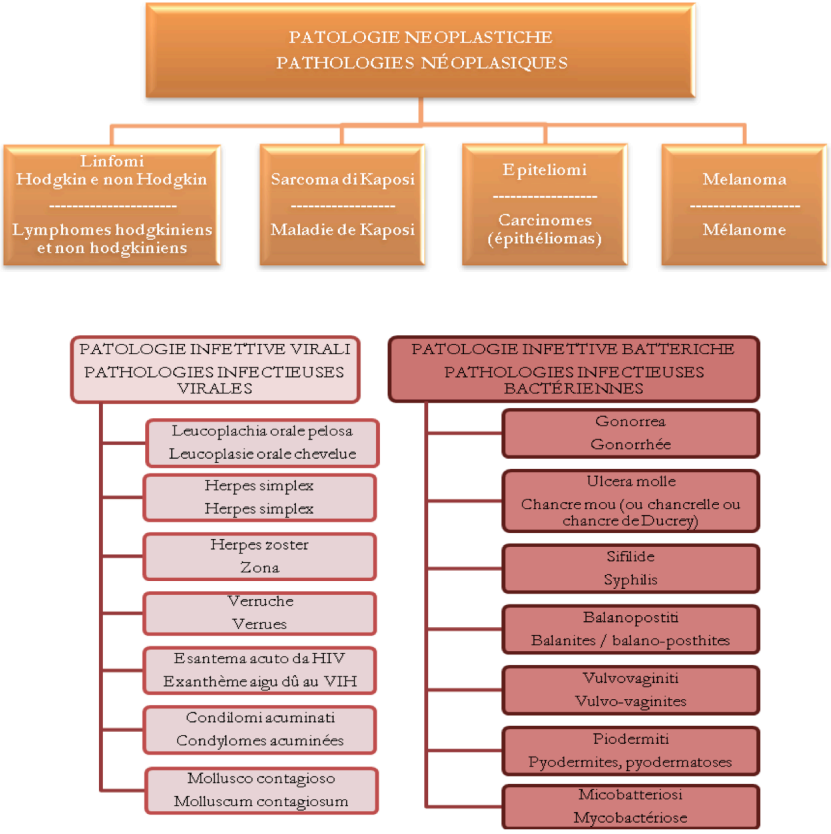
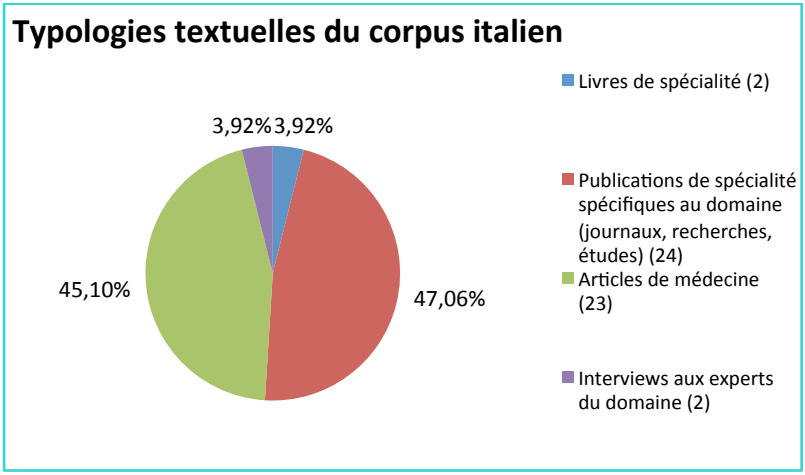


FIG. 1 – pathologies néoplasiques, pathologies infectieuses virales et bactériennes dues à l'infection par le VIH.

Ensuite, il a été nécessaire de repérer un ou plusieurs experts du secteur. Leur expertise a été essentielle pour choisir les documents pour la création des corpus terminologiques et la rédaction d'un discours *ad hoc* pour tester l'ontologie produite durant l'interprétation simultanée. Les deux corpus créés ont été composés comme suit :



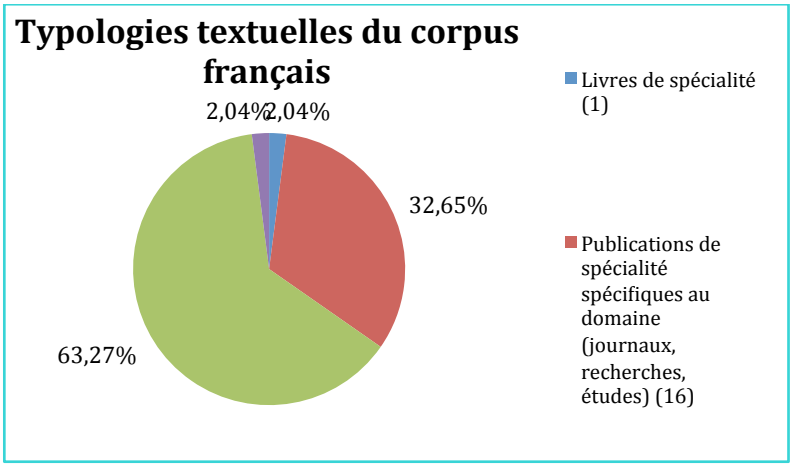


FIG. 2 – Diagrammes présentant les typologies textuelles des documents composant les corpus italien et français

Les corpus sont homogènes en termes de fiabilité, niveau de spécificité et destinataire (experts ou semi-experts) ; mais ils ne le sont pas complètement en termes de typologie textuelle. Ils incluent en effet des livres, des articles et des publications spécialisés sur papier, ainsi que des recherches, des articles et des études fouillés spécifiques publiés sur des sites Internet de médecine.

a) L'extraction des candidats termes et l'identification des termes et des relations pour l'ontoterminologie

Nous sommes ensuite passés à la phase d'extraction terminologique bilingue, pour laquelle nous avons utilisé Wordsmith, un logiciel qui permet d'identifier les

candidats termes d'un domaine de spécialité à travers une comparaison statistique de la fréquence d'apparition des mêmes termes à l'intérieur d'un corpus de spécialité et d'un corpus de type générique. Les deux corpus ont les dimensions suivantes :

- Corpus de langue de spécialité en italien: 551.257 mots
- Corpus de référence en italien: autour de 2 millions de mots
- Corpus de langue de spécialité en français: 552.079 mots
- Corpus de référence en français: autour de 2 millions de mots

L'analyse à travers des paramètres configurés par nous-mêmes au sein de Wordsmisth nous a permis d'extraire une longue liste de candidats termes simples et composés (syntagmes et relations sémantiques) dans les deux langues. Nous avons extraits des listes de un à cinq termes regroupés (*clusters*), formant des termes composés ou des syntagmes. L'extraction a affiché une correspondance intéressante entre les positions des candidats dans les listes italiennes et françaises: les termes définissant le même concept sont très souvent placés à des positions similaires dans le classement. Cet aspect nous a aidés lors de la recherche des équivalents interlinguistiques et démontre qu'il est tout à fait possible de schématiser un domaine scientifique et suffisamment spécifique par une ontologie bilingue, puisqu'en principe les mêmes concepts sont présents dans les deux langues. Comme on peut le voir dans le tableau qui suit, très souvent les candidats termes ont une position et une fréquence très similaires.

N	WORD	FREQ.		N	WORD	FREQ.
1	HIV	4.659		1	VIH	4.036
2	INFEZIONE	2.537		2	TRAITEMENT	3.613
3	PAZIENTI	2.370		3	CHEZ	2.598
4	TERAPIA	2.482		4	ÊTRE	2.277
5	VIRUS	1.805		5	PATIENTS	1.992
6	AIDS	1.570		6	INFECTION	1.888
7	FARMACI	1.605		7	PEUT	1.552
8	ANTIRETROVIRALE	1.293		8	SIDA	1.352
9	INFEZIONI	1.225		9	VIRUS	1.441
10	CELLULE	1.071		10	PERSONNES	1.300
11	LESIONI	940		11	ÉTÉ	1.232
12	RISCHIO	1.330		12	LÉSIONS	1.035
13	VIRALE	865		13	PATIENT	1.249
14	SOGGETTI	1.202		14	VIRALE	898
15	HAART	821		15	INFECTIONS	875

16	PAZIENTE	756		16	CHARGE	1.343
17	LINFOCITI	764		17	RISQUE	1.279
18	CASI	1.391		18	CAUSE	1.856
19	PATOLOGIA	1.067		19	PEUVENT	720
20	TRATTAMENTO	1.199		20	CELLULES	642
21	DIAGNOSI	712		21	PEAU	639
22	CUTE	575		22	IMMUNITAIRE	589
23	CUTANEE	2.010		23	THÉRAPEUTIQUE	577
24	KAPOSI	430		24	TRAITEMENTS	734
25	MALATTIA	415		25	PORTE	7.670
26	OPPORTUNISTICHE	411		26	CUTANÉES	638
27	SARCOMA	410		27	AYANT	544
28	TOSSICITÀ	439		28	LYMPHOCYTES	535
29	CLINICA	515		29	PRÉVENTION	511
30	CAUSA	419		30	CAS	1.815

FIG. 3 – Les 30 premiers résultats de l'extraction des candidats termes simples (unigrammes) par Wordsmith, en italien et en français. Le montant total de candidats extraits s'élève à 500 termes (simples, complexes et composés)

b) L'identification des relations sémantiques fondamentales

Les relations essentielles pour la conceptualisation ont été repérées lors de nos entretiens avec les experts, ainsi que par l'extraction des termes des corpus de langue de spécialité. L'ontoterminologie inclut:

- Des relations hiérarchiques, et en particulier des relations génériques (surtout la relation de subsomption) et des relations partitives;
- Des relations non hiérarchiques, et donc associatives pragmatiques, et en particulier des relations séquentielles (temporelles) et causales.

Dans notre cas, nous avons dû "dilater" et adapter sensiblement les standards d'identification et d'emploi des relations sémantiques. Tout d'abord, parce que le sous-domaine est plutôt complexe. Ensuite, puisque, vu la nécessité de produire un outil, même si encore prototypique, qui réponde aux besoins des interprètes, l'exigence s'est imposée d'adapter de façon pertinente la structure ontologique tout entière. La biomédecine est en effet un domaine tellement articulé et complexe que la théorie formelle de la classification médico-biologique

est encore à l'état embryonnaire[3]. De suite, les relations principales employées pour l'ontoterminologie:

Italien	**Français**
Causa	Cause
Colpisce, interessa	Atteint, touche (à)
È	Est
È caratterizzato/a da	Est caractérisé/e par
Si manifeste, si esprime con / tramite	Se manifeste, s'exprime par
Si tratta, si cura con	Est traitée à / par Est soigné(e) à / par Le traitement repose sur
Dimostra, prova	Démontre, prouve Fait porter le diagnostic de
Origina, scatena, porta a	Cause, engendre, déclenche, porte à
Si differenzia / distingue in Può differenziarsi / distinguersi in	Se divise, se différencie en Peut se diviser, se différencier en
Sviluppa, porta allo sviluppo di	Développe, Porte à, conduit à développer

TAB. 1 – Tableau des principales relations sémantiques employées pour l'ontoterminologie

3.2. La création d'un glossaire thématique bilingue

Une fois les candidats termes extraits, nous avons procédé à les analyser avec l'aide des experts du secteur, afin de les valider et détecter les relations sémantiques et logiques les plus importantes. En mettant en relation les termes simples et composés identifiés, nous avons construit un glossaire très spécifique, à partir duquel nous avons ensuite modélisé notre ontologie *ad hoc*. La méthode

[3] Smith (2003: 90) affirme en fait: "Biological classes are marked always by an opposition between standard or prototypical instances and a surrounding penumbra of non-standard instances (not all the instances of the class human being are marked by the presence of amputation stumps or pituitary tumors)".

de sélection terminologique et de réalisation du glossaire a suivi les étapes suivantes:

- identification des termes de domaine présents dans les textes analysés;
- analyse du signifié et de la compréhension des concepts véhiculés par les termes, ainsi que de leur intension et extension;
- analyse contrastive entre les systèmes conceptuels des deux langues et recherche des équivalents dans la langue d'arrivée de l'interprétation simultanée (le français).

Manifestazioni cutaneo-mucose in presenza di HIV	**Manifestations cutanéo-muqueuses en présence de VIH**
Accesso	Poussée
Affezione cutanea	Affection cutanée
Aftosi, stomatite aftose afta - forma *minor* - forma *maior* - forma erpetiforme	Stomatite aphteuse, aphtose aphte - forme mineure - forme majeure - forme herpétiforme
Alone ecchimotico	Halo ecchymotique
Angiolipoma	Angiolipome
Angioma	Angiome
Angiomatosi epitelioide	Angiomatose épithélioïde
Anorettale	Anorectal
Balanopostite manifestazioni eritemato-essudative	Balanite manifestations érythémateuses exsudatives
Benignità	Bénignité
Benigno	Bénin
Bolla rilievo cutaneo circoscritto, costituito da una cavità a contenuto liquido, di dimensioni superiori a quelle della vescicola	Bulle dermatologique phlyctène rempli d'une sérosité contenant ou non du sang, ou l'ampoule due aux frottements et aux brûlures, plus grand qu'une vésicule

TAB. 2 – Extrait du glossaire réalisé à partir des termes extraits

4. La réalisation et la formalisation de l'ontologie

Aujourd'hui, les langages artificiels et les logiciels utilisés pour la création automatique d'ontologies (tels que Protégé, OWL, RDF) portent principalement sur deux relations sémantiques : la relation de subsomption (*is-a*) et la relation méréologique (partitive, *part-of*). La raison en est qu'il s'agit des relations principales qui lient les termes insérés dans les bases de données traditionnelles. Ces types de bases de données cataloguent les termes d'un domaine donné de façon hiérarchique suivant la direction hyperonyme ⇨ hyponyme (de « vin blanc » à « Chardonnay ») ou bien classe supérieure ⇨ classe inférieure (de « mammifère » à « chien »). Toutefois, ces relations ne suffisent pas à conceptualiser un vrai discours, lors duquel un orateur utilise des liens sémantiques et logiques beaucoup plus nombreux pour expliquer ou décrire tel ou tel sujet.

En d'autres mots, ce dont nous avons besoin est de pouvoir formaliser « informatiquement » les données ontologiques elles-mêmes. Il existe pour cela différents langages et modèles de représentation qui, toutefois, ne permettent pas de représenter ces informations en gardant une différence graphique « rationnelle » entre les différentes relations sémantiques à la base d'un domaine de connaissance. De surcroît, un logiciel adapté aux besoins des interprètes devrait permettre de garder et formaliser ce qu'on pourrait appeler « la valeur ajoutée » conférée à toute conceptualisation par son concepteur, c'est-à-dire la façon dont l' « ontoterminologue » a représenté visuellement les éléments de l'ontologie, en les approchant, les éloignant, les mettant sur une feuille plutôt qu'une autre, etc.

La modélisation manuelle des arbres ontologiques. La représentation visuelle d'une ontoterminologie pour interprètes de conférence est donc un élément déterminant de la qualité du résultat. C'est justement là, à notre avis, la plus grosse différence avec ce qui existe déjà dans l'état de l'art : les différents outils et modèles que nous avons testés ne s'intéressent qu'à la modélisation des relations et des concepts, avec plus ou moins de richesse de reproduction ; mais la représentation graphique qui peut en découler est, *a priori*, calculée par l'ordinateur selon des algorithmes assez basiques, qui n'envisagent pas la complexité de visualisation. Ces logiciels sont configurés selon des algorithmes très efficaces pour créer des arbres géométriquement bien structurés, qui calculent la largeur des branches inférieures pour positionner les parents et les instances supérieures de façon équilibrée et correcte. Toutefois, peu de contrôle est exercé sur le rendu visuel effectif qui, très probablement, sera moins efficient que celui que l' « ontoterminologue » a déterminé manuellement pour l'usage de l'interprète. Et la moindre altération du graphe lui-même sera susceptible de

déséquilibrer l'algorithme, et donc de produire un résultat sensiblement différent, pouvant rendre incorrecte ou invalide la conceptualisation. Pour cette raison, pour produire notre prototype nous avons créé manuellement les ontologies nécessaires par le logiciel graphique Visio 2007 de Microsoft, qui permet de réaliser des arbres et des graphes très variés.

4.1. La construction de l'ontoterminologie

Pour la construction et la structuration des diagrammes nous avons suivi l'approche suivante:

1) Construction d'un lexique « orthonormé », composé par les termes identifiés comme fondant le domaine (et dénotant donc un concept spécifique) et, éventuellement, par leurs synonymes (selon les lexiques employés par les utilisateurs) à partir de la documentation fournie par les spécialistes du secteur.
2) Identification des typologies épistémologiques des termes simples et composés repérés (en prenant en compte l'essence des termes et la définition de leur signification par l'expression de leurs différences spécifiques, des termes étant par exemple les véritables instances conceptuelles du domaine, d'autres étant leurs attributs ou des relations).
3) Modélisation et conséquente construction des diagrammes conceptuels schématisant le domaine sur l'empreinte de l'« Arbre de Porphyre », qui est à la base des ontologies modernes.

Nous nous sommes donc basés sur une combinaison et un usage concerté des méthodes suivantes:

1) Application du modèle épistémologique.
2) Application du modèle computationnel.
3) Formalisation logique *a posteriori*, sur la base de nos connaissances du domaine et surtout des indications fournies par les médecins spécialistes.

Ainsi, nous avons été à même de confectionner des ontologies très spécifiques et pertinentes en mettant en évidence les relations sémantiques les plus utiles et appropriées selon le discours que l'on prévoit d'interpréter.

4.2. Critères de développement des diagrammes

Représentation des relations. Les instances sont liées par des relations sémantiques de différent type (hiérarchiques et non hiérarchiques, verticales et horizontales), explicitées par des « satellites » pour faciliter la lecture. Nous avons repéré ces relations lors du choix des candidats termes pour l'ontologie,

puisqu'elles définissent et caractérisent le domaine de la biomédecine autant que les relations d'identité/appartenance.

Structuration graphique. Les formes, les couleurs et l'aménagement des diagrammes et des instances se basent sur des critères logiques ainsi qu'épistémologiques. Toute instance est caractérisée par une forme, une couleur et une position qui dépendent de son niveau, ses attributs, son essence et sa fonction. Nous présentons de suite une feuille de la conceptualisation en guise d'exemple. Les correspondants en français apparaissent tous, pour mieux montrer ce que l'interprète pourrait voir en choisissant de visualiser le correspondant d'un terme simple ou composé.

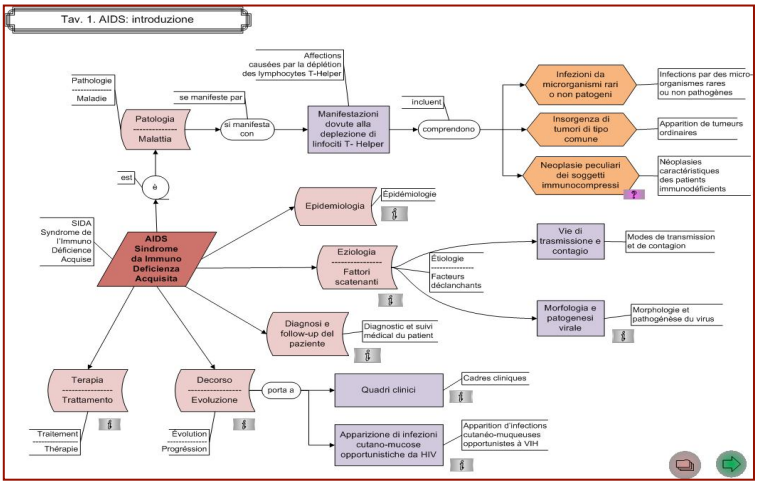


FIG. 4 – Page d'accueil de l'ontoterminologie, de type « top level » par rapport à la spécificité des concepts inclus.

Visualisation des équivalents interlinguistiques des instances. Nous avons inséré les équivalents français à l'intérieur de fenêtres de type *tooltip*, pour faire en sorte qu'ils s'affichent seulement au passage de la flèche de la souris. Ceci permet à l'interprète de visualiser les équivalents des termes le plus rapidement possible. Ces équivalents sont accompagnés, le cas échéant, d'un hyperonyme, un synonyme et/ou un hyponyme (en français et/ou en italien) pour que l'interprète ait une compréhension globale de la phrase qu'il est en train de reformuler. Nous avons choisi de faire apparaître les termes en français seulement si l'interprète en a besoin pour éviter que des graphiques trop compliqués ou étendus ne le fassent s'égarer, l'empêchant ainsi de suivre aisément la démarche conceptuelle du discours à travers les arbres.

Aménagement et organisation des diagrammes et des feuilles. La structuration des arbres a été basée sur l'évolution la plus vraisemblable d'un discours prononcé au sein d'une conférence spécialisée de médecine pour experts et/ou semi-experts (déterminée grâce à l'aide des experts du domaine). Selon le sous-domaine et la partie du discours à conceptualiser, donc, la structure suit les parcours suivants (l'un après l'autre ou bien en même temps):

- introduction ⇨ corps du discours ⇨ conclusion (pour l'ordre d'apparition des feuilles et des instances au sein des arbres);
- concepts supérieurs - à bas ou moyen niveau de granularité ⇨ sous-concepts spécifiques - à haut niveau de granularité (pour la représentation des pathologies, de leur étiologie et des traitements) ;
- origine ⇨ évolution ⇨ fin (pour la progression chronologique des pathologies).

L'ontologie se compose de plusieurs feuilles de travail et de différentes sections, à cause des limites de représentation que Visio nous a imposées. Chaque page contient donc une partie de l'arbre ontologique du domaine, qui se développe en suivant la démarche typique d'un discours scientifique. Les instances et les concepts relatifs sont présentés dans l'ordre de parution le plus probable d'un point de vue logique et conceptuel – ordre proposé aux experts et validé par ceux-ci. D'un point de vue graphique, l'ordre logique et chronologique attribué aux concepts est mis en exergue par l'emploi d'une forme et une couleur spécifique pour chaque niveau de la conceptualisation.

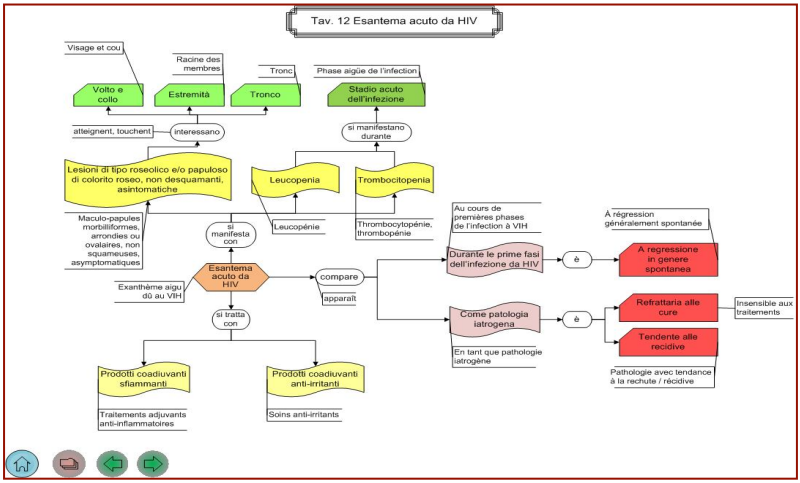


FIG. 5 – Fiche relative à l'exanthème aigu dû au VIH. Les instances faisant partie d'une même classe sont insérées à l'intérieur de cases de la même forme et couleur. La structure de l'arbre suit l'ordre logique et chronologique qui sera sans doute suivi par l'orateur.

Navigation de l'ontoterminologie. Pour rendre notre ontoterminologie prototypique bien navigable et facile à consulter, nous avons inséré des liens intertextuels qui permettent à l'usager de se déplacer à l'intérieur des modules de la conceptualisation. Les liens hypertextuels ont été employés pour:

- faciliter le passage d'une feuille à l'autre et la consultation des diagrammes;
- produire un support souple et adaptable à plusieurs type d'usages /recherches (même si encore prototypique) ;
- permettre aux usagers de suivre différents parcours discursifs et logiques lors de la conférence ou de la préparation au travail.

Plus précisément, les hyperliens relient:

- les instances des arbres de type *top-level* aux feuilles de conceptualisation spécifiques;
- les termes spécifiques aux fiches terminologiques relatives;
- tous les diagrammes à l'arbre introductif de type *top-level* (*homepage*) et au tableau présentant le sommaire des pathologies cutanéo-muqueuses conceptualisées en détail.

Pour la mise en exergue et l'activation des liens intertextuels, nous avons utilisé différents boutons. Le bouton « Home » (bleu, avec le symbole d'une maison) permet de revenir à la page d'accueil, pour pouvoir changer de volet quand l'orateur passe à la phase suivante de son discours. Le bouton « Informations » (gris, avec le symbole « i » ou celui d'une loupe) permet d'accéder directement à la conceptualisation de l'instance à laquelle il se réfère (par exemple, la schématisation de l'épidémiologie du Sida). Le bouton « Fiche » (rose, avec un point d'interrogation) conduit directement à la fiche terminologique créée pour les termes les plus intéressants, pour lesquels il était souhaitable de mener une analyse linguistique plus fouillée. Enfin, sur toutes les feuilles sont présentes les flèches « Suivant » et « Précédent », pour en faciliter la consultation.

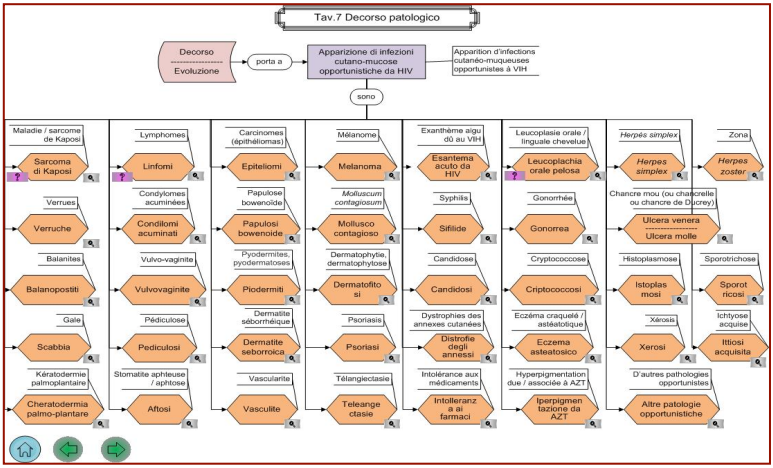


FIG. 6 – L'arbre des maladies cutanéo-muqueuses dues au VIH. Le bouton gris avec une loupe permet d'accéder directement à la fiche relative à chaque pathologie recensée et analysée.

Comme on peut le voir, l'ontologie a été structurée pour qu'elle soit non seulement une représentation conceptuelle graphique d'un domaine spécifique, mais pour qu'elle suive aussi de près la démarche de la conférence à interpréter, pour faire en sorte que l'ontoterminologie soit un véritable support pour le travail de l'interprète.

4.3. Quelques chiffres…

Notre prototype d'ontoterminologie bilingue se compose de 46 feuilles, chacune contenant un différent diagramme ontologique. Pour chaque langue, l'ontoterminologie inclut approximativement:

- 800 termes simples;
- 1000 syntagmes ou termes composés;
- 500 phrases définitoires / descriptives (descriptions ou définitions strictement liées à un terme ou concept);
- 15 relations sémantiques (de quatre types: de subsomption, méréologiques, causales, temporelles).

5. Essai empirique de l'outil ontoterminologique

Au terme de ce travail, nous avons testé l'ontologie ainsi créée de façon empirique. Nous avons demandé à deux types d'utilisateurs différents d'interpréter un discours créé *ad hoc* sur le VIH et ses symptômes dermatologiques en cabine, pour tester l'utilité concrète de notre ontologie. Un premier groupe qui a testé l'outil se composait d'interprètes de conférence professionnels, alors qu'un deuxième rassemblait des étudiants expérimentés en interprétation. Cette démarche nous a permis de vérifier également si l'expérience joue un rôle dans l'usage d'outils numériques de support à l'interprétation en cabine.

De surcroît, les deux groupes ont été ultérieurement divisés en un groupe A et un groupe B. Une semaine avant le test, les interprètes/étudiants du groupe A ont reçu un glossaire traditionnel sur papier contenant, entre autres, les termes que l'orateur utiliserait dans le discours d'épreuve. Par contre, les interprètes/étudiants du groupe B ont appris à consulter et utiliser l'ontoterminologie, qu'ils ont employée pour se préparer au test, ainsi que pendant l'interprétation en cabine. Les deux groupes ont effectué le test de la façon suivante:

1) **Groupe A:** composé de 7 personnes, les membres ont interprété le discours tous seuls, sans l'aide d'un collègue passif;
2) **Groupe B:** composé de 8 personnes, les membres ont été divisés en 4 couples. Dans chaque couple, un interprète a interprété activement le texte en français en simultanée, tandis que l'autre a fait fonction de collègue passif et a aidé le collègue à consulter l'ontoterminologie. En

outre, le collègue passif a précédemment interprété le texte préparé *ad hoc* à l'aide du seul glossaire thématique.

Lors du test, les participants ont enregistré leur performance, que nous avons collectée et analysée. Enfin, nous leur avons demandé de remplir un formulaire spécifiquement rédigé pour vérifier si l'ontologie a été utile ou pas, et quels changements pourraient améliorer l'outil ainsi conçu. L'analyse des performances des deux groupes et la lecture des formulaires nous a permis d'évaluer l'utilité effective de l'ontologie et le niveau de difficulté maximal qu'un support pour l'interprétation simultanée devrait afficher pour qu'il soit une aide et non une contrainte en cabine.

5.1. Analyse des opinions des interprètes

En général, le test empirique de l'ontoterminologie a donné des résultats satisfaisants. Les interprètes ont déclaré à l'unanimité qu'il s'agit d'un excellent outil de préparation au travail. En outre, les participants ont été d'accord, même si à différents niveaux, pour affirmer que l'ontoterminologie peut représenter aussi un bon support pour l'interprétation simultanée, en particulier en présence d'un collègue passif qui aide la consultation. Les réponses les plus significatives :

1) Pensez-vous que l'ontoterminologie est un bon support pour la préparation au travail?

Très bon (13)
Assez bon (2)
Pas très bon / médiocre (0)
Inutile (0)

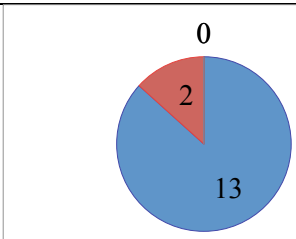


FIG. 7 – L'ontoterminologie en tant que support pour la préparation au travail

2) L'ontoterminologie est-elle représentative du domaine analysé?

Beaucoup	(11)
Assez	(4)
Peu	(0)
Pas du tout	(0)

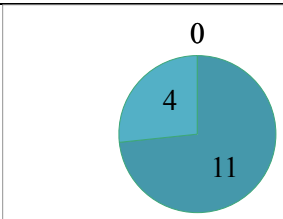
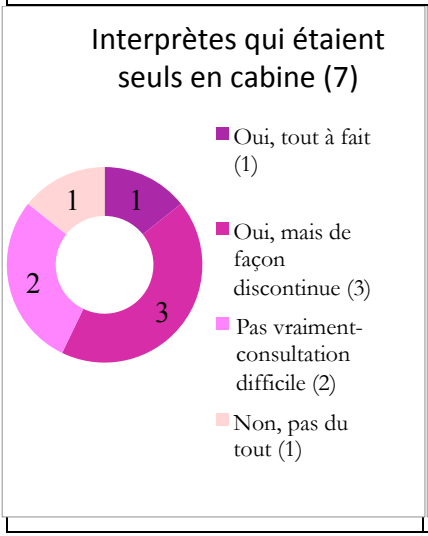


FIG. 8 – La représentation et l'adhérence au domaine analysé

3) A-t-il été possible de consulter et employer efficacement l'ontoterminologie en cabine?



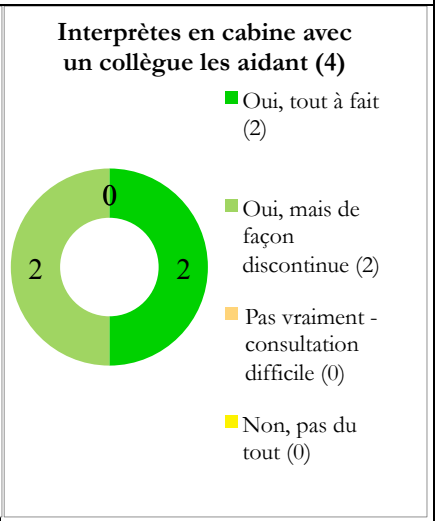


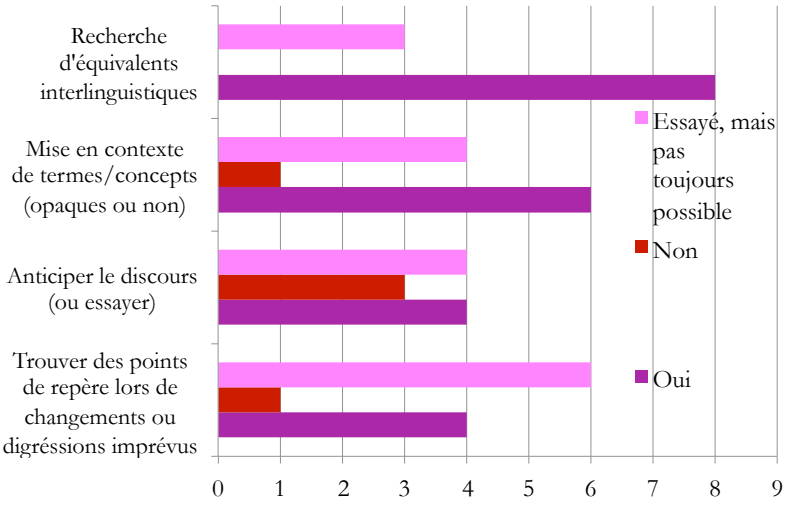
FIG. 9 – L'emploi de l'ontoterminologie en cabine (facilité d'usage).

Page suivante :

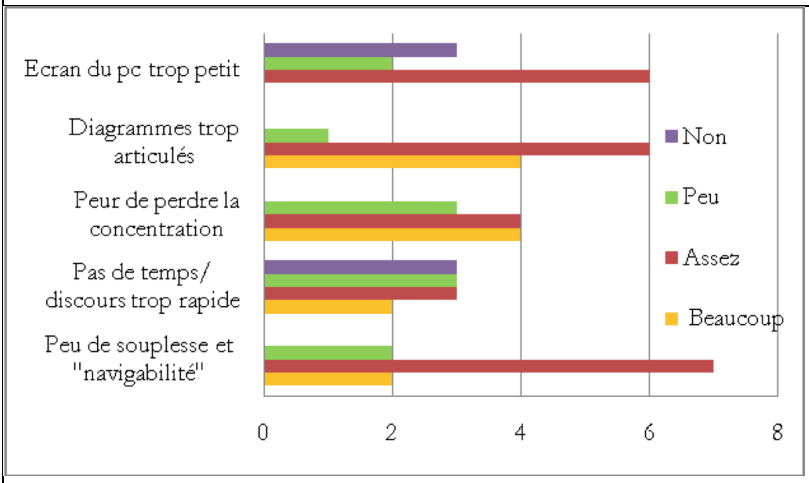
FIG. 10 – L'emploi de l'ontoterminologie en cabine (opérations)

FIG. 11 – L'emploi de l'ontoterminologie en cabine (difficultés)

4) Pour quel type d'opération avez-vous utilisé l'ontoterminologie en cabine?

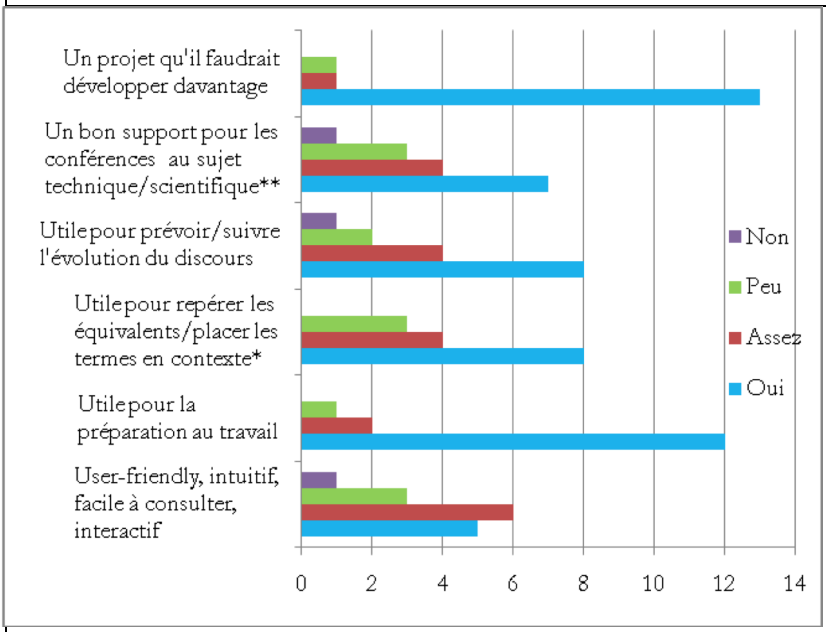


5) Quelles sont les difficultés que vous avez rencontrées lors de l'usage en cabine?



NB. Les participants qui ont rencontré plus de problèmes ont été les interprètes qui ont travaillé seuls en cabine. Ils ont eu beaucoup ou assez de difficultés à cause de diagrammes trop articulés, d'un manque de temps ou de la crainte de perdre la concentration en consultant l'outil de recherche.

6) Vous estimez que le prototype que vous avez testé est:



***** Commentaire de la majorité des participants: « Surtout si en cabine est présent un collègue qui peut aider ». ****** Commentaire de 5 participants: « Il faudrait toutefois que quelqu'un puisse le créer pour nous, faute de temps ».

FIG. 12 – Évaluation générale de l'ontoterminologie

Considérations générales. Le niveau de complexité de certains diagrammes devrait être diminué par des logiciels permettant d'en cacher/réduire certaines portions et afficher seulement la partie souhaitée de l'arborescence. Ces problématiques relèvent toutefois du fait que le support ontoterminologique a été développé avec un logiciel de création graphique non approprié pour ce type d'usage et de projet.

De toute façon, le prototype a déjà apporté des bénéfices en cabine, notamment grâce aux liens intertextuels insérés, à l'application aux instances d'une fenêtre *tooltip* pour faire apparaître l'équivalent au passage de la souris, et à la fonction *Find* (« Trouver ») de Visio, qui permet de rechercher un terme spécifique à l'intérieur de toutes les tables. Nous avons donc pu déjà atteindre plusieurs objectifs, puisque notre support permet, si les circonstances contingentes le consentent, de :

- anticiper ou, du moins, suivre l'évolution d'un discours technique/scientifique;
- mettre en contexte un terme/concept opaque;
- rechercher de façon ponctuelle un terme et/ou son équivalent interlinguistique, ainsi qu'un concept, à l'intérieur des diagrammes.

6. Conclusions

L'utilité d'un support ontoterminologique pour interprètes de conférence. L'évaluation de l'ontoterminologie de la part des utilisateurs a été généralement très ou assez positive, surtout de la part des interprètes qui étaient en cabine avec un collègue et pouvaient donc compter sur son aide pour la consultation et la recherche d'un terme spécifique à l'intérieur des arbres. En effet, le discours livré par les interprètes qui ont utilisé l'ontoterminologie pour la préparation et lors du test a été généralement plus précis et assuré d'un point de vue de registre et d'exactitude terminologique, affichant moins de fautes conceptuelles ou de grammaire.

Toutefois, des limites extrinsèques demeurent malheureusement, puisque le logiciel utilisé pour la création des diagrammes (Microsoft Visio 2007) n'est pas conçu pour la création automatique et formelle d'arbres ontologiques. Puisque l'outil utilisé n'est pas spécifiquement créé pour ce type de réalisation graphique, la conceptualisation peut paraître peu lisible, et l'apparition des équivalents linguistiques est parfois trop lente.

Pourtant, la majorité des interprètes, et surtout les professionnels, se sont dits très favorables au développement d'un tel support à l'interprétation de conférence, puisqu'il serait très utile non seulement pour la préparation au travail, mais aussi pendant l'interprétation même. Un tel outil permettrait en fait de récupérer les équivalents linguistiques des termes plus opaques et techniques, ainsi que de suivre l'orateur au fur et à mesure qu'il présente son discours. Ainsi, l'interprète serait à même de placer un terme qu'il ne connaît pas ou dont il ne se souvient pas dans le correct contexte conceptuel et d'usage, ce qui lui permettrait de trouver des points de repère même si le discours affiche un haut degré de spécificité.

6.1. Petit cahier des charges pour la création d'un logiciel adéquat pour ontoterminologies

Création des corpus linguistiques et extraction des candidats termes. Pour ce qui est de la création des corpus et de l'extraction automatique des candidats termes, il serait souhaitable de pouvoir compter sur un logiciel qui

fournisse une meilleure interopérabilité entre un système d'analyse statistique et les résultats d'une analyse distributionnelle. On pourrait par exemple intégrer une extraction à deux niveaux: une extraction des formes du texte pour la validation humaine des données, et une extraction de la forme lemmatisée des textes et des unités lexicales sélectionnées, qui seraient stockées dans la base de données.

Un logiciel pour la création d'ontoterminologies pour interprètes. Pour ce qui est de l'outil « idéal » de structuration ontologique du savoir, les énoncés définitoires des classes et des instances devraient pouvoir être réalisés sur la base de différentes relations sémantiques, qui soient aussi transversales et, donc, plus complexes par rapport à l'hyperonymie ou la synonymie. Il serait donc souhaitable d'avoir des logiciels qui permettent la réalisation de ce qu'on appelle « heavyweight ontologies », c'est-à-dire des ontologies fondationnelles ou régionales basées sur l'emploi d'un langage formel très expressif et à même d'aboutir à des structurations mieux articulées des arbres ontologiques, pour mieux représenter le domaine conceptualisé. Le logiciel dont nous aurions besoin pour créer un outil conçu pour les interprètes de conférence devrait donc:

- fournir un ensemble riche de distinctions conceptuelles (au moins, par rapport au domaine de l'application) ;
- être à même de calculer et modéliser clairement et correctement les relations sémantiques temporelles et de cause-effet par rapport aux relations d'identité et d'appartenance (ou faire en sorte qu'elles soient « remodélisables » manuellement, après la création automatique de l'arbre).

De surcroit, un outil adéquat pour la représentation de réseaux ontologiques pour interprètes devrait permettre de:

- appliquer aux instances l'option de fenêtre *tooltip*, pour insérer plus d'informations et les afficher seulement si nécessaire;
- créer automatiquement un index avec aperçu des tables, auxquelles pouvoir accéder par des liens intratextuels ou par une *overview* ;
- élargir et réduire des parties des diagrammes par une option de zoom, pour que les diagrammes se présentent aérés et bien lisibles. Il serait souhaitable que la navigation soit réalisable via une sorte de « loupe » qui « zoome/dé-zoome » des portions de l'arborescence au fil des clics de l'utilisateur (suivant tes hyperliens), ou suivant une action explicite (comme l'emploi de la roulette de la souris) ;
- effectuer une recherche directe des termes, en tenant compte de toutes les tables de l'ontoterminologie ;

- modéliser de façon automatique ainsi que semi-automatique les relations sémantiques, pour que le concepteur ou l'utilisateur puisse décider si les rendre ou non explicites.

Nous estimons que la production d'un logiciel basé sur notre cahier de charge permettrait de réaliser des ontoterminologies adaptées aux besoins spécifiques non seulement des interprètes de conférence et des traducteurs, mais également et plus génériquement des linguistes et des terminologues. Cela, parce que les difficultés auxquelles il faut faire face pour conceptualiser des domaines de spécialité très spécifiques sont encore considérables, vu qu'à présent les logiciels de structuration du savoir permettent de créer automatiquement et formellement des ontologies ne représentant que des relations sémantiques verticales et de simple inclusion/exclusion.

Il serait donc souhaitable de développer des outils à même de traiter des données plus complexes et, surtout, basés sur des algorithmes qui prennent en considération des paramètres de structuration « logiques », afin de créer une représentation graphique vraiment représentative du domaine analysé et dont les éléments soient agencés de façon rationnelle. Ainsi, la position, la forme et les couleurs des instances, des arbres et des ramifications pourraient, elles-mêmes, fournir aux utilisateurs des informations importantes sur les connaissances conceptualisées.

6.2. Considérations finales

La nécessité de pouvoir compter sur des outils plus appropriés pour un classement conceptuel de la terminologie de domaine est ressentie non seulement par traducteurs et interprètes, mais aussi par les spécialistes-mêmes de médecine et de biomédecine (Yapi J.H. *et al.*, 2003). Il serait donc souhaitable de poursuivre ce projet (ou la recherche en ce domaine) afin d'améliorer et développer davantage le prototype que nous avons produit et pouvoir construire des ontoterminologies appropriés pour le travail des terminologues, des linguistes et des interprètes / traducteurs.

Dans cette perspective, s'il était possible de développer des logiciels adéquats, il serait même licite de postuler la naissance de la profession de l'"ontoterminologue": un professionnel spécialisé dans la construction d'ontoterminologies spécialisées complexes, de type fondationnel, modulaires et dynamiques, applicables à plusieurs domaines et très souples et expressives, pour les tailler sur les besoins spécifiques des utilisateurs. Les ontoterminologies ainsi créées pourraient être réunies à l'intérieur de véritables bases de données ontoterminologiques, qui apporteraient des bénéfices à plusieurs professionnels

(interprètes, traducteurs, linguistes, terminologues, spécialistes des domaines conceptualisés, etc.). Étant produits de façon formelle, ces outils pourraient être créés de façon automatique ou semi-automatique, mis à jour, partagés et réutilisés. S'il en sera ainsi, il sera possible dans un futur proche de réaliser des ontoterminologies « fondationnelles » complètes et bien représentatives de notre réalité (ou, au moins, d'une partie de celle-ci). Nous pourrons donc dire d'être enfin arrivés à réunir le moderne sens du mot « ontologie » avec sa signification la plus ancienne: l'étude de « l'être en tant qu'être » d'Aristote.

Bibliographie

Bachimont, B., (2004) : *Arts et sciences du numérique : ingénierie des connaissances et critique de la raison computationnelle,* habilitation à diriger des recherches, Université de Technologie de Compiègne, in Malaisé V., (2005) : Méthodologie linguistique et terminologique pour la structuration d'ontologies différentielles à partir de corpus textuels, thèse de doctorat en linguistique, université Paris 7 – Denis Diderot, version en ligne.

Charlet, J,. (2002) : *L'ingénierie des connaissances : développements, résultats et perspectives pour la gestion des connaissances médicales,* habilitation à diriger des recherches, CHU Pitié-Salpêtrière, 10 décembre 2002.

Golbreich, C., Dameron, O., Gibaud, B., Burgun, A., *Web Ontology Language Requirements w.r.t Expressiveness of Taxonomy and Axioms in Medicine.,* in "International Semantic Web Conference", 2003, pp. 180-194.

Roche, C., *Le terme et le concept : fondements d'une ontoterminologie, TOTh 2007 :* «Terminologie & Ontologie : Théories et Applications », Annecy, 2007 .

Roche, C., Marty, J-C., Lacroix_S., *Ontologie et terminologie: le modèle OK,* in Rint - Réseau international de néologie et de terminologie , «Terminologies nouvelles », 19, 6-1999, pp. 101-110.

Wright, S.E., (2007) : *The Once and Future ISO 704*, in eDITion 2007/1, pp. 1-4.

Yapi, J.H., Lasquellec, A., Gibaud, B., *Evolution de Schéma et Migration d'Instances. Prise en compte des Besoins d'une Application Médicale.,* BDA, 1996, pp. 353-372.

A propos des auteurs

Veronesi Elisa

Elisa Veronesi est interprète de conférence et traductrice sectorielle. Elle travaille en collaboration avec le professeur Bertaccini et mène plusieurs recherches dans le domaine de la terminologie, de l'interprétation et de la traduction.

Via Nello Buzzi 4, 48124 Piangipane / Ravenna (Italy)

everonesi@gmail.com

Bertaccini Franco

Franco Bertaccini est professeur de terminologie à l'université de Bologne, responsable du département de traduction et directeur du laboratoire de recherche terminologique. Parmi ses succès il compte la publication de plusieurs dizaines d'articles en différents domaines.

Campus Universitario di Forlì - Via G. della Torre, 5 – 47100 Forlì (FC) (Italy)

bertacci@sslmit.unibo.it

http://terminologia.sslmit.unibo.it/

Semiotic Triangle Revisited for the Purposes of Ontology-based Terminology Management

Igor Kudashev, Irina Kudasheva

Abstract: In this paper, we examine the limitations of the traditional semiotic triangle from the point of view of ontology-based, multipurpose terminology management and suggest an alternative model based on the concept of terminological lexeme. The new model is being tested in the *TermFactory* project aimed at creating a platform and a workflow for distributed collaborative ontology-based terminology work.

Keywords: Ontology-based terminology management, semiotic triangle, terminological lexeme.

1. Introduction

In this paper, we describe the principles of modelling of the core schema for an ontology-based terminology management platform called TermFactory (TF). In ontology work the focus is usually on the relations between concepts and their instances while the designations of concepts are given less attention. Often they are just listed as alternative string labels attributed to concepts. TermFactory, in contrast, represents designations, too, as named ontology resources, globally identified by URIs.

TermFactory has been designed not only as a terminology management system for human users but also as a source of terminological and linguistic information for different applications and processes, such as text parsing and generation, speech recognition, machine translation, etc. Due to this multipurpose character of the TermFactory platform, traditional data models used in terminology management have not been readily applicable while designing its schema. In particular, we felt the need to modify and extend the semiotic triangle which had been the starting point in the data modelling of the TF.

At the time of writing, TermFactory has reached the stage of a working demo but there is still a lot of testing and evaluation ahead, so the data model presented in this paper may still undergo some changes. Besides, some features described in this paper, in particular those related to the layout of the headwords,

are supposed to be implemented later. Terminology used in the naming of classes and data types may still be revised at the later stages.

In the sections to follow we discuss the benefits of ontologization of terminological and linguistic data, describe the TermFactory project and platform in more detail, review the traditional semiotic triangle and its interpretations in the terminology theory, examine some limitations of the semiotic triangle from the point of view of multipurpose terminology management, and propose an alternative model.

2. Benefits of ontologization of terminological data

Ontologization of terminological data has several benefits. Ontologies have proved a powerful instrument of creating and sharing common understanding about different domains. Defining a term as a concept instance requires precise thinking and negotiations between the parties which are going to commonly use it. Thanks to this, the parties can in the future be certain that they are referring to the same, globally identified object.

Furthermore, information about terms, just like information about the domain, becomes machine-readable. Ontologies can be automatically checked for logical errors. This helps finding mistakes and inconsistencies in existing and newly created terminology.

In ontologies, a lot of data can be inferred automatically using ontology reasoning. Concept and property inheritance lets developers of terminological collections build their work on top of other terminologies. Information about transitive and symmetric relations too gets propagated automatically. A typical TF terminology project does not start from scratch, but imports bridge concepts and properties from more general terminologies and ontologies.

When inferences are inductive (for instance, in the case of partial synonyms and cross-language equivalence relations), a reasoner can generate educated guesses to be verified by a human user. Both types of propagation speed up the input of data in a terminology management system and assists automatic management of links.

An interesting feature of the TF term ontology is that it is not bound to the notion of an entry as the mandatory rigid "container" that keeps individual data elements together. Instead, each object of the description and each element of the description are represented as OWL statements which spell out their relations

explicitly and unambiguously. The same data can then be presented to the end users in many different ways. Static entries become dynamic orientations and views that result from different traversals and serialisations of the term ontology graph.

3. TermFactory Project and Platform

3.1. TermFactory project

TermFactory (TF) is a part of a larger project *ContentFactory* (2008–2010) carried out at several departments of the University of Helsinki and Helsinki University of Technology. The project is financed by the Finnish Funding Agency for Technologies and Innovation (Tekes) and a number of enterprises specialising in language technologies. The initiator and responsible leader of the project is professor Lauri Carlson (University of Helsinki). The *TermFactory* work package of the project aims at creating a platform and a workflow for distributed collaborative ontology-based terminology work.

TermFactory's mission is to allow companies, organizations and individual contributors to collaboratively produce multi-domain special language vocabularies and ontologies (see *Figure 1*). TermFactory can be used to organize content and standardize communication in global multilingual organisations as well as to boost the exchange of ideas and innovations and support education across language barriers.

TermFactory can also serve as a source of terminological and linguistic information for different applications and processes, such as text parsing and generation, speech recognition, machine translation, etc.

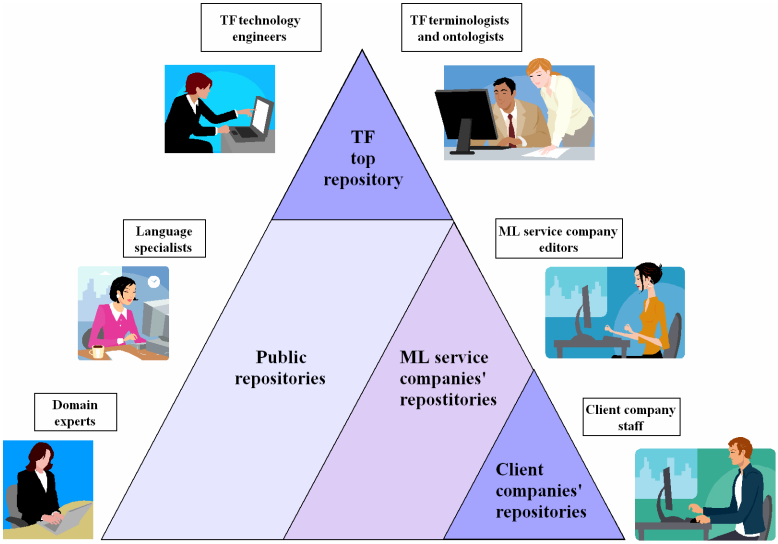


FIG. 1 – TermFactory's layers and users.

3.2. Collaborativeness

TermFactory is designed to support collaborative, Wikipedia-like terminology work by communities. The tools for collaborative terminology work are based on the content management systems like *Mediawiki*. They can also be implemented as plug-ins for the wiki platforms already used by companies, organisations or web communities.

3.3. TermFactory architecture

TermFactory is designed as a distributed resource. TermFactory network consists of OWL repository servers and collaborative wiki / forum platforms connected by a common directory (see *Fig. 2*). The nodes communicate in a peer-to-peer fashion on the web service layer. Collaborative platform servers are loosely coupled to the TermFactory repositories.

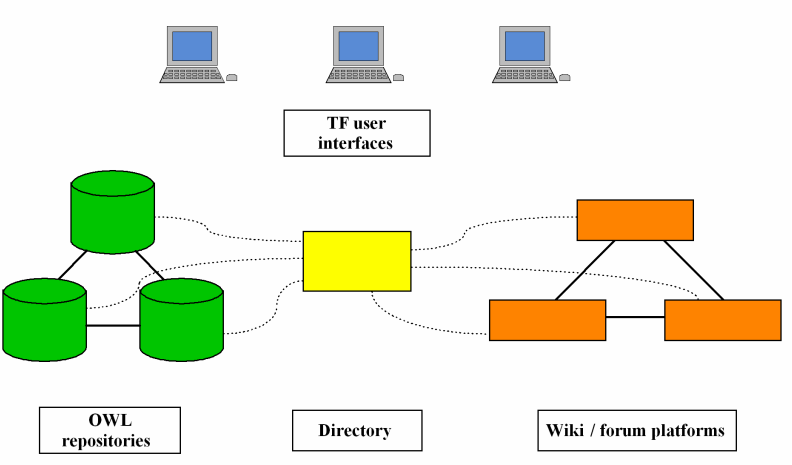


FIG. 2 – TermFactory architecture.

3.4. Multipurpose and federated character

TermFactory has been designed as a source of terminological and linguistic information for both human users and computer agents. TF accumulates and imports data from multiple sources, including wiki platforms. The multipurpose and federated character of the TermFactory has several important implications for the data modelling.

TermFactory has to be tolerant of a wide range of objects of description and different ways of data presentation. The range of described objects is not restricted to classic terms but includes other LSP designations as well. There is no notion of "minimum entry" in TF. Any set of triples conformant to the TF schema can constitute a TF document.

TF does not require that a particular set of data categories is used. Instead, it tries to federate heterogeneous content with the help of the list of upper-level "metacategories" (see Kudashev, 2009). For example, such class as "information about meaning" is general enough to accommodate all kinds of data related to meaning, including verbal and non-verbal definitions, descriptions, notes, etc. At the same time it is more precise for the purposes of search than "text fields" or "full text of the term record" used in many federated terminological resources.

In many cases computer agents need very specific and well-structured information about LSP designations. This is one of the reasons why TermFactory

allows providing separate descriptions for concepts and words. For example, it is possible to describe a concept which has not been attributed a term yet or a word which is not a part of any term presented in the term bank.

Among other things this approach considerably reduces the duplication of data. For example, as the same word can be a component of dozens and hundreds of terms, it is reasonable to describe its linguistic form once and refer to this description as many times as needed.

In spite of its multipurpose nature, TermFactory is first of all a reference resource which contains more or less generalized, context-independent linguistic data. This means, for example, that TF mostly deals with stable, generalized meanings of LSP designations, so there is no need to take into account occasional, contextual or idiolect meanings.

4. Semiotic triangle and its interpretations in terminology theory

The starting point in the modelling of the TF schema was the so-called "semiotic triangle" (*Fig. 3*). The triangle is often referred to as the "Ogden and Richards' triangle" as it was famously depicted by these authors in their book "The meaning of meaning" (Ogden and Richards, 1923: 11). However, the idea of the triangle has its roots already in the works by Aristotle (Seuren, 2006: 469).

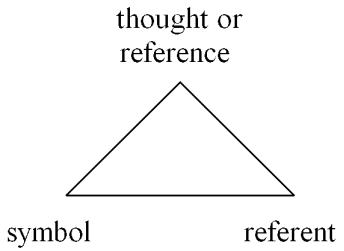


FIG. 3 – The original semiotic triangle (Ogden & Richards, 1923).

The triangle has undergone a long history of modifications and interpretations in different theories and by different authors. Terminology theory has adapted the semiotic triangle to explain the relationship between objects, concepts and terms (e.g. Schmitz, 2006: 579). The triangle is sometimes also represented as a tetrahedron with definition as the fourth vertex (e.g. Suonuuti, 2001: 13; Sanastotyön käsikirja, 1989: 24).

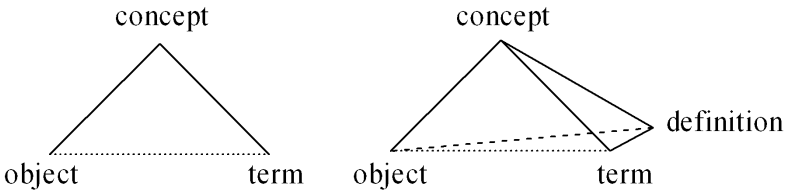


FIG. 4 – Interpretations of the semiotic triangle in terminology theory.

When we tried to apply the semiotic triangle to the TF schema, we realized that from the point of view of multipurpose terminology management the triangle had some limitations and required modifications and extensions.

5. Limitations of the semiotic triangle from the point of view of multipurpose terminology management

5.1. Range of designations

Terms are probably the most important object of description in terminology management systems but they are not the only object. At least the following LSP units should be taken into consideration as potential objects of a terminology management system in addition to terms (see Kudashev, 2010, forthcoming):

Appellations, i.e. designations of individual objects/concepts.

Nomenclature. This class has multiple interpretations in different terminology schools and theories. One interpretation is that nomenclature is designations of "primitive classes" which are formed by listing necessary characteristics without "closing the class" by specifying sufficient characteristics. A typical example of nomenclature in such interpretation is objects of mass production. It is hard to define them with a classical genus-species definition but it is possible to describe them precisely enough with the help of specifications. However, original specifications may become insufficient for singling out the product as new similar objects come to the market.

Names of objects of mass production, such as "Dreamland Soft" mattress, "Delux Beauty Relax" pillow or "Ecomoods Fabia" lamp, are often neglected in terminology theory and terminology management but in many companies there is a need to deal with them in a multilingual environment (for example, to translate and localize names of company's products). For more information about the

presentation of different classes of nomenclature in LSP reference resources see (Kudashev, 2005).

Names of goods should not be mixed with alphanumeric designations assigned to objects in manufacturing and inventory control systems. Stock-keeping units usually refer to *consignments* of products rather than products themselves. Linking between a material management system and a terminology management system is possible but material management categories are unlikely to become objects of terminological description in a term bank.

Formal notations. Special concepts and objects may be referred to with the help of means of formal notation, such as

- special symbols: §, €, °, ∑, ∞;
- formulae: H_2O, $[As@Ni_{12}As_{20}]^{3\square}$;
- international scientific names: "*Salix starkeana* subsp. *cinerascens*";
- codes: "ABHD12", "C20orf22", "DKFZP434P106", "dJ965G21.2", "BEM46L2", "ABHD12A" are code names of a gene with the official name "abhydrolase domain containing 12";
- catalogue names: "Messier 31", "M31", "NGC 224" are catalogue names of the Andromeda Galaxy.

Formal notations are often used in LSP texts interchangeably with the corresponding appellations and terms and sometimes they are the only existing designation of a special object or concept.

Lexicalized LSP expressions, i.e. single-word or multi-word expressions which have a relatively stable form and function in a particular LSP or special area of application. Here are a few examples of lexicalized LSP expressions:

- Instructions: "handle with care", "this end up" (ISO 12620:1999: 10);
- Military commands: "Stand at ease!", "Eyes right/left!", "Double march!";
- Set phrases used in radio and signalling: "More to follow", "How copy?", "Solid copy!"

Term elements, i.e. components of LSP designations which have a relatively stable specific meaning in a given LSP. Classical examples of term elements can be found in domains of medicine and chemistry. Here are examples from the domain of medicine:

- Prefixes "a-", "an-" mean an absence of something (apathy, analgia);
- Suffix "-ac" means "pertaining to something" (cardiac);
- Root "aur(i)-" means "pertaining to the ear" (aural).

In addition to morphemes adjectives, participles and components of complex words often function as term elements. When described as objects in their own right, term elements can be provided with a more comprehensive description of their meaning, etymology, usage as well as term formation models.

In TF, difference is made between *term elements* and *term components* because not all words which make up a term have stable specific meaning in a given LSP and deserve a separate terminological description. However, all term components can be provided with a general linguistic description.

5.2. Relations within the triangle

For some LSP designations, in particular for appellations, lexicalized expressions and term elements, the model of semiotic triangle is applicable only with certain assumptions or not applicable at all.

For appellations, it can be argued that they designate objects directly and not via concepts. Indeed, if conceptualization is used to group various objects together on the basis of their essential characteristics, with individual objects there is nothing to group.

Term elements denote properties rather than objects so they have "morphological" rather than conceptual meaning.

Lexicalized expressions, such as instructions, are usually statements about specific situations, so they neither denote special concepts nor refer to objects, at least directly.

5.3. Additional components of meaning

While working on the Finnish-Russian Forestry Dictionary (Suomalais-venäläinen metsäsanakirja, 2008, awarded by EAFT in 2008) we noticed that in many cases the meaning of an LSP designation was broader than the intention of the corresponding concept. In addition to the *conceptual meaning* an LSP designation may contain other components of meaning, such as:

- Different connotations, i.e. evaluative components of meaning;
- Inner form of expression (its "literal", morpheme-by-morpheme meaning);

- Components of meaning induced by other LSP or LGP meanings of the same designation;
- Components of meaning resulting from antonymous, synonymous, paronymous and other systematic relations of the designation;
- Different kinds of associations;
- Components of meaning resulting from consonance, rhymes, etc.

These components of meaning can be called *induced meaning* because they result from the attitude of language users towards the form and/or referent of the given LSP expression and/or units related to it by systematic relations or association.

Induced components are welcome and even cultivated on purpose when they create positive associations or allow users to express their attitude to the subject in informal communication. However, in most cases they only distract the users' attention from the logical meaning which is supposed to lie at the core of LSP communication.

This is probably one of the reasons why components of induced meaning have to a great extent been neglected in terminology theory. However, taking them into consideration is an important prerequisite for successful terminological nomination and LSP communication. In our opinion, this topic deserves deeper investigation; both in LGP and LSP (cf. Rigotti & Rocci, 2006, 444). For more information about the presentation of induced meaning in LSP reference resources see (Kudashev, 2006).

5.4. Synonymy, polysemy and homonymy

The traditional semiotic triangle abstracts from the fact that the same designation can have several meanings (polysemy, homonymy) and the same meaning can be denoted by several designations (synonymy). However, in terminology management it is very important to anchor the object of terminological description to exactly one form-meaning pair as meaning and domain of application may influence the formal and pragmatic characteristics of an LSP unit (e.g. pronunciation, inflection and combinatory power) and vice versa, the inner form and relations of a unit may influence its meaning and usage.

5.5. Inflected forms

An LSP designation usually has multiple inflected forms. In languages with rich morphology the number of such forms can be quite substantial.

The traditional way of dealing with inflected forms in reference resources is to choose the so-called "canonical form" which represents the whole paradigm. Non-canonical forms are usually ignored but irregular forms may be provided as reference articles.

Division of forms into those included in the reference resource and ignored is important from the point of view of terminology management.

6. Alternative data model

Below are suggested several modifications to the original semiotic triangle and a few extensions to the vertexes of the triangle which help overcoming the problems mentioned above.

6.1. Removing the object

The *object* vertex of the semiotic triangle is of little significance for terminology management systems because the focus in them is mostly on the designations and concepts. Elements of encyclopaedic information may be considered supplementary information about the meaning of LSP units.

6.2. Introducing modified lexeme

Most of the problems described above (the need to deal with a wide range of LSP designations, term components, lexicalized LSP expressions, synonymy, polysemy, homonymy and inflected forms) can be solved if we introduce a class which is general enough to embrace a wide range of objects of description and which represents a union of a set of canonical and non-canonical forms with exactly one meaning.

A very close concept can be found in general lexicography, namely that of a *lexeme*. ISO 24613 defines lexeme as an "abstract unit generally associated with a set of forms sharing a common meaning" (2008: 4). This definition would be completely suitable for our needs unless lexemes were associated in general lexicography only with words and word-like units. However, we also need to take into account LSP designations (designations of special objects and concepts), lexicalized LSP expressions and term elements.

We suggest that the lexicographical interpretation of lexeme (set of forms of a word or a word-like unit sharing a common meaning) should be called *lexicographical lexeme* while the terminological extension (set of forms of an LSP designation, lexicalized LSP expression or a term element sharing a common

meaning) should be called *terminological lexeme*. The concept of lexeme can thus be broadened to embrace both lexicographical and terminological lexemes (*Fig. 5*).

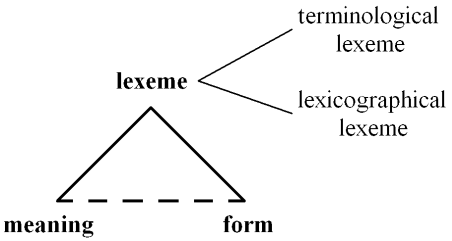


FIG. 5 – Introducing modified lexeme.

Lexeme is a tight union between meaning and form, so lines between lexeme and form as well as lexeme and meaning are solid. Meaning and form are often interconnected, too, but their ties are looser, which is depicted with a dashed line. As was mentioned before, meaning and form can even be described independently from lexeme and each other in the TermFactory platform. These descriptions serve as "shared resources" from which descriptions of lexemes sharing the same or similar form or meaning may be built up.

6.3. Extending the meaning

Meaning is divided in our model into denotative meaning and induced meaning. Denotative meaning covers the whole range of objects, concepts, properties, situations, conditions, processes, etc., which are referred to by the sign directly and not via associations. Induced meaning includes components of the meaning which are imposed on the sign via associations and relations of different kinds.

Denotative meaning is a broader class than *concept* as it also covers specific types of meaning characteristic of appellations, nomenclature, term elements and lexicalized LSP expressions.

Denotative meaning may be further divided into core denotative meaning and supplementary denotative meaning. This division roughly corresponds to the division made in terminology work between essential characteristics and non-essential but pragmatically important characteristics (see also Kudashev, 2006).

We would like to illustrate the interconnection between form and meaning and the need to differentiate between different types of meaning with the

acronym *PIGS* which became popular in economics during the financial and economic crisis 2008–2010. First, a few quotations from the Wikipedia (*http://en.wikipedia.org/wiki/PIGS_(economics)*; a mix of versions from 19.5.2010 and 30.6.2010; references and hyperlinks removed):

"*PIGS*, the original acronym referred to in 1997 to Portugal, Italy, Greece and Spain. […] The acronym has long been used by bank analysts, bond and currency traders […] and is used by some analysts, academics and commentators as a concise way to refer to the Eurozone countries of southern Europe noted for similar economic environments. […] Similar terms, such as "the Olive Belt" or "Club Med", have also been used for the same or similar groupings of countries in southern Europe. […] Ireland, previously known as the Celtic Tiger, became associated with the acronym in 2007 […]. The acronym thus became *PIIGS*, or remained *PIGS* with Ireland replacing Italy. […] The acronym is understood by many to be pejorative, but has been used as a term of art by some. It was denounced as racist in 2008 by the then Portuguese finance minister, Manuel Pinho, following a headline in the Financial Times that read, "Pigs in muck".*http://en.wikipedia.org/wiki/PIGS_(economics) - cite_note-24* Some variants of the acronym have been criticised because economies with similar financial problems, often notably the United Kingdom, are arbitrarily excluded. This has raised some doubts about a possible hidden agenda behind the acronym that would in reality correspond, according to some interpretations, to a wish to deviate the world's attention away from the delicate financial and budgetary situation after 2008's crisis in the UK and the US. GIPSY … refers to the same group as PIIGS, and has the same derogatory sense. It was adopted after protests against the PIGS acronym. It hasn't arrived to substitute the term, though, since it incorporates clear racist connotations*http://en.wikipedia.org/wiki/PIGS_(economics) - cite_note-23#cite_note-23*".

The case of *PIGS* is interesting in several respects. First, it has undisputable pejorative connotations resulting from its inner form which relates it to an LGP meaning of the word that has the same negative connotations. Second, the meaning of the acronym is highly dependant on its referent and form: new countries can not be added to the group or dropped unless the acronym is changed. As the form is so colourful, it is more tempting to change the meaning, which becomes blurry.

In fact, unlike the 1990s when the acronym was first introduced and had a relatively strict special meaning (four Eurozone countries of Southern Europe with particular economic problems like high or rising government debt levels and a high government deficits relative to annual GDP), in the 2000s the acronym

seems to be used more often just to point to the "weak link" in Europe's economy without referring to particular region or economic problems (cf. inclusion of Ireland and arbitrary exclusion of the UK from the list). The core denotative meaning is shrinking and gets more and more shaded by the induced meaning while the supplementary denotative meaning (background information needed for correct understanding of the acronym) is growing.

Synonymous terms and acronyms (the Olive Belt, Club Med, GIPSY, etc.) demonstrate other possible shades of induced meaning for the same referent – from neutral and even slightly positive to ironic and racist.

6.4. Extending the form

As was mentioned above, in terminology management and dictionary work it is important to divide the set of forms into the canonical form and non-canonical forms. Non-canonical forms may also be divided into those included in the reference resource and ignored. Extensions to the meaning and form vertexes of the triangle are depicted in *Figure 6*.

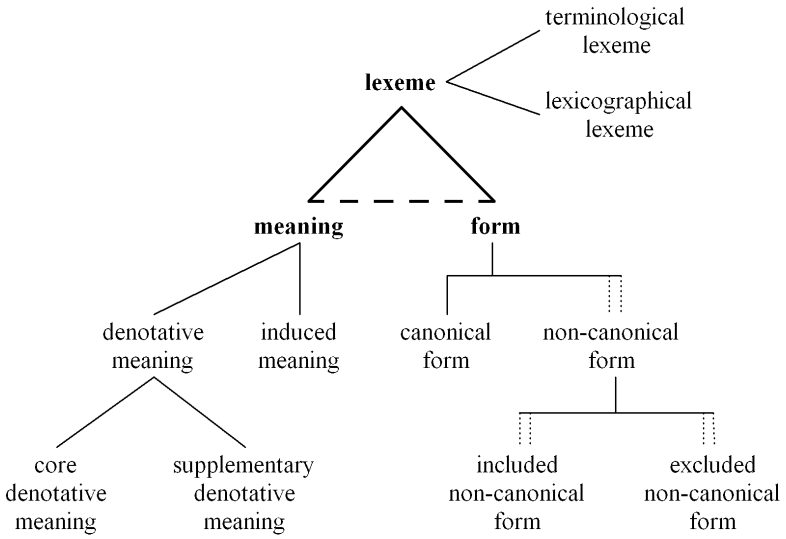


FIG. 6 – Extensions to the meaning and form vertexes of the triangle.

6.5. Additional extensions to the form

Headwords in a terminological database may also be divided into several classes according to their layout. This can speed up querying and displaying the records in a terminology management system. Such division has not yet been implemented in the TermFactory platform but it was used in MyTerMS terminographic processor in which the Finnish-Russian Forestry Dictionary and several other dictionaries were prepared (see Kudashev & Kudasheva 2006).

Headwords in the entry may contain elements of inline formatting, some additional comments and even other data categories, for example, hyphenation. In printed dictionaries a tilde sign or other means of compression can be used to substitute some portions of the headword. "Entry layout" takes into consideration possible elements of inline formatting as well as "foreign" elements.

In different lists, for example in term lists, hitlists and indexes, additional components of headwords are usually no longer needed but inline formatting has to be preserved.

For the purposes of search headwords usually have to be cleared from all the extra elements and inline formatting. Besides, headwords can be further optimized for the search by bringing them to an upper or lower case. *Figure 7* shows the additional subclasses for canonical and non-canonical forms included in the reference resource as headwords and index words.

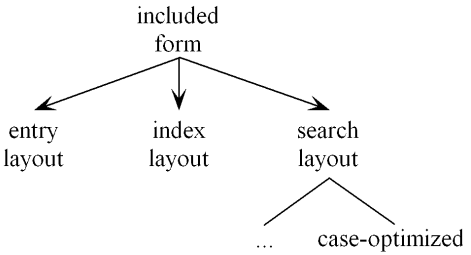


FIG. 7 – Additional extensions to the form vertex of the triangle.

Let us illustrate possible differences between the layouts with the help of a term *CO_2 laser*. In the entry layout it will have an element of inline formatting – a lower index. Besides, let us suppose that the term also has its hyphenation marked (CO_2 la-ser). In the index layout we preserve the lower index but remove the hyphenation (CO_2 laser). In the search layout we remove the lower index, too

(CO2 laser). If we also bring the term to the lower case, it will be best optimized for the search (co2 laser).

7. Conclusion

We have suggested a data model for the purposes of ontology-based multipurpose terminology management which helps overcoming some limitations of the traditional semiotic triangle. In particular, the model takes into account a wider range of LSP designations which can be objects of terminology management, additional types and components of meaning of LSP designations as well as issues which are pragmatically important in terminology management, such as lemmatization and resolving of homonymy, polysemy and synonymy.

Besides, our model partially breaches the gap between reference resources created within terminological and lexicographic frameworks and facilitates data exchange between them.

Bibliography

ISO 12620:1999(E). *Computer Applications in Terminology – Data Categories*. Geneva: ISO.

ISO 24613:2008(E). *Language Resource Management – Lexical Markup Framework (LMF)*. Geneva: ISO.

Kudashev, Igor (2005). *On the Treatment of Nomenclature in Specialised Dictionaries*. In Mård-Miettinen, Karita & Niemelä, Nina (eds) Erikoiskielet ja käännösteoria. Vakki-symposiumi XXV. Vöyri 12.–13.2.2005. Vaasa: Vaasan yliopisto, 176–183.

Kudashev, Igor (2006). *Additional Meaning Components of Terms and their Treatment in LSP Dictionaries*. In Lehtinen, Esa & Niemelä, Nina (eds) Erikoiskielet ja käännösteoria. VAKKI-symposiumi XXVI. Vaasa 11.–12.2.2006. Vaasa: Vaasan yliopisto, 143–149.

Kudashev, Igor (2009). *Improving Compatibility of Terminological Collections with a Bridging Classification of Data Categories*. Terminologija. Vol. 16. Vilnius: Lietuvi kalbos institutas, Terminologijos centras, 36–55.

Kudashev, Igor (2010) *What Can be an Object of Terminological Description in a Terminology Management System?* In Proceedings of the XXX Vakki symposium, Vaasa, Finland, 12-13.2.2010 (forthcoming).

Kudashev, Igor & Kudasheva, Irina. *Software Demo: The Terminographic Processor MyTerMS*. In Schryver, G.-M. de (ed.) DWS 2006: Proceedings of the Fourth International

Workshop on Dictionary Writing Systems, Tuesday 5th September 2006, Turin, Italy (Pre-EURALEX 2006). Pretoria: (SF)2 Press, 35–40.

Ogden, C.K. & Richards, I.A. (1923). *The Meaning of Meaning. A study of the influence of language upon thought and of the science of symbolism*. London: Routledge & Kegan Paul.

Rigotti, E. & Rocci, A. (2006). *Denotation versus Connotation*. In Keith Brown (ed.) Encyclopedia of Language and Linguistics. 2nd ed. Vol. 1. Oxford: Elsevier, 436–444.

Sanastotyön käsikirja (1989). Soveltavan terminologian periaatteet ja työmenetelmät / Toimittanut Tekniikan sanastokeskus. Helsinki: Suomen Standardoimisliitto.

Schmitz, Klaus-Dirk (2006). *Terminology and Terminological Databases*. In Keith Brown (ed.) Encyclopedia of Language and Linguistics. 2nd ed. Vol. 1. Oxford: Elsevier, 578–587.

Seuren, Pieter A.M. (2006). *Aristotle and Linguistics*. In Keith Brown (ed.) Encyclopedia of Language and Linguistics. 2nd ed. Vol. 1. Oxford: Elsevier, 469–471.

Suomalais-venäläinen metsäsanakirja (2008) / Kudasheva, I. & Kudashev, I. (authors); Vehmas-Lehto, I. & Gerd, A. (eds). Helsinki: Metsäkustannus.

Suonuuti, Heidi (2001). *Guide to Terminology*. Helsinki: Tekniikan Sanastokeskus.

About the authors

Kudashev Igor

Affiliations : University of Helsinki

Adresse postale : P.O. Box 239 (Paraatikenttä 6), FI-45100 Kouvola, Finland

Adresse électronique : igor.kudashev@helsinki.fi

Site web : http://www.helsinki.fi/palmenia/kouvola/index.htm

Kudasheva Irina

Affiliations : University of Helsinki

Adresse postale : P.O. Box 239 (Paraatikenttä 6), FI-45100 Kouvola, Finland

Adresse électronique : irina.kudasheva@helsinki.fi

Site web : http://www.helsinki.fi/palmenia/kouvola/index.htm

L'ontoterminologie pour la recherche d'information sémantique

Luc Damas, Christophe Tricot

Résumé :

En dehors d'une gestion électronique de documents à base de thesaurus, les nouvelles technologies (linked data) amènent de nouveaux usages et de nouveaux utilisateurs. Les bases documentaires grandissent de pair avec les capacités de stockage. Classer les documents manuellement devient irréalisable et trouver une information se complexifie. La recherche par mots-clés s'impose comme une alternative incontournable à la navigation. Les moteurs de recherche actuels, connus principalement sur Internet, se basent sur les mots-clés pour fournir des documents répondant aux besoins de l'utilisateur. Si le besoin se fait précis, généralement dans un contexte métier, l'imprécision de la recherche devient gênante, autant par l'apparition de documents hors-sujet que par l'absence de certains documents pertinents. La gestion sémantique de ces documents devient nécessaire. L'ontoterminologie, regroupant une dimension conceptuelle et une dimension terminologique, est une théorie intéressante sur lequel les processus documentaires peuvent s'appuyer : La compréhension des concepts du domaine et les mots pour en parler.

Mots-clés : Recherche d'information sémantique, ontoterminologie

1. Introduction

La taille des bases documentaires augmente de pair avec les capacités de stockage. Les usages ont évolué vers le tout électronique, sans souci de quantité. Tout est conservé. Le classement manuel de ces documents est une tâche qui dépend de nombreux facteurs et dont le résultat peut varier fortement. Pour une même personne, un document peut avoir de multiples utilités, et se retrouver dans des catégories différentes en fonction des besoins. De manière collaborative, au sein d'une même organisation, plusieurs personnes peuvent posséder des points de vue variés, et donc classer un même document de manières radicalement différentes. Enfin, un même document, en fonction de son volume, peut traiter de plusieurs thèmes et ainsi appartenir à plusieurs catégories. Ces considérations ont mené à une variation des usages, qui

deviennent de moins en moins normatifs. Le classement des documents devient sommaire et les utilisateurs se fient d'avantage aux outils de recherche.

Cette tendance est confortée par une habitude de plus en plus grande à effectuer des requêtes sur Internet. Formuler un besoin de manière textuelle plutôt que de naviguer dans un espace documentaire devient la norme, au sein d'un système d'information, mais aussi, de manière individuelle, au sein d'un logiciel de gestion de courriels ou d'un stockage physique sur disque dur.

Un système de recherche d'information (SRI) cherche à mettre en correspondance des besoins plus ou moins bien exprimés par un individu et des réponses à ces besoins (Chevallet, 2009). La formulation de la requête est un processus cognitif dans lequel l'individu exprime sous la forme de mots une compréhension d'un thème. Le défi consiste à lui fournir les documents correspondant à cette compréhension.

Notre approche se base sur l'ontoterminologie (Roche, 2007). L'avantage du modèle ontoterminologique est de distinguer clairement les aspects conceptuels, regroupant et structurant les idées, des aspects terminologiques, regroupant les mots et usages. Dans notre vision d'une recherche d'information (RI) qui vise à combler un « besoin cognitif » exprimé à l'aide de mots, l'ontoterminologie semble adaptée.

L'article présente dans un premier temps les éléments fondamentaux de la RI, de l'indexation à la recherche en passant par les mesures de pertinence. Nous présentons ensuite l'ontoterminologie en nous appuyant sur un exemple concret. Nous détaillons enfin notre système de recherche sémantique en l'illustrant d'exemples caractéristiques.

2. Eléments de RI

2.1. Définitions et principes généraux

Un Système de Recherche d'Information (SRI) est un outil informatique permettant la mise en correspondance d'un besoin, exprimé sous la forme d'une requête, et d'un élément d'information (*Information item* (Baeza-Yates et. al, 1999). Un élément d'information, au sens des Systèmes d'Information, est tout élément enregistré dans le système, en général un fichier : un document PDF, une page web, une image, une vidéo… La recherche d'information fournit donc des documents, conteneurs de l'information recherchée. La différence entre recherche documentaire et recherche d'information est ténue, et provient principalement des différentes disciplines qui manipulent les documents

(bibliothécaires, informaticiens). Ainsi, à la requête « quel temps fera-t-il demain ? », un SRI ne dira pas « Il fera beau », mais : « Cette information est disponible dans tel document »[1].

a) Requête

Une requête est généralement un ensemble de mots. Sous cette forme non ordonnée, il est supposé que l'utilisateur cherche un document contenant tous les mots. La plupart des moteurs de recherche proposent une syntaxe simple à base d'opérateurs logiques permettant d'affiner une recherche. La conjonction (ET logique) est la formulation d'une recherche de tous les mots. Elle est la forme par défaut (sans opérateur) proposée par les moteurs actuels. La disjonction (OU logique inclusif) est la formulation d'une recherche d'un mot parmi une liste de mots. Dans google par exemple, le symbole à utiliser est `or`. L'expression « terminologie or ontologie » permet de trouver tous les sites web contenant le mot « terminologie » ou le mot « ontologie » ou les deux. Enfin, la négation permet de trouver des résultats ne contenant pas un mot donné. L'expression « terminologie –ontologie » est une requête permettant d'accéder aux documents contenant le mot « terminologie » et ne contenant pas « ontologie ».

b) Recherche

La méthode la plus simple pour trouver un document permettant de répondre à un besoin et de compter le nombre de fois où les mots de la requête apparaissent dans les documents, et de trier les résultats par ordre décroissant. Cette méthode n'est toutefois pas utilisée car les temps de traitement sont beaucoup trop grands. Un parcours complet de tous les documents à chaque requête est consommateur en ressources informatiques (calculs, mémoire). Néanmoins, le comptage des occurrences des mots de la requête reste le critère essentiel d'évaluation de la pertinence d'un document (voir paragraphe 2.3).

c) Indexation

Afin de résoudre le problème des temps de parcours des documents, et plus généralement, pour pouvoir faire correspondre une requête et un document quelconque (non textuel), l'index s'avère un moyen idéal. Un index est un tableau associant mots et documents. Pour chaque mot, il est possible d'obtenir la liste des documents qui le contiennent. L'index peut être enrichi de nombreuses informations comme le nombre d'occurrences d'un mot donné dans un document, la distribution d'un mot dans un document, … La recherche se

[1] Répondre précisément à une question est un champs ouvert de l'intelligence artificielle : Les systèmes de question-réponse (Bellot, 2008)

trouve désormais grandement simplifiée et accélérée, puisque le jeu de documents n'est plus parcouru. La difficulté du travail se reporte alors sur la construction de l'index et sur l'évaluation des résultats de recherche.

Le processus d'indexation est un parcours de tout nouveau document, une seule fois. Tout mot du document n'étant pas contenu dans l'index devient une nouvelle entrée. Toute entrée de l'index qui est contenue dans le document pointe désormais sur ledit document. Ce principe d'indexation est valable aussi bien dans un système d'information local (Système de fichiers sur un ordinateur, serveur documentaire) que sur le web. Dans ce dernier cas, les moteurs de recherche « envoient » des robots (petits programmes, *bot* ou *crawler* en anglais) chargés de répertorier le contenu des pages web. Les algorithmes d'indexation associés effectuent des pondérations selon des critères qui forment des secrets bien gardés par chaque acteur.

d) Modèle vectoriel

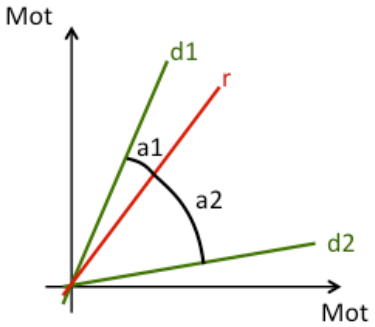
La forme la plus utilisée en RI pour l'index est le modèle vectoriel (Salton, 1971) et son amélioration LSI (Favre et. Al, 2006). Chaque document est représenté par un vecteur dont les indices sont les mots de l'index. La valeur de chaque dimension du vecteur est le nombre d'occurrence du mot correspondant dans le document.

	Mot 1	Mot 2	Mot 3	…
Document 1	0	2	1	…
Document 2	1	0	0	…
Document 3	0	0	3	…
…	…	…	…	…

Table 1 : La base documentaire est représentée sous la forme d'une matrice dans laquelle chaque document est un vecteur des nombres d'occurrences de chaque entrée de l'index.

La similarité entre documents, et la correspondance entre un document et une requête deviennent alors des mesures de distance entre vecteurs. La mesure la plus couramment utilisée est le cosinus. Il s'agit d'une mesure de différence entre les angles (les directions) des vecteurs pondérée par la taille de ces vecteurs. Deux documents proches sont représentés par deux vecteurs de directions

similaires (même sens). Cela s'applique de la même façon pour mesure la similarité entre un document et une requête (fig. 1).



Figre 1 : Deux documents d1 et d2 sont plus ou moins similaires à une requête r dans le repère formé par deux mots (projection à deux dimensions).

Le processus de recherche revient mesurer la distance entre la requête et chaque document du système. La réponse (l'ensemble des documents retournés) du SRI est triée selon une mesure de pertinence, détaillé ci-après.

2.2. Métriques

Un des objectifs majeurs de la recherche d'information est de mesurer la qualité des réponses d'un système.

a) Qualité d'une réponse

La qualité d'une réponse sert à classer les résultats d'une requête. Cette qualité doit tenir compte de l'importance du mot dans un document et de l'importance relative de ce même mot dans l'ensemble des documents. Un mot apparaissant dans tous les documents n'est pas discriminant. La plus connue des mesures de pertinence est le TF.IDF.

TF (Term Frequency) : Nombre d'occurrences d'un mot donné dans un document, avec une pondération sur la taille du document. TF indique si le document traite du mot donné.

IDF (Inverse Document Frequency) : Inverse du nombre de documents contenant un mot donné. IDF précise si c'est un mot discriminant (rare dans l'ensemble des documents).

b) Qualité d'un SRI

En amont, les concepteurs d'un SRI doivent être capables de mesurer la qualité globale des réponses. En phase de test, ils simulent un jeu de requêtes sur un ensemble de documents dont ils ont au préalable évalué la pertinence pour chaque requête. Cela permet d'obtenir les taux de bonnes et mauvaises réponses et ainsi affiner les résultats.

Le taux de rappel d'un SRI est une mesure du silence. Il s'agit du nombre de documents pertinents retournés lors d'une recherche par rapport au nombre total de documents pertinents pour cette recherche. Un faible taux de rappel signifie que le SRI a oublié des résultats importants.

Le taux de précision est une mesure du bruit. Il s'agit du rapport entre le nombre de documents pertinents retournés lors d'une recherche et le nombre total de documents retournés. Un faible taux de précision signifie que de nombreux résultats ne sont pas pertinents.

2.3. Améliorations

Afin d'améliorer la qualité des résultats, il existe de nombreux moyens. Le premier d'entre eux est de ne pas considérer l'égalité stricte de deux mots. La fonction binaire (égale, non égal) utilisée initialement compare les lettres deux à deux. Un mot au singulier et son pluriel sont considérés comme différents. L'usage actuel veut que la reconnaissance des mots se fasse par similarité, la formule la plus connue étant celle de Levenstein. Elle consiste à compter le nombre d'opérations (ajout, suppression, déplacement de lettre) pour passer d'un mot à l'autre. Cette mesure n'étant pas spécifiquement adaptée à la langue, elle a été adaptée pour répondre à certains besoins spécifiques. Ainsi, le coût de transformation d'un « é » en « e » est moins important que pour d'autres lettres et permet de compenser certaines petites fautes d'orthographe. De même, la présence d'un « s » en fin de mot a un impact réduit et permet de considérer comme équivalent le singulier et le pluriel. Néanmoins, ce principe a tendance à apporter du bruit, et les réglages sont délicats et dépendants de la langue. Dans nos applications, nous avions par exemple trouvé un amalgame entre les mots « solaire » et « polaire ». Certains moteurs effectuent une analyse morpho-syntaxique pour affiner les ressemblances.

Une deuxième amélioration plus utilisée consiste à proposer à l'utilisateur une complétion de la requête en fonction des requêtes les plus effectuées. Très efficace dans un cadre général (sur le web), le principe à tendance à être pauvre à l'initialisation du système et trop fourni au fil du temps. En particulier dans les cadres restreints (SI d'entreprise), les requêtes autour du même thème vont être

systématiquement complétées de la même façon. Ce principe n'est d'aucune utilité pour les requêtes précises faisant référence à peu de documents.

2.4. Limitations des moteurs à mots-clés

Les moteurs de recherche à base de mots-clés fonctionnent relativement bien et offrent une précision assez intéressante. Ils possèdent par contre un défaut de rappel dans la mesure ou la recherche sur les mots ne permet pas de retrouver des documents traitant d'un thème sans utiliser les mots classiques. C'est tout le problème de l'utilisation de synonymes ou de figures de styles. Le cas de l'ellipse, très utilisée dans les domaines techniques, ne permet pas à un moteur de recherche de retrouver un document contenant « relais de tension » si l'utilisateur requière des documents contenant « relais à seuil » (<Relais à seuil de tension>)

Pour augmenter le taux de rappel sans perdre en précision, il est nécessaire de baser l'indexation et/ou la recherche sur une structure de données référençant les relations entre mots. C'est d'autant plus nécessaire si le moteur de recherche se veut multilingue.

2.5. Solutions intermédiaires

La solution du réseau lexical est souvent utilisée pour augmenter la qualité des moteurs, en particulier par la prise en compte de la synonymie. Wordnet en particulier est au cœur de projets d'indexation dite sémantique (Chevallet, 2008).

En marge, la folksonomie (Hotho, 2006) est un principe à la mode avec le caractère social que prennent les systèmes d'information et le web en général. Dans ce cas, ce sont les utilisateurs qui indexent les documents à la volée. Le principal avantage, c'est que cette indexation est le reflet d'une compréhension du document, et l'index ne contiendra pas nécessairement les mêmes mots. Le principal inconvénient est que l'indexation se fait par des points de vues particuliers dépendant d'usages particuliers. On trouve généralement des niveaux d'expertise variés qui décrivent les documents soient en termes généraux (pour le néophyte), soient en aspects précis (pour l'expert qui s'attache souvent à ce qui est considéré comme détail par le néophyte). A l'extrême, les communautés de pratiques différentes peuvent mener à des descriptions radicalement déconnectées. Les points de vues n'ayant parfois rien de commun, les résultats peuvent être surprenant.

Il est nécessaire de prendre en compte la compréhension d'un document dans la recherche d'information. La compréhension est le résultat d'un processus cognitif, spécifiquement humain. Elle mobilise des connaissances qui ne sont pas

uniquement linguistiques. Nous pensons donc qu'il faut adosser le processus de recherche à une conceptualisation du domaine détachée des considérations linguistiques, tout en prenant en compte la langue qui reste le point d'accès aux documents. Notre solution passe par une formalisation ontoterminologique.

3. Ontoterminologie

3.1. Concepts et termes : systèmes sémiotiques distincts

« *L'ontoterminologie est une terminologie dont le système notionnel est une ontologie formelle* » (Roche, 2007). Elle est composée de deux systèmes sémiotiques distincts. D'une part, l'ontologie structure les concepts du domaine et fournit une compréhension indépendante de la langue. D'autre part, les termes fournissent les moyens d'exprimer les idées. La relation entre termes et concepts est décrite ci-après.

3.2. Méthodologie

La méthodologie de construction de l'ontoterminologie est discutée dans (Roche, 2007). Nous la rappelons ici pour insister le principe important de la séparation des systèmes de signes.

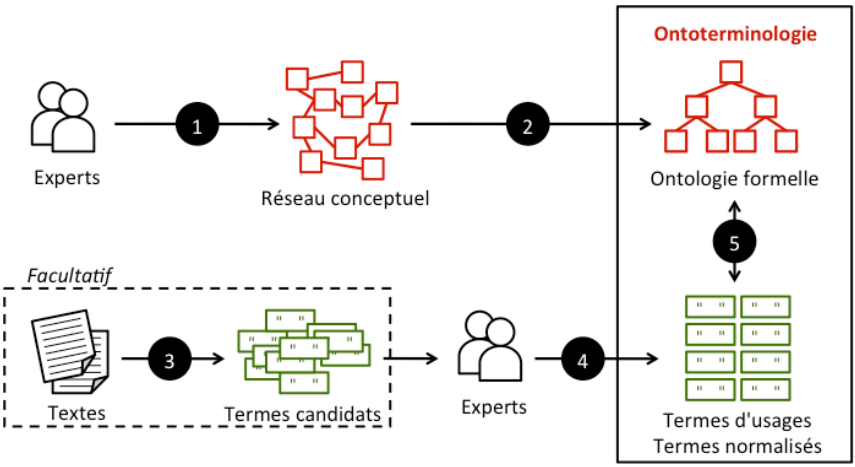


Figure 2 : Ontoterminologie : Une méthodologie mettant l'expert au cœur de du processus. Les concepts, en haut, sont séparés des termes, en bas. La difficulté du travail de modélisation réside dans la séparation des deux systèmes lors de leur construction.

La démarche se décompose en cinq parties révisables. La première vise à lister les concepts du domaine. Les experts sont au cœur de la démarche, interviewés et guidés par l'ingénieur cogniticien. Ils sont invités à définir précisément les objets sur lesquels ils travaillent. Rapidement, les informations se structurent sous la forme d'un réseau semi-formel. L'étape 2 consiste à formaliser ce premier résultat. L'ingénieur cogniticien, guidé par l'expert, rectifie le réseau selon une configuration formelle. Elle est ensuite révisée et validée. L'étape 3, facultative, utilise un extracteur de candidats termes pour lister une première série d'expressions faisant référence aux concepts. Les experts sont invités (étape 4) à fournir les termes utilisés dans le domaine par les différentes communautés de pratiques. Enfin, chaque terme est associé au concept qu'il désigne (étape 5).

3.3. Représentation, exemple

Afin de présenter la structure de l'ontoterminologie, nous nous appuyons sur des exemples tirés d'une ontologie qui traite de l'escalade. L'exemple présenté en figure 3 se focalise sur les différents ancrages. *Un ancrage sert au grimpeur à assurer sa sécurité. Il s'agit d'un dispositif d'assurage inséré, souvent fixé, dans le rocher auquel le grimpeur s'attache au fur et à mesure de sa progression.* L'exemple ne définit pas l'ancrage, mais les différentes sortes d'ancrage en fonction de la notion générale.

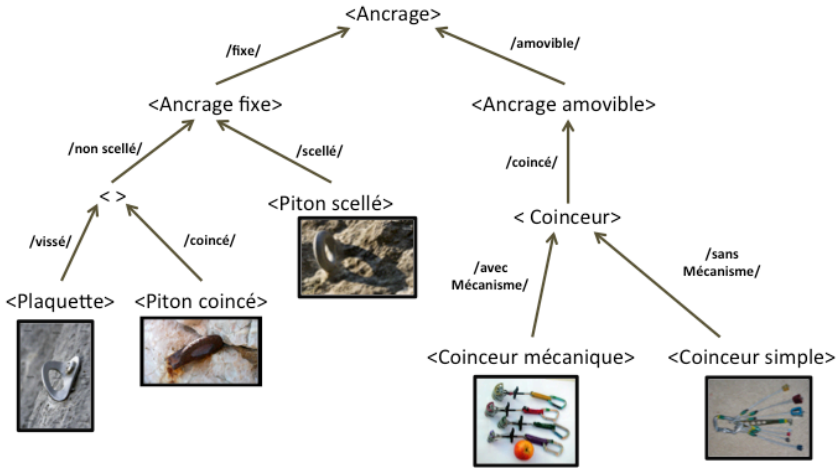


Figure 3 : Exemple partiel d'ontologie (simplifiée) : Les ancrage en escalade. Le terme « piton » en escalade désigne un piton coincé. Le <Piton scellé> n'est jamais appelé « piton » ni « piton scellé », mais plutôt « goujon ». Le terme « goujon » a quant à lui un sens plus général en mécanique.

Les concepts sont notés entre chevrons et nommés par une expression en majuscule. L'<Ancrage> est le concept parent et constitue la base de la définition formelle des concepts plus spécifiques. Les flèches représentent la relation de subsomption. Ainsi, <Ancrage fixe> est une sorte de <Ancrage>. La relation entre les deux concepts est étiquetée par une différence spécifique (Notation : encadrée par /) (Roche, 2001). La définition formelle de <Ancrage fixe> est <Ancrage> /fixe/ (*Un "ancrage fixe" est un "ancrage" qui a la caractéristique d'être "fixe"*). Cette définition semble apporter peu d'information car les expressions désignant les concepts s'incluent. La structure a ici plus d'importance que les mots utilisés, car c'est sur cette structure que les raisonnements s'appuieront.

L'arbre ontologique étends ses branches sur cette base jusqu'au niveau de précision requis. Les concepts les plus généraux (de premier niveau dans l'arbre ontologique) sont nommés « catégories ».

<Plaquette> $=_{def}$ <Ancrage>/fixe//non scellé//vissé/

<Coinceur mécanique> $=_{def}$ <Ancrage>/amovible//coincé//avec mécanisme/

La dimension terminologique regroupe l'ensemble des termes et les lie aux concepts. « goujon » est le mot généralement utilisé en escalade pour désigner un <Piton scellé> (<Ancrage>/fixe//scellé/), le nom du concept n'étant quasiment jamais employé. En mécanique, le terme « goujon » désigne une catégorie de matériel plus générale, incluant le goujon de l'escalade, d'où l'importance de définir formellement les concepts. Le terme « coinceur » n'est généralement pas utilisé pour désigner un <Coinceur>, mais plutôt <Coinceur simple>, le <Coinceur mécanique> étant plutôt appelé « friend »[2].

La figure 4 présente trois termes très utilisés pouvant apparaître dans le même texte, voir dans un même paragraphe. L'expression « huit » en escalade désigne autant un <Matériel d'assurage>, un <Nœud> qu'un <Niveau> de difficulté.

[2] Friend est une marque commerciale qui n'existe plus. Le terme « friend » est resté et désigne tous les coinceurs à mécanisme, quelque soit leur marque.

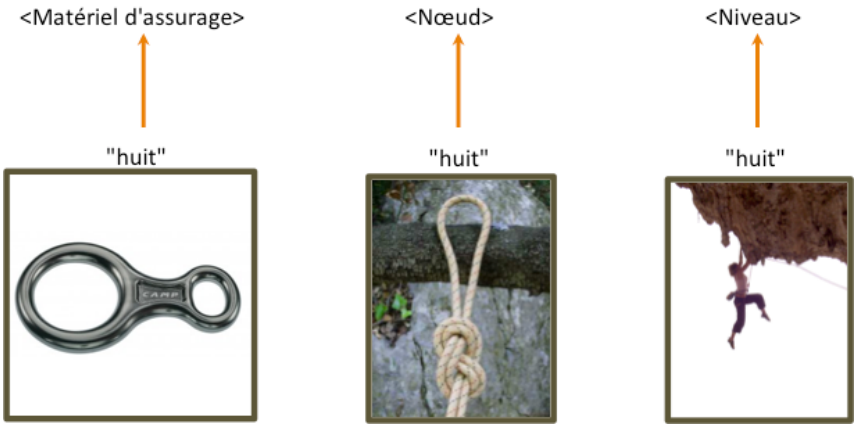


Figure 4 : Polysémie.

Le modèle ontoterminologique propose ainsi une architecture formelle à base de concepts associée à un répertoire de termes. Il est donc possible, d'une part d'effectuer des inférences, et d'autre part de faire référence aux contenus de documents. L'ontoterminologie constitue donc un outil idéal pour la recherche d'information sémantique.

La présentation du modèle n'est ici que partielle et ne conserve que certains fondamentaux nécessaires à la recherche d'information. L'ontoterminologie présente une dimension conceptuelle plus riche, avec des relations conceptuelles variées, ainsi qu'une dimension terminologique détaillée, avec définitions, catégorie grammaticales, réseau lexical…

4. RI Sémantique

La recherche d'information sémantique s'appuie en général sur une ressource extérieure de type thesaurus ou ontologie pour augmenter la qualité de la réponse. On dit que le SRI s'appuie sur une compréhension du domaine. Plus celle-ci est fine et validée, meilleurs sont les résultats. Nous présentons ici la RI sémantique telle que nous la concevons, basée sur l'ontoterminologie.

4.1. Marqueur sémantique

Comme dans la plupart des moteurs de recherche, notre proposition se base sur un index. Il est ici constitué de ce que nous nommons « marqueur sémantique ».

Le marqueur sémantique est la signature d'un document dans le contexte de l'ontoterminologie. Un même document peut avoir différents marqueurs sémantiques selon qu'il est considéré dans un domaine ou un autre. Produit de l'indexation, il contient les concepts identifiés dans le document, ou plus précisément, il est la liste des différences des concepts dont traite le document.

De manière plus formelle, un marqueur sémantique est un vecteur (tel que présenté au paragraphe 2.1) sur la base des différences spécifiques et des catégories de l'ontologie. Les valeurs de chaque dimension du vecteur décrivent la présence de la différence associée dans le document : valeur entre 0 et 1 (0 pour absent, 1 pour présent).

<Ancrage>	/fixe/	/Amovible/	/non scellé/	/scellé/	/vissé/	…
1	1	0	1	0	1	…

Table 2 : Partie de marqueur sémantique d'un document traitant de <Plaquette>. La définition complète du concept apparaît dans le marqueur.

La définition complète de chaque concept identifié apparaît dans le marqueur sémantique. Un document traitant d'un concept donné s'intéresse indirectement à ses concepts fils. Cette information est implicite dans le vecteur et pourra s'avérer utile pour la recherche.

4.2. Indexation

La construction du marqueur sémantique obéit aux mêmes règles générales que la construction d'un vecteur classique. Le document, généralement lemmatisé, est parcouru dans son intégralité. Chaque terme de l'ontoterminologie y apparaissant est identifié et les définitions des concepts associés sont ajoutées au marqueur sémantique. Si un terme est polysémique, le poids des différences dans le vecteur est divisé par le nombre de concepts les partageant. Ainsi, il est possible d'exprimer l'incertitude de présence d'un concept. Cette incertitude est relativement légère lorsqu'il s'agit de concepts parents, formulés dans le document à l'aide de métonymies. L'incertitude est plus lourde lorsque les termes sont « plus » polysémique (le nombre de différence partagées entre les concepts est faible). C'est le cas, à l'extrême, des trois versions du terme « huit » présenté plus haut. L'indexation est sujette à de nombreux paramètres et procédures que nous ne détaillons pas ici.

4.3. Recherche

La recherche en elle même ne présente pas de réelle différence par rapport à celle basée sur le modèle vectoriel initial (similarité vectorielle et tri). Les calculs s'effectuent sur les marqueurs sémantiques. La requête est travaillée pour

augmenter la quantité de résultats pertinents et diminuer le bruit. Dans un premier temps, les concepts correspondant aux termes de la requête sont extraits pour obtenir un marqueur sémantique identique à celui des documents, issu du même algorithme d'indexation.

La requête est implicitement simplifiée et augmentée sous la forme du marqueur sémantique, simplifiée car des redondances genre + espèce ne sont conservées que les espèces, augmentée car tous les documents contenant des termes désignant les concepts requis sont impliqués, y compris ceux ne comprenant pas les termes de la requête.

Ainsi, une requête de la forme « ancrage fixe coincé » sera signée par <Ancrage>/fixe//non scellé//coincé/ et sera interprétée naturellement comme traitant du concept spécifique <Piton coincé>. Les concepts encore plus spécifiques (possédant plus de différences) partageraient ces différences et seraient naturellement impliquée dans la recherche.

Ce travail sur la requête peut se résumer à deux actions principales :

- D'une part, tous les synonymes de tous les termes présents dans la requête sont ajoutés parce qu'ils sont autant de référence au concepts.

- D'autre part, tous les termes désignant des concepts plus spécifiques sont ajoutés, de manière à multiplier les résultats pertinents, même plus précis. Ceci est dû au partage des différences spécifiques. Ils ont en commun une partie de la définition.

Le classement des résultats par ordre décroissant de pertinence s'effectue selon une nouvelle mesure similaire au TF.IDF, mais actualisée pour la dimension conceptuelle. Le premier membre est la fréquence d'une concept "dans" un document : le nombre de fois où un des termes le désignant apparaît. Le second membre est l'inverse du nombre de documents où apparaissent les termes désignant le concept.

5. Meta-moteur

La section précédente présente le principe général de la recherche d'information appliquée à une base documentaire. La recherche d'information sémantique peut aussi être appliquée à Internet avec la différence majeure qu'il n'est pas possible d'indexer les documents.

Une solution consiste à développer ce qui est communément appelé un meta-moteur. Il s'agit d'utiliser les moteurs de recherche classiques et d'y injecter des aspects sémantiques. L'ontoterminologie peut être utilisée en amont sur la requête et en aval sur le traitement des résultats.

La requête de l'utilisateur est indexée comme précédemment (génération d'un marqueur sémantique). Les concepts requis sont identifiés et éventuellement désambiguïsés. Le système reformule la requête en utilisant tous les termes désignant les concepts identifiés. Ceci permet aux moteurs du web de retourner plus de documents, y compris des documents ne contenant pas les mots de la requête initiale. Une demande sur le mot « coinceur » effectuera aussi une recherche sur le mot « friend ». L'ajout de ces termes amène un bruit qui est supprimé par l'augmentation de la requête avec les différences dénomination du domaine. Ceci permet d'éliminer du bruit. Il est donc possible, par exemple, de retourner des documents traitant de « goujon » sans parler de mécanique. Enfin, les termes désignant les différents sous-concepts des concepts identifiés peuvent être recherchés afin de trouver des documents très spécifiques pouvant répondre au besoin de l'utilisateur. Ansi, une personne recherchant des informations sur les « pitons » peut être intéressée par des documents traitant de « plaquette ».

6. Conclusion, perspectives

La recherche par mot-clé présente des limites, même si l'habitude est bien ancrée et que les essais successifs permettent d'obtenir de bons résultats. Ceux-ci restent incertains et la qualité est difficilement vérifiable. La recherche sémantique s'impose dans des cadres techniques. L'ontoterminologie est un support efficace à la recherche d'information puisqu'elle permet de prendre simultanément en compte la compréhension du domaine et les mots pour en parler.

Notre approche a été validée et reste en cours de validation de deux manières. D'une part, nous avons mené diverses simulations sur des documents indexés manuellement pour vérifier que notre outil est efficace. Ces simulations devront s'accompagner d'une expérimentation réaliste pour un passage à l'échelle de la vérification expérimentale. D'autre part, l'outil est en usage dans diverses organisations et offre une satisfaction aux utilisateurs. Cette validation reste informelle, mais montre une certaine qualité de l'approche.

Malgré tout, il reste des évolutions possibles. Actuellement, le système ne prend en compte qu'une seule relation conceptuelle. Même si la relation « sorte-de » reste à nos yeux capitale car elle représente la définition des concepts,

d'autres types de relation devraient être pris en compte. La composition décrit des objets formés à partir d'autres objets. Chercher un document à propos d'un concept devrait dans un certaines mesure chercher ses composants. Il en va de même pour la relation de fonction.

Du point de vue de la structure du document, il existe des documents très pertinents à propos d'un concept ne contenant un terme le désignant que dans le titre. Avec les méthodes s'appuyant sur la fréquence des mots, ce document serait considéré comme non pertinent. Il faudrait ici accorder un poids lors de l'indexation à certaines parties du document. Cette tâche est délicate et requiert un certain nombre d'études, en particulier inter-langues.

Enfin, du point de vue de la langue, il restera toujours le biais des certaines figures de style, en particulier la métaphore qui va chercher les termes dans des domaines disjoints. Plus pragmatiquement, la lemmatisation des documents et la diversité des langues reste une barrière qui laisse ouvert le thème de la recherche d'information.

Bibliographie

Baeza-Yates R., Ribeiro-Neto B. (1999). *Modern Information Retrieval.* Ad- dison Wesley, New York, USA.

Bellot P., Boughanem M., "*Recherche d information et systèmes de questions-réponses*", 2008 in " La recherche d'informations précises : traitement automatique de la langue, apprentissage et connaissances pour les systèmes de question-réponse (Traité IC2, série Informatique et systèmes d'information)", sous la direction de B.Grau, Hermès-Lavoisier, chapitre 1, p. 5-

Chevallet J-P. (2009). "*Ressources endogènes et exogènes pour une indexation conceptuelle intermédia*", Habilitation à diriger des recherches

Favre B., Béchet F., Bellot P., Boudin F., El-bèze M., Gillard L., Lapalme G., Torres-Moreno J.-M. (2006) *The LIA-Thales summarization system at DUC-2006*, Document Understanding Conference (DUC-2006), New York (USA).

Hotho A., Jäschke R., Schmitz C., Stumme G. (2006) *Information Retrieval in Folksonomies:* Search and Ranking. ESWC 2006: 411-426

Roche C. (2001) : "T*he 'Specific-Difference' Principle : a Methodology for Building Consensual and Coherent Ontologies*", Actes de la conference IC-AI'2001, Las Vegas , USA

Roche C. (2008) : "*Le terme et le concept : fondements d'une ontoterminologie*", Actes de la première conférence TOTh 2007, Terminologie & Ontologie : Théories et applications, Christophe Roche éd., Annecy, Institut Porphyre, pp. 1-22, 2007

Salton G. (1971). T*he SMART Retrieval System—Experiments in Automatic Document Processing.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

A propos des auteurs

Damas Luc

Equipe Condillac, Université de Savoie

Campus Scientifique, 73370 Le Bourget-Du-Lac

Luc.damas@univ-savoie.fr

http://www.condillac.org

Tricot Christophe

Société Onomia

36 rue du clos d'Orléans, 94120 Fontenay-Sous-Bois

christophe.tricot@onomia.com

http://www.onomia.com

Modélisation des dénominations ontologiques

Benjamin Diemert, Marie-Hélène Abel, Claude Moulin

Résumé : L'échange de l'information numérique et sa réutilisation dans de nouveaux contextes pose un problème d'écart des représentations entre le producteur et le récepteur. La production télévisuelle se tourne de plus en plus vers des modes de production collaboratifs incluant des amateurs ne maîtrisant pas le jargon audiovisuel. Dans cet article, nous introduisons un modèle conceptuel capable d'articuler différentes dénominations autour d'un même élément ontologique (concept, instance, propriété). Nous définissons notre motivation à partir d'exemples d'échange d'information entre professionnel et amateur puis montrons comment le même problème s'applique également à la recherche multilingue. Nous rappelons l'apport des normes OAIS, FRBRoo et du modèle AXIS à la conceptualisation de l'échange d'information et les manques concernant la modélisation des écarts d'usages de la langue entre différentes communautés. Alors que ces modèles se concentrent sur l'échange de données, nous proposons de décrire et relier l'information, l'utilisateur et son contexte d'action. A partir d'une description d'un utilisateur et de son contexte, nous pouvons alors sélectionner la dénomination qui sera la plus adaptée. La définition de notre modèle permettra de mettre en évidence la généralisation de ce principe à des documents. Nous finissons par discuter nos choix de représentation en OWL.

Mots-clés : Représentation des connaissances, terminologie, archivage numérique, dénomination

1. Introduction

La numérisation massive des contenus et leur mise en réseau ont considérablement augmentées la quantité de documents et de connaissances disponibles dans les organisations. Cependant, si la numérisation doit rendre les contenus plus manipulables et la mise en réseau les rendre plus accessibles, aucun de ces deux procédés n'a pour autant facilité leur exploitation (Abel, 2009).

Le milieu de la production télévisuelle ne fait pas exception. L'exploitation des contenus y est d'autant plus ardue que le processus de création est fortement segmenté. De la définition de l'intention éditoriale à la production en passant par le tournage de scènes et leur montage, chaque étape fait appel à des métiers très spécialisés et donc à des savoir-faire très différents. Remarquons que le matériel et les documents produits à une étape sont réutilisés lors des étapes suivantes.

Par ailleurs, les systèmes informatiques qui ont été développés jusque là répondent à des besoins très divers et se concentrent essentiellement sur le traitement du contenu lui-même. Ils négligent la préservation des informations nécessaires au bon déroulement de la production, ainsi qu'à l'archivage des contenus en vue d'une réutilisation ultérieure. En particulier, manquent les métadonnées décrivant la structure du contenu, le contexte de production, les intentions éditoriales ou en encore les droits concernant les personnes filmées.

Le peu de descriptions présentes est produit de diverses manières. Certaines descriptions sont générées automatiquement par les systèmes de la chaîne de production (coordonnées GPS, caractéristiques de l'objectif d'une caméra). D'autres sont extraites par analyse automatique du contenu (détection de changements de plan, retranscription écrite d'un dialogue par reconnaissance vocale). Certaines enfin sont renseignées manuellement par les différentes communautés d'acteurs de la chaîne de production. Actuellement, ces descriptions sont perdues ou cloisonnées dans des formats ou des schémas de description propriétaires (Gervais et al., 2006, p.7). Le manque d'intégration des systèmes amène à une déperdition de l'information au fur et à mesure de l'avancement dans la chaîne de production (Delvin, 2002).

Il apparaît nécessaire de concevoir un système qui permette à chaque étape d'encapsuler dans une entité commune à la fois le contenu et les descriptions qui seront nécessaires pour une exploitation future, soit durant une étape ultérieure du même processus de production, soit lors d'une recherche dans une archive d'émissions.

Notons que la diversité des communautés présentes amène une diversité de vocabulaires de description même si elles partagent les mêmes concepts. Ces écarts d'usages de la langue posent un problème d'articulation des dénominations (signifiant) au sens des terminologies (Roche, 2005).

Il est donc indispensable de construire un modèle capable d'intégrer ces vocabulaires quel que soit la syntaxe ou le jargon employé. Une conceptualisation avancée, sous forme d'ontologies par exemple, pourra convenir à condition qu'elle prenne en compte les différents contextes d'usages de ces vocabulaires. Le cas du multilingue sera traité comme un cas à part de conceptualisation plus générique.

Dans cet article, nous introduisons les avantages d'une modélisation des dénominations pour l'échange d'information entre communautés ne partageant pas le même jargon ainsi que pour la recherche multilingue (Section 2). Nous présentons ensuite les apports et les manques des modèles dont nous nous inspirons pour notre travail (Section 3). Nous définissons ensuite le modèle et ses principes et reprenons les exemples précédents pour illustrer son utilisation (Section 4). Nous finissons par discuter certains points de notre représentation OWL du modèle (Section 5).

2. Motivations

Ce travail s'effectue dans le cadre du projet MediaMap[1] qui utilise les technologies du Web Sémantique pour favoriser la production collaborative et la réutilisation de contenu audiovisuel. Le consortium comporte notamment les chaînes de télévision publiques belges[2] qui contribuent à définir les besoins du projet. De ce fait, la problématique du multilinguisme est primordiale pour notre approche.

De plus, l'évolution du marché pousse fortement ces organisations à s'orienter vers des modes de production collaborative dans lesquels les sources de contenu se diversifient (tournage, archivage interne, achat). La description, la recherche, l'intégration de contenus provenant de partenaires différents et même d'amateurs deviennent des enjeux cruciaux. L'échange de contenu et de documents ne se fait plus seulement entre différents métiers d'une même organisation, mais entre différents métiers de différentes organisations et même

[1] http://www.mediamapproject.org/

[2] La Radio Télévision Flamande (VRT), la Radio Télévision Belge Francophone (RTBF).

avec des amateurs. Ainsi, l'indexation au cours du processus de production devient nécessaire pour supporter ces échanges.

2.1. Collaboration entre amateur et professionnels

L'appel à la contribution amateur pour la production de contenu professionnel pose en particulier un problème de qualité du contenu. Une solution est de guider l'amateur dans sa tâche à partir d'indications de tournage écrites par un réalisateur dans le jargon de l'audiovisuel. Ces indications sont transmises et traduites en termes compréhensibles par l'amateur. Professionnels et amateurs forment deux communautés distinctes, la première produit l'information, la seconde y accède en vue d'effectuer une tâche. Un exemple d'indication serait un cadrage particulier pour une interview. Voici différentes valeurs de plan utilisées classiquement (voir FIG.1).

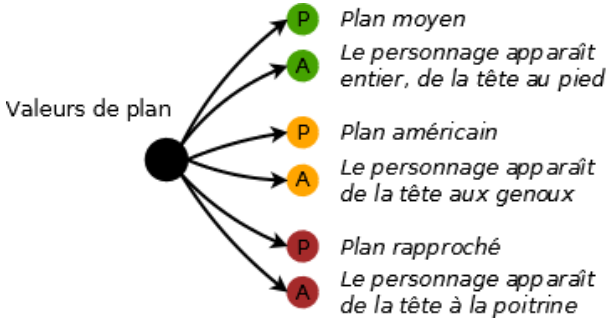


FIG.1 – Valeurs de plan : dénomination dans le jargon Professionnel (P) et pour des Amateurs (A)

Le même concept s'exprime de deux manières, la première (P) est utilisée par les professionnels, la seconde (A) par les amateurs. La distinction entre les deux dénominations à disposition se fait sur le critère de la connaissance du jargon de l'audiovisuel qui définit une communauté par rapport à l'autre.

2.2. Recherche d'information multilingue

De la même manière, on peut considérer le problème du multilinguisme (Ghebghoub, 2009), (Moulin, 2009) en caractérisant les variations orthographiques d'un même mot ou d'une même instance. Par exemple, un nom propre peut avoir différentes orthographes reconnues et chacune sert à identifier la même personne. Le nom du compositeur du Lac des Cygnes possède des dizaines de variantes dont voici trois exemples : Pyotr Ilyich Tchaikovsky (orthographe anglaise); Piotr Ilitch Tchaïkovski (orthographe française); (orthographe russe, voir FIG.1). Imaginons un utilisateur français lançant une

requête sur le compositeur. Du fait de l'orthographe utilisée il risque de ne pas voir les résultats indexés en anglais ou en russe (sans compter les autres variantes) même s'il parle ces langues.

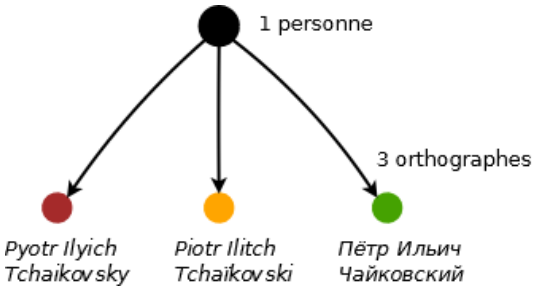


FIG.1 – Différentes orthographes pour le nom du compositeur du Lac des Cygnes

Il apparaît important d'identifier ce qui permet de distinguer ces trois orthographes. Si les deux premières partagent le même alphabet latin, la dernière version est écrite en cyrillique. Si l'on veut différencier les deux premières versions, il faut s'intéresser à la langue. Dans d'autres cas il faut aussi s'intéresser à la localisation géographique (orthographe française différente de l'orthographe québécoise, orthographe anglaise différente de l'américaine etc.).

3. Travaux liés

Parmi les travaux que nous avons examinés, le modèle « Acquisition, eXchange, Indexing and Structuring » (Axis) nous offre la meilleure base conceptuelle pour décrire la production de contenu audiovisuel.

Axis est un modèle et une architecture développés dans le cadre du projet Memories[3] pour traiter le problème de l'échange et l'archivage de ressources média. Axis se veut une implémentation de la norme « Open Archive Information System » (OAIS) qui pose un certain nombre de principes pour l'archivage numérique. Axis se sert de ces principes d'archivage pour résoudre les problèmes d'échanges de ressources entre systèmes différents.

[3] http://www.memories-project.eu/

3.1. L'archivage numérique et l'échange d'information

L'archivage peut être défini dans une perspective documentaire comme une activité intentionnelle de constitution et préservation d'une mémoire. Sa réalisation comprend deux activités indissociables : (1) la conservation physique des documents, (2) la préservation de la tradition de lecture liée aux documents qui assure ainsi leur compréhension (Bachimont, 2000). Ces deux activités sont couvertes à différents degrés par la norme OAIS. Nous présentons d'abord la manière dont OAIS pose le problème de l'archivage puis explicitons l'organisation des échanges d'information.

a) L'archivage comme service d'échange

La norme OAIS définit l'archivage numérique comme un problème d'échange de données entre une communauté productrice et une communauté demandeuse ou réceptrice (CCSDS, 2002). Elle formule des recommandations pour la mise en place d'une organisation rendant un service traitant ce problème (voir FIG.3).

L'enjeu d'un tel service est avant tout d'anticiper les changements de technologies de représentation de l'information dans l'un et l'autre monde. OAIS recommande de conserver certaines informations pour assurer la transformation des données d'une représentation à l'autre, suivant les apports des producteurs et les besoins des consommateurs. Il s'agit donc :

- de préserver les données de toute altération matérielle
- de documenter les données afin de permettre leur utilisation et leur compréhension par la communauté réceptrice.

Notons que cet objectif correspond à une certaine interprétation de la définition de (*Bachimont* 2000). Il faut comprendre ici que la tradition de lecture n'est pas seulement considérée comme la possibilité de voir le document (accès à une forme matérielle de restitution) mais également de le comprendre (accès à une forme sémiotique intelligible) et de l'exploiter pour son usage propre.

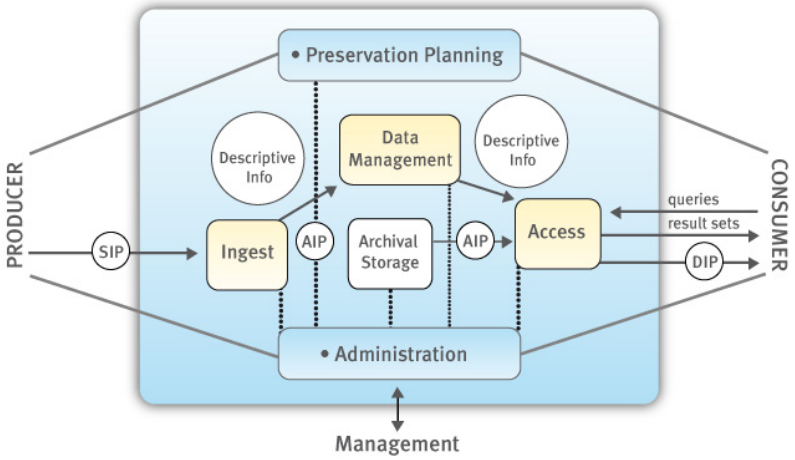


FIG.3 – Un système d'archivage OAIS, organisation fonctionnelle et flux de données

Pour jouer ce rôle d'intermédiaire entre deux mondes et gérer les flux de données la norme pose ainsi une organisation fonctionnelle des échanges et de l'archivage ainsi qu'un format d'encapsulation des données, les Information Package (IP).

b) Gérer l'échange de données

Nous détaillons l'organisation fonctionnelle proposée par OAIS (voir FIG. 3). D'abord l'*Ingest* réceptionne les informations fournit par le *Producer*. Il les transforme en « Archival Information Package » (AIP) répondant aux standards définit par l'*Administration* est constitué à partir des « Submission IP » (SIP).

Le contenu de l'AIP est pris en charge par l'*Archival Storage*, sa description par le *Data Management*. Le contenu et sa description sont donc stockés séparément du fait de besoins opérationnels différents. En effet, la recherche d'information se fait à partir des descriptions (elles sont donc fréquemment consultées) alors que l'accès au contenu se fait uniquement au moment de la livraison ou lors de contrôle d'intégrité.

L'*Access* utilise ces descriptions pour répondre aux recherches de la communauté de consommateurs. Au moment de la livraison, on prépare un « Dissemination IP » (DIP) à partir de l'AIP suivant les exigences des consommateurs. Ainsi, chacun des IP répond à des exigences différentes (la livraison, l'archivage, la diffusion) ce qui nécessite parfois des transformations

(formats) ou des réarrangements (composition). Ces modifications des IP ne sauraient se faire sans un modèle d'information qui structure et documente le contenu.

c) Documenter les données

Le modèle d'information préconisé dans OAIS distingue quatre types d'information :

- *Description du contenu* : Le contenu à archiver doit être accompagné d'une description qui permette de traduire les séquences de bits en information intelligible pour la communauté réceptrice. La description porte sur les structures de données, sur ce qu'elles représentent et les connaissances nécessaires pour interpréter correctement le contenu. Pour un fichier PDF contenant une partition musicale, il faut d'une part définir le format PDF pour accéder à la partition (accès à une forme matérielle) et d'autre part définir la notation musicale (accès à une forme intelligible).
- *Information pour la préservation* : La préservation du contenu passe avant tout par une identification robuste de tout ou partie du contenu ainsi qu'une description de son cycle de vie. En partant du contexte de production qui permet d'éclairer les intentions du producteur, un historique des transformations doit être tenu. Des indicateurs de qualité doivent être définis pour tester l'intégrité des données au bit près si nécessaire (identification, gestion du cycle de vie).
- *Description du contenu d'un IP* : L'information décrivant le tout et les parties du contenu à archiver en vue d'une recherche d'information.
- *Description de l'organisation d'un IP* : L'information indiquant le lien entre chaque partie de l'IP et un emplacement sur un support physique.

d) Manques

OAIS permet d'identifier des exigences d'organisation fonctionnelle et de documentation des données. Cependant, aucun modèle de données n'est proposé. De plus, la norme n'aborde pas la question des différences de pratiques des communautés que sous l'angle de la représentation de l'information. La question des usages de la langue et de la représentation des connaissances des communautés est simplement mentionnée comme étant à traiter.

3.2. FRBRoo

« Functional Requirements for Bibliographic Records » (FRBRoo) est un modèle conceptuel développé pour faciliter l'échange d'information entre les librairies numériques et les musées (CIDOC, 2009). Il permet de représenter les personnes participant aux différentes étapes de construction d'un objet culturel, depuis l'idée jusqu'à la réalisation matérielle. Chaque objet possède trois types de caractéristiques :

- Les idées abstraites (Work) n'ayant pas pris corps dans une matérialité externe à un sujet (une mélodie ou une histoire).
- Les différentes expressions (Expression) possibles pour une idée (une nouvelle écrite, ses traductions, une adaptation de la nouvelle en scénario etc.).
- Les porteurs physique d'information (Information Carrier) qui portent les expressions (livre, partition, cd-rom etc.). Dans ce cas, le modèle distingue entre l'original (Manifestation Singleton), des copies manufacturées (Item) issues d'un modèle de publication (Manifestation Product Type).

3.3. Axis

Axis reprend les exigences d'OAIS tout en les adaptant pour l'échange d'information et de ressources médias entre organisations. Les conséquences sont multiples. D'abord l'architecture du système se doit d'être flexible et distribuée. Chaque organisation peut remplir une ou plusieurs fonctions, parfois de manière redondante (*Ingest*, *Access*).

Ensuite, Axis propose de n'utiliser qu'un seul type de paquet d'information, l'« Autonomous eXchange Entity » (AXE), pour permettre l'échange. La fusion des paquets doit répondre aux exigences de l'archivage et du même coup à celui de l'interopérabilité entre systèmes.

La construction du modèle Axis s'inspire également du modèle conceptuel FRBRoo pour décrire les ressources médias comme des objets culturels. Le modèle Axis se réapproprie ces concepts en les considérant comme trois niveaux distincts ; le niveau conceptuel, le niveau des signes, le niveau des supports physique. Expression et Manifestation deviennent ainsi des relations entre objets distincts. Ainsi pour une idée, on peut associer différentes expressions sémiotiques, chacune pouvant être inscrites sur un ou plusieurs supports.

Au niveau des signes, Axis utilise le principe des thésaurus de manière similaire au standard du W3C « Simple Knowledge Organization System » (SKOS). Une approche de représentation de type thésaurus permet en effet de relier chacune des dénominations possibles (alternatives) à une dénomination de référence. SKOS par exemple permet de les grouper en schémas et de représenter les relations entre dénominations (spécificité, généralité, utilisé pour). La modélisation en OWL Full de SKOS permet de représenter les thésaurus existants et de les intégrer au Web Sémantique avec toutes les possibilités d'alignement qui s'en suivent (Summers et al., 2008).

Cependant cette approche ne permet pas de caractériser les contextes d'usages des dénominations. Le seul attribut disponible est « xml:lang » qui permet d'ajouter le code d'une langue ce qui permet de couvrir le cas du multilingue, mais pas du tout le cas des jargons métiers (voir Section 2). Une solution consiste à avoir un schéma par contexte d'usage et de les relier manuellement.

4. Modélisation

Nous présentons d'abord les principes généraux qui fondent notre modélisation (Section 4.1), puis nous montrerons la modélisation des exemples précédemment décrits (Section 4.2). Nous finirons par une définition détaillée des concepts et relations de notre modèle (Section 4.3).

4.1. Principes

Nous avons trouvé dans Axis un modèle générique qui partage une bonne partie des problématiques du projet MediaMap. Les concepts et les principes d'Axis servent de fondation à notre travail qui s'étend à la modélisation des contextes d'usages des dénominations (signifiant) attachées à un même concept, une même propriété ou une même instance. Notre approche tire partie de façon intéressante des modèles existants.

Notre modélisation permet de faire un choix informé parmi différentes dénominations à disposition. Elle intègre la description de la situation de production et d'accès à l'information en prenant en compte l'utilisateur et son contexte. Ces éléments contextuels servent d'une part à identifier des usages de la langue, et d'autre part à décrire un utilisateur et le relier à des usages identifiés. De cette manière, on peut proposer automatiquement une dénomination adaptée à un utilisateur connu.

Le modèle se divise en trois parties (voir FIG. 4). La première définit les capacités de l'utilisateur (Modèle Utilisateur). La deuxième regroupe les dénominations selon leurs caractéristiques (Modèle d'Information). La dernière est composée d'éléments généraux qui permettent de spécifier et de relier le contexte d'usage des dénominations au contexte d'action d'un utilisateur (Modèle du Contexte).

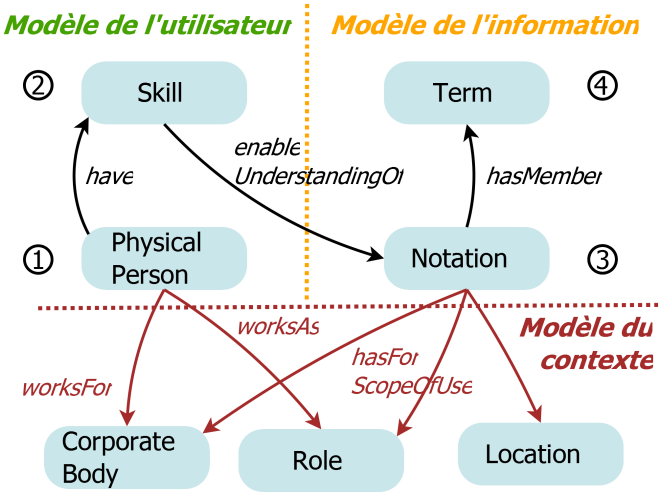


FIG. 4 – Concepts principaux et Relations entre Modèle Utilisateur ; Modèle d'Information ; Modèle du Contexte

Ainsi, la sélection d'une dénomination (instance de Term) se fait d'abord en fonction des capacités (instance de Skill) d'un utilisateur (instance de PhysicalPerson). Ces capacités permettent de comprendre un jargon métier, une langue ou un codage (instance d'une spécialisation de Notation). D'autres éléments contextuels peuvent permettre de sélectionner plus finement une dénomination et d'identifier une communauté d'usage de la langue. Par exemple l'appartenance à une organisation (instance de CorporateBody) qui a défini un thésaurus métier. Le métier de l'utilisateur (instance de Role) peut également apporter des indications précieuses sur le vocabulaire et les codes utilisées. Le lieu de vie ou le lieu de travail (Location) permet également d'identifier l'appartenance à une communauté linguistique (alphabet, langue).

Finalement, il faut préciser que le modèle peut s'étendre à la caractérisation de tous types de support d'information (instance de Document).

4.2. Modélisation des exemples

Nous présentons une vue simplifiée de la modélisation des exemples présentés en Section 2. Leurs représentations en OWL sont détaillées en Section 5. D'abord, les dénominations de Valeur de Plan pour amateurs et professionnels. La FIG. 5 présente une vue simplifiée du réseau des relations entre deux utilisateurs et deux dénominations différentes pour désigner la même valeur de plan.

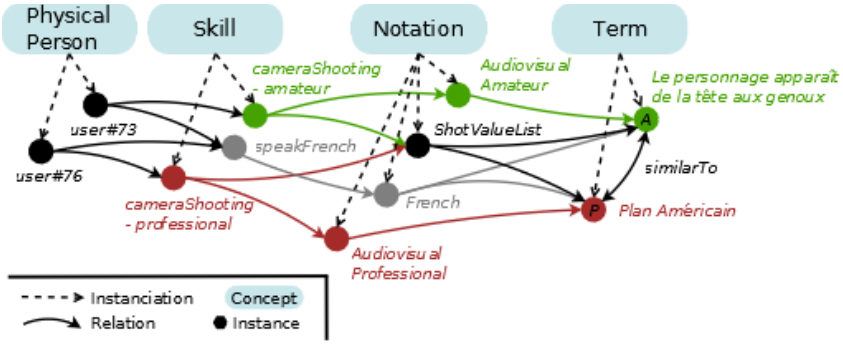


FIG. 5 – Vue simplifiée du réseau de relations entre deux utilisateurs et deux dénominations d'une même valeur de plan

L'exemple considère deux caméramans francophones, l'un amateur, l'autre dont c'est le métier. Tous deux doivent donc comprendre les indications de tournage qu'on leur donne, dont la liste des valeurs de plan. On les distingue par leur niveau de compétence (amateur, professionnel) qui se traduit par l'accès à deux vocabulaires différents (instance de Notation) pour désigner les valeurs de plan. Chaque valeur de plan de la liste est exprimée par deux dénominations distinctes (instance de Term) appartenant à l'un ou l'autre vocabulaire. Comme il n'y a pas un concept par valeur de plan, mais juste des dénominations, une relation de similarité est nécessaire pour regrouper celles qui correspondent à la même valeur de plan.

L'exemple de Tchaïkovski (voir FIG. 6) permet de montrer un exemple de recherche d'information multilingue. Supposons que la requête d'un utilisateur soit exprimée en français et aboutissent à retrouver le compositeur (instance de Personne). Dans ce cas, tous les documents (A, B, C) relatifs au compositeur seront retrouvés. Grâce à la description des capacités linguistiques de l'utilisateur, le système pourra mettre en avant ceux qui lui sont compréhensibles.

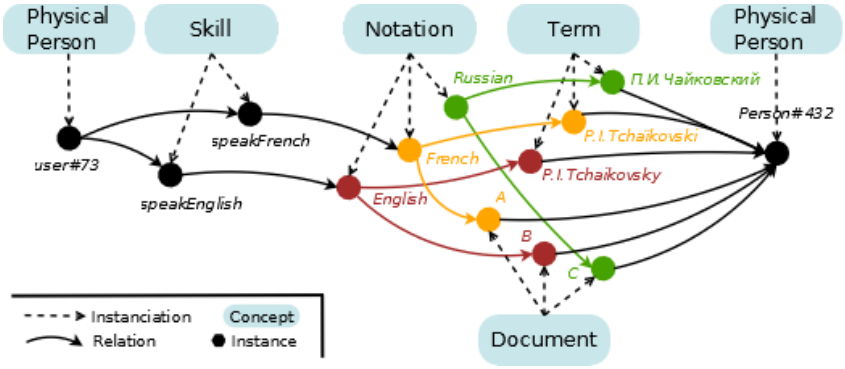


FIG. 6 – Recherche multilingue et sélection de document

On peut complexifier l'exemple en rajoutant à la description des documents l'encodage des caractères (ISO 8859-1 pour l'alphabet latin, ISO 8859-5 pour le cyrillique). Cet ajout peut servir lorsqu'on décrit également le terminal d'accès utilisé par un utilisateur et que l'on renseigne ses capacités à lire tel ou tel encodage.

4.3. Description de la conceptualisation

Nous définissons les concepts et les relations des différentes parties de notre modèle. Nous finissons par présenter les relations qui permettent de relier un utilisateur à des dénominations ou d'autres types d'information.

a) Concepts et Relations du Modèle d'information

Un terme (instance de la classe **Term**) encapsule la valeur d'une propriété qui dans une modélisation commune serait généralement de type chaîne de caractères. Nous lui adjoignons un ensemble de caractéristiques qui précisent son contexte d'utilisation. Le véritable type de la valeur peut en fait être quelconque : une chaîne, mais aussi un nombre, une date, etc.

Un document (instance de la classe **Document**) désigne un support d'information en relation avec un concept. Les concepts sont des entités abstraites qui ne sont accessibles qu'à travers des manifestations physiques que sont les documents. Ainsi contrairement à FRBRoo, une manifestation est une relation (**isAManifestationOf**) entre un document et une entité.

Une notation (instance de la classe **Notation**) dénote un ensemble de termes partageant des caractéristiques communes. La classe Notation possède un ensemble de spécialisations que nous détaillons ci-dessous.

- Un langage naturel (instance de la classe **NaturalLanguage**) désigne l'appartenance d'un terme à un langage naturel, (français, anglais, flamand etc.). Le code d'un pays modélise les variations géographiques d'une langue.
  Par exemple l'anglais est parlé dans différentes régions du monde et des distinctions existent entre l'anglais du Royaume-Uni et celui des Etats-Unis d'Amérique.

- Un schéma d'encodage syntaxique (instance de la classe **SyntaxEncodingScheme**) désigne un codage particulier d'octets permettant de représenter une valeur.
  Ce peut être un codage de caractères (UTF-8), un encodage de type de données (xsd:string), un codage avec une syntaxe complexe tel un format de fichier (jpg, pdf) ou des formats de dates (ISO 8601 : 1988 (E)[4]).

- Un schéma d'encodage de termes d'un vocabulaire (instance de la classe **VocabularyEncodingScheme**) désigne un ensemble d'appartenance de termes tel qu'un thésaurus métier. Cela permet plus généralement de modéliser une convention de nommage dans un contexte métier, organisationnel ou géographique. Un schéma de description de métadonnées tel que Dublin Core[5] est un exemple.

- Une liste d'autorité (instance de la classe **AuthorityList**) désigne un ensemble fini de termes dont les valeurs sont arbitrairement choisies. Ainsi, la norme « ISO 639-2[6] » est une liste de codes à trois lettres dénotant le nom des langues. Par exemple, « eng » pour l'anglais.

Un **terme** est membre (**isMemberOf**) d'une notation lorsqu'il remplit les caractéristiques qui la définissent. L'appartenance à une notation n'est pas exclusive ce qui permet de raffiner les descriptions jusqu'au niveau de détails souhaité. Cette relation se spécialise par la relation d'appartenance à une langue (**inLanguage**).

Une **notation** se conforme à (**conformsTo**) à un schéma d'encodage syntaxique (SES) lorsqu'elle utilise ses règles de codage. Une notation peut se

[4] Les formats de date conseillés par le W3C, http://www.w3.org/TR/NOTE-datetime

[5] http://dublincore.org/documents/dcmi-terms/

[6] http://www.loc.gov/standards/iso639-2/

conformer à plusieurs schémas ce qui permet là encore de détailler la description à souhait.

b) Concepts et Relations du Modèle utilisateur

Un agent (instance de la classe **Agent**) désigne une entité agissante. Cette classe possède trois spécialisations :

- Une personne (instance de la classe **PhysicalPerson**) désigne un individu, « Tchaïkovski » par exemple.
- Une personne morale (instance de la classe **CorporateBody**) désigne une organisation composée d'autres agents, le « Minneapolis Symphony Orchestra ».
- Un système (instance de la classe **Proxy**) désigne un système agissant pour le compte d'un autre agent, un serveur web.

Une compétence (instance de la classe **Skill**) désigne la capacité d'un agent à effectuer une tâche ou à comprendre une information en rapport avec cette tâche (indications de réglage, mode d'emploi pour un nouvel instrument etc.). Cette classe possède deux spécialisations :

- Une compétence linguistique (instance de la classe **LinguisticSkill**) désigne la capacité d'un agent à comprendre une langue.
- Une compétence professionnelle (instance de la classe **BusinessSkill**) désigne la capacité d'un agent à effectuer des tâches professionnelles et à comprendre des notations métiers.
  « SavoirCadrer » désigne ainsi la compétence d'une personne à régler une caméra en fonction d'une indication de valeur de plan telles que plan américain, gros plan, plan d'ensemble. Ces différentes valeurs étant des termes regroupés dans une liste d'autorité.

Une personne est définit comme compétente (**isProficient**) lorsqu'on lui reconnaît cette capacité.

c) Concepts et Relations du Modèle du Contexte

Les compétences permettent de caractériser des métiers (instance de la classe **Role**). Un rôle désigne alors les droits et les responsabilités que l'on attribue à une personne ou une organisation dans un projet ou un processus.

Une personne travaille (**workFor**) pour une organisation lorsque cette dernière l'emploie contractuellement. De même, un emploi implique qu'on attribue (**workAs**) une position ou un métier à cette personne.

Un lieu (instance de **Location**) désigne une entité géographique ou politique. Location est un concept général qui peut être spécialisé suivant les besoins.

d) Relations entre parties du modèle

Il y a trois relations très importantes dans le modèle. Celle qui permet de relier l'utilisateur à des notations, celle qui relie les notations à des éléments contextuels et enfin celle qui permet de relier un terme à un élément ontologique :

- Une compétence désigne la capacité d'un agent à comprendre une notation (**enablesUnderstandingOf**)
- Une notation a pour contexte d'usage (**hasForScopeOfUse**) une organisation (instance de CorporateBody), un lieu (instance de Location) ou un type de métier (instance de Role).
- La relation de nommage (**isNamedBy**) permet d'attacher un terme à n'importe quel autre élément du modèle. Ainsi, on généralise les mécanismes de caractérisations des dénominations aux concepts, instances, propriétés.

5. Représentation OWL

La représentation de notre modèle sous forme d'ontologie nous semble profitable sur plusieurs points. Tout d'abord le fait d'expliciter la sémantique des concepts utilisés permet de faciliter la compréhension du modèle, donc son appropriation par les utilisateurs et sa maintenance pour les concepteurs. De plus, le fait de modéliser des objets d'un domaine sans concentrer son attention sur une tâche particulière permet d'obtenir des conceptualisations plus flexibles et extensibles (Spyns et al., 2002). Lorsqu'il s'agit ensuite d'effectuer des recherches sur les données produites à partir d'une ontologie, on mesure les bénéfices d'un tel détachement grâce aux extensions de requêtes (Guarino, 1998). Enfin, les capacités de raisonnement et d'intégration de données entre modèles différents sont des avantages particulièrement précieux lorsqu'on traite une problématique d'échange d'information (Garcia et al., 2005).

Les langages de représentation de connaissances telle que RDF et OWL nous offrent une manière d'exprimer notre ontologie sous une forme répandue et exploitable. Nous avons choisi d'utiliser la variante OWL-DL pour s'assurer une expressivité forte sans pour autant sacrifier à la calculabilité. Nous discutons de certains détails à partir des exemples présentées en Section 2 et modélisé en Section 4.2.

5.1. Dénominations pour amateur et professionnel

Voici la modélisation de l'exemple de la liste de Valeur de Plan en notation N3[7]. Une liste d'autorité a pour membre des instances de termes dont certains renvoient à la même valeur (un plan américain). Cependant cette valeur n'est pas représentée en tant que telle, elle n'existe que par les termes qui appartiennent à l'un ou l'autre vocabulaire (amateur, professionnel).

Il faut donc pouvoir d'une part caractériser chaque terme de la liste individuellement (par les relations isMemberOf, inLanguage) et d'autre part identifier la similarité entre deux termes (par la relation similarTo).

```
// ==== NOTATIONS ===== //
mm:shotValueList
    a      mm:AuthorityList ;
    mm:hasMember mm:planAmericain-a , mm:planRapproche , mm:planRapproche-a ,
mm:planAmericain .
mm:Audiovisual-Professional
    a      mm:VocabularyEncodingScheme ;
    mm:hasMember mm:planAmericain , mm:planRapproche .
mm:Audiovisual-Amateur
    a      mm:VocabularyEncodingScheme ;
    mm:hasMember mm:planAmericain-a , mm:planRapproche-a .
mm:French-NL
    a      mm:NaturalLanguage .

// ==== TERM ===== //
mm:planAmericain
    a      mm:Term ;
    mm:inLanguage mm:French-NL ;
    mm:isMemberOf mm:shotValueList , mm:Audiovisual-Professional ;
    mm:similarTo mm:planAmericain-a ;
    mm:value "Plan Américain"^^xsd:string .
mm:planAmericain-a
    a      mm:Term ;
    mm:inLanguage mm:French-NL ;
    mm:isMemberOf mm:shotValueList  , mm:Audiovisual-Amateur ;
    mm:similarTo mm:planAmericain ;
    mm:value "Plan qui fait apparaître le personnage de la tête jusqu'aux genoux."^^xsd:string .
```

[7] http://www.w3.org/DesignIssues/Notation3.html

La description des utilisateurs se fait par les compétences et leur niveau. Le niveau de compétence est un attribut qui peut prendre deux valeurs : amateur ou professionnel. Ainsi, chaque compétence peut avoir deux variantes qui renvoient à deux vocabulaires distincts (Audiovisual-Amateur, Audiovisual-Professional). Ces variantes permettent également d'accéder à des notations communes, ici une liste d'autorité de valeur de plan.

```
// ==== PHYSICAL PERSON ===== //
mm:user_73
    a      mm:PhysicalPerson ;
    mm:isProficient mm:cameraShooting-a , mm:speakFrench .
mm:user_76
    a      mm:PhysicalPerson ;
    mm:isProficient mm:cameraShooting-p , mm:speakFrench .

// ==== SKILLS ===== //
mm:speakFrench
    a      mm:LinguisticSkill ;
    mm:enablesUnderstandingOf
          mm:French-NL .

mm:cameraShooting-p
    a      mm:BusinessSkill ;
     mm:enablesUnderstandingOf
          mm:Audiovisual-Professional , mm:shotValueList ;
    mm:SkillLevel "professional"^^xsd:string .
mm:cameraShooting-a
    a      mm:BusinessSkill ;
    mm:enablesUnderstandingOf
          mm:Audiovisual-Amateur , mm:shotValueList;
    mm:SkillLevel "amateur"^^xsd:string .
```

5.2. Dénominations et recherche d'information multilingue

L'exemple avec Tchaïkovski permet de discuter de la représentation en OWL de la relation de nommage (isNamedBy). Il s'agit nécessairement d'une propriété d'annotation (AnnotationProperty) car elle doit pouvoir attacher un terme à n'importe quel élément ontologique (concept, propriété, instance). De ce

fait, elle se substitue à l'usage de rdfs:label tout en le généralisant. Notre choix de représentation en OWL-DL implique certaines contraintes[8] qui n'ont pas de conséquences pour l'usage que nous avons choisi. Les triplets RDF que nous construisons sont systématiquement de la forme suivante : « sujet mm:isNamedBy terme » avec comme sujet une classe, une propriété ou un individu et comme objet une instance de Term.

Voici la représentation des trois variantes de termes nommant Tchaïkovski, des documents se rapportant au compositeur (par la relation isAManifestationOf) et des langues dans lesquels le contenu des documents ou les dénominations sont exprimées (par la relation inLanguage).

```
// ==== NOTATIONS ===== //
mm:French-NL
    a      mm:NaturalLanguage .
mm:Russian-NL
    a      mm:NaturalLanguage .
mm:English-NL
    a      mm:NaturalLanguage .

// ==== TERM ===== //
mm:Tchaikovsky-EN-Term
    a      mm:Term ;
    mm:inLanguage mm:English-NL ;
    mm:value "Pyotr Ilyich Tchaikovsky"^^xsd:string .
mm:Tchaikovsky-FR-Term
    a      mm:Term ;
    mm:inLanguage mm:French-NL ;
    mm:value "Piotr Ilitch Tchaïkovski"^^xsd:string .
mm:Tchaikovsky-RU-Term
    a      mm:Term ;
    mm:inLanguage mm:Russian-NL ;
    mm:value "Пётр Ильич Чайковский"^^xsd:string .

// ==== DOCUMENT ===== //
mm:Document_A
    a      mm:Document ;
    mm:inLanguage mm:French-NL ;
```

[8] Dans la variante OWL-DL, il n'est pas possible de définir de domaine, portée et de sous-propriété à une propriété d'annotation.

```
    mm:isAManifestationOf
        mm:Tchaikovsky .
mm:Document_B
    a      mm:Document ;
    mm:inLanguage mm:English-NL ;
    mm:isAManifestationOf
        mm:Tchaikovsky .
mm:Document_C
    a      mm:Document ;
    mm:inLanguage mm:Russian-NL ;
    mm:isAManifestationOf
        mm:Tchaikovsky .
```

Une instance de personne peut se nommer d'autant de manières souhaitées (par la relation isNamedBy). Un utilisateur lançant une requête à partir de l'orthographe française du compositeur aboutit à retrouver la personne et tous les documents associés. Suivant la description des documents et les compétences de l'utilisateur, un système peut mettre en avant les contenus qu'il est le plus susceptible de comprendre (ici les documents en français et en anglais).

```
// ==== PHYSICAL PERSON ===== //
mm:Tchaikovsky
    a      mm:PhysicalPerson ;
    mm:  isNamedBy mm:Tchaikovsky-FR-Term , mm:Tchaikovsky-RU-Term , mm:Tchaikovsky-EN-
Term .
mm:user_73
    a      mm:PhysicalPerson ;
    mm:isProficient mm:speakEnglish , mm:speakFrench .

// ==== SKILLS ===== //
mm:speakFrench
    a      mm:LinguisticSkill ;
    mm:enablesUnderstandingOf
        mm:French-NL .
mm:speakEnglish
    a      mm:LinguisticSkill ;
    mm:enablesUnderstandingOf
        mm:English-NL .
```

Conclusion

Dans cet article nous avons décrit un modèle conceptuel permettant de caractériser la portée d'usage des dénominations se rapportant à tous types d'éléments ontologiques (concept, instance, propriété). Cette caractérisation se fait sur la base d'éléments contextuels (métier, organisation, lieu) qui servent de critères communs à une description de l'information et de l'utilisateur voulant y accéder. Ce lien permet de sélectionner la dénomination ou la forme d'information la plus pertinente pour une communauté d'utilisateur donné.

Nous avons d'abord montré la pertinence de cette approche dans le contexte d'une production télévisuelle collaborative où les contributions d'amateurs sont amenées à prendre de l'importance. La capacité à échanger de l'information et à la réutiliser dans de nouveaux contextes devient un enjeu important pour les organisations. Notre étude d'un modèle conceptuel d'archivage numérique permet de poser le problème comme celui d'un échange d'information entre communautés dont les représentations varient. Cependant, les efforts se concentrent sur l'échange matériel des données plutôt que sur la représentation des connaissances et des usages de la langue nécessaires à leur interprétation. L'étude de modèles de représentation des objets culturels et des ressources médias nous a permis de poser une distinction entre concept, signe et porteur physique d'information. Cette distinction nous a servi de base pour le développement de notre modèle qui aboutit à une ontologie en OWL-DL.

Notre modèle permet de sélectionner la forme d'information la plus pertinente pour un utilisateur souhaitant y accéder. La recherche d'information ainsi que la compréhension du contenu s'en trouve facilitée. Nous travaillons actuellement à la modélisation du processus de production télévisuelle pour affiner notre modèle du contexte. L'objectif est de faciliter la communication entre les différents acteurs participant au processus de production en précisant les informations qu'ils produisent et qu'ils échangent.

Bibliographie

Abel M., Leblanc A. (2009) : *Knowledge Sharing via the E-MEMORAe2.0 Platform,* In 6th International Conference on Intellectual Capital, Knowledge Management & Organisational Learning (pp. pp. 10-19). Montreal, Canada.

Bachimont Bruno (2000) : *L'archive numérique: entre authenticité et interpretabilité.* archivistes.qc.ca

Centre Internation de la DOCumenation (CIDOC) (2009) : *FRBR object-oriented definition and mapping to FRBRer*, http://cidoc.ics.forth.gr/frbr_inro.html.

Roche C. (2005) : *Terminologie & Ontologie,* Langages, 157, 1-11.

Consultative Committee for Space Data Systems (CCSDS) (2002) : *Reference Model for an Open Archival Information System*,
http://public.ccsds.org/publications/archive/650x0b1.pdf

Devlin B. (2002) : *What is MXF?,* Ebu Technical Review, (July), 1-7.

Garcia R., Celma O. (2005) : *Semantic Integration and Retrieval of Multimedia Metadata,* In 5th International Workshop on Knowledge Markup and Semantic Annotation (SemAnnot'05). Ireland.

Ghebghoub O., Moulin C., Abel M. (2009) *: Création d'une ontologie des ressources numériques, In* 3èmes Journées Francophones sur les Ontologies (JFO) (pp. 65-71). Poitiers, France.

Gouyet J., Gervais J. (2006) : *Gestion des médias numériques*, (INA) Audio-Photo-Video (p. 328). Paris, France: Dunod.

Moulin C., Sugawara K., Fujita S., Wouters L., Manabe Y. (2009) : *Multilingual Collaborative Design Support System,* In 13th International Conference on Computer Supported Cooperative Work in Design (CSCWD 2009) (pp. 312-318). Santiago, Chile.

Summers E., Isaac A., Redding C., Krech D. (2008) : *LCSH, SKOS and Linked Data,* CoRR.

A propos des auteurs

Diemert Benjamin (A)
benjamin.diemert@hds.utc.fr
http://www.hds.utc.fr/~bdiemert/perso/

Abel Marie-Hélène (A)
marie-helene.abel@hds.utc.fr
http://www.hds.utc.fr/~mhabel/perso/

Moulin Claude (A)
claude.moulin@hds.utc.fr
http://www.hds.utc.fr/~cmoulin/perso/

(A) Heudiasyc, UMR 6599
Université de Technologie de Compiègne
Centre de Recherches de Royallieu
BP 20529
60205 COMPIEGNE cedex FRANCE

Filtrage des Entités Nommées par des méthodes de Fouille de Textes

Mathieu Roche

Résumé : Cet article présente une approche de Traitement Automatique du Langage (TAL) afin de filtrer les Entités Nommées à partir d'une liste de candidats à la collocation. La méthode proposée s'appuie uniquement sur des mesures statistiques associées aux ressources du Web. L'évaluation à partir de candidats à la collocation de type Nom-Nom issus d'un corpus en français (corpus de CVs) permet de valider l'approche et de discuter des limites de cette dernière.

Mots-clés : Traitement Automatique du Langage (TAL), Fouille de Textes, Collocations, Entités Nommées

1. Introduction

Dans cet article, nous nous intéressons à l'étude des groupes de mots souvent appelés des collocations. Ces groupes peuvent être extraits par des méthodes de TAL (Traitement Automatique du Langage). Plus formellement, [Clas, 1994] donne deux propriétés définissant une collocation. Premièrement, une collocation est définie comme un groupe de mots ayant un sens global qui est déductible des unités (mots) composant le groupe. Par exemple, « lumière vive » est considéré comme une collocation car le sens global de ce groupe de mots peut être déduit des deux mots « lumière » et « vive ». En nous appuyant sur cette définition, l'expression « tirer son chapeau » n'est pas une collocation car son sens ne peut pas être déduit de chacun des mots. De telles formes sont appelées des **combinaisons figées**. Une deuxième propriété est ajoutée par [Clas, 1994] pour définir une collocation. Le sens des mots qui composent la collocation doit être limité. Par exemple « acheter un chapeau » n'est pas une collocation car le sens de « acheter » et de « chapeau » n'est pas limité. En effet, de multiples objets, voire des personnes, peuvent être achetés. De tels groupes de mots sont appelés des **combinaisons libres**. Notons cependant qu'il reste très difficile de différencier par des méthodes automatiques issues du TAL les locutions figées, libres et les collocations.

La définition générale des collocations étant donnée, elle peut être enrichie avec deux caractéristiques supplémentaires : les aspects sémantiques et syntaxiques [Heid, 1998; Laurens, 1999]. Le premier point s'appuie sur des caractéristiques sémantiques communes de certaines collocations. Par exemple, « lait tourné » et « beurre rance » ont des sens très proches liés à un phénomène de dégradation. Les aspects sémantiques définissant formellement les collocations sont pris en considération dans de nombreux travaux [Melcuk *et al.,* 1984-1999; Heid, 1998; Laurens, 1999]. Ainsi, [Melcuk *et al.,* 1984-1999] ont introduit les fonctions lexicales qui s'appuient sur des caractéristiques sémantiques pour définir les relations entre les unités des collocations. La deuxième caractéristique est liée à la structure syntaxique des collocations. A titre d'exemple, « lumière vive » et « marque distinctive » ont une même structure syntaxique de type Nom-Adjectif. Une classification de la structure syntaxique des collocations que nous donnons ci-dessous est proposée dans de nombreux travaux [Clas 94; Laurens 1999] : Nom-Verbe (par exemple, « interpréter un film »), Nom-Adjectif (par exemple, « cinéma muet »), Nom-Nom/Nom-Préposition-Nom (par exemple, « plateau de cinéma »), Verbe-Adverbe (par exemple, « boire goulûment »), Adverbe-Adjectif (par exemple, « gravement malade »).

Même si les méthodes de TAL ne permettent pas toujours d'identifier les collocations proprement dites qui s'appuient sur les définitions linguistiques énoncées, dans cet article, nous allons nous intéresser à l'extraction des *candidats à la collocation* qui respectent les patrons syntaxiques suivants : Nom-Nom, Nom-Préposition-Nom, Adjectif-Nom et Nom-Adjectif. La terminologie nominale de ce type est par exemple étudiée dans [Daille, 1996; Bourigault et Jacquemin, 1999]. Cependant, l'originalité de l'approche décrite dans cet article réside dans l'**identification automatique des *Entités Nommées*** à partir des candidats extraits.

Les Entités Nommées (EN)

Les EN sont classiquement définies comme les noms de Personnes, Lieux et Organisations. Initialement, une telle définition est issue des campagnes d'évaluation américaines MUC – *Message Understanding Conferences* qui furent organisées dans les années 90. Cette série de campagnes consistait à extraire des informations telles que les EN dans différents documents (messages de la marine américaine, récits d'attentats terroristes, etc). Aujourd'hui, de telles campagnes d'évaluation couvrent des tâches très variées sur la base de textes de différents domaines (textes spécialisés en biologie, dépêches d'actualités, blogs, etc). Nous pouvons, entre autres, citer les challenges TREC – *Text REtrieval Conference* (international) et DEFT – *DEfi Fouille de Textes* (francophone) qui sont aujourd'hui très actifs dans la communauté « fouille de textes ».

Comme le précisent [Daille *et al.,* 2000], les classes de base d'EN définies dans le cadre de MUC doivent être enrichies. Par exemple, outre les classes relatives aux Personnes, Lieux et Organisations, [Paik *et al.*, 1994] définissent de nouvelles classes telles que *Document* (logiciels, matériels, machines) et *Scientifique* (maladie, médicaments, etc).

Pour identifier les EN, de nombreux systèmes s'appuient sur la présence de majuscules [Daille *et al.,* 2000]. Cependant ceci peut se révéler peu efficace dans le cas d'EN non capitalisées et pour le traitement de textes non normalisés (mails, blogs, textes ou fragments de textes inégalement en majuscule ou minuscule, etc). A titre d'exemple, certaines données du défi DEFT'06 étaient constituées de discours politiques entièrement capitalisés [Azé *et al.,* 2006] (corpus disponible à l'adresse suivante : http://deft.limsi.fr/). Ainsi, nous avons choisi dans nos travaux de ne pas exploiter ce type d'informations pour identifier les EN. Notons cependant que de telles caractéristiques pourraient être intéressantes à associer à l'approche essentiellement statistique présentée dans cet article.

Plus formellement, pour caractériser les EN, les critères d'unicité référentielle (c'est-à-dire, un nom propre renvoie à une entité référentielle unique) et une stabilité dénominative (c'est-à-dire, peu de variations possibles) sont notamment précisées par [Fort *et al.*, 2009]. Nous allons nous appuyer sur ce dernier critère pour identifier les EN parmi les candidats à la collocation dont l'extraction est décrite dans la section suivante.

2. Méthode de fouille de textes pour l'extraction de candidats à la collocation

Nous présentons ci-dessous un processus automatique pour extraire les candidats à la collocation. Dans un premier temps, il est nécessaire de rassembler les textes à traiter. Ces textes devront être homogènes dans la spécialité étudiée (textes sur la biologie, les régimes politiques, etc). Nous appelons l'ensemble de ces textes des *corpus*. Après l'acquisition d'un corpus, l'étape suivante du traitement consiste à normaliser les textes. Cette tâche consiste par exemple à supprimer les caractères qui peuvent provoquer des erreurs dans les traitements automatiques (tirés d'énumérations, balises HTML, etc). Avec les textes normalisés nous pouvons appliquer un étiqueteur grammatical qui appose une étiquette grammaticale à chacun des mots du corpus. Par exemple, dans la phrase *L'ouvrier règle la machine outil*. Les mots *ouvrier, machine* et *outil* auront une étiquette Nom, le mot *règle* aura une étiquette Verbe. Ces étiquettes sont apposées en utilisant des systèmes d'étiquetage morpho-syntaxique automatique [Brill, 1994]. L'étude effectuée dans cet article concerne l'extraction des candidats à la

collocation à partir des textes étiquetés. Par exemple, avec le fragment étiqueté *L'/Article ouvrier/Nom règle/Verbe la/Article machine/Nom outil/Nom*, nous pouvons extraire le candidat à la collocation « *machine outil* » qui est de type Nom-Nom. Les candidats ainsi extraits sont utilisés pour des tâches précises : extraction d'informations, traduction automatique, classification de documents, etc. Notons que le filtrage grammatical ainsi appliqué permet de désambiguïser le mot « *règle* » qui peut avoir plusieurs rôles grammaticaux (nom, verbe).

La section suivante décrit une **méthode statistique** de sélection des EN à partir des candidats à la collocation obtenus après l'application du processus de Fouille de Textes.

3. Filtrage des Entités Nommées

Principe général

Il est fréquent que les candidats à la collocation de type Nom-Nom aient des formes variées comme nous le montrerons dans la section « Expérimentations » de cet article. Par exemple, la collocation « fichier clients » peut se décliner sous les formes Nom-Préposition-Nom : « fichier de clients », « fichiers pour clients », etc.

A contrario, les EN sont peu sujettes aux variations [Fort *et al.*, 2009] telles que les « variations prépositionnelles ». Nous allons nous appuyer sur cette constatation pour identifier avec des méthodes de TAL les EN nominales à partir d'une liste de candidats à la collocation de type Nom-Nom.

Pour cela, pour chaque candidat Nom-Nom, l'approche que nous décrivons dans cet article va consister à :

(1) Construire artificiellement une collocation prépositionnelle de type Nom-Préposition-Nom à partir du candidat Nom-Nom.

(2) Mesurer la « pertinence » de la collocation prépositionnelle construite en mesurant la dépendance entre chaque mot par des méthodes statistiques.

(3) Sélectionner les collocations prépositionnelles ayant des scores faibles (c.-à-d. collocation construites peu pertinentes). En effet, si les possibilités de variations du candidat Nom-Nom sont faibles, nous pouvons supposer que ce candidat à la collocation peut potentiellement être une EN.

Nous allons maintenant décrire de manière précise chacune de ces étapes en nous appuyant sur les exemples « fichier clients » et « logiciel ciel ». Rappelons que le but de nos travaux est de déterminer automatiquement que le second candidat est en fait une EN.

Description du processus

Etape 1 – Construction

Nous allons dans une première étape construire des candidats prépositionnels en nous appuyant, dans ces travaux, sur la préposition « de » qui demeure la plus courante. En appliquant ce principe avec nos deux exemples, nous obtenons les résultats suivants :

fichier clients $_{\mathbf{NN}}$ → fichier de clients $_{\mathbf{N\text{-}Prep\text{-}N}}$

logiciel ciel $_{\mathbf{NN}}$ → logiciel de ciel $_{\mathbf{N\text{-}Prep\text{-}N}}$

Notons que lorsque le second Nom du terme de base commence par une voyelle, la préposition qui sera appliquée sera « d' » :

mission intérim $_{\mathbf{NN}}$ → mission d'intérim $_{\mathbf{N\text{-}Prep\text{-}N}}$

Etape 2 – Mesure

Le but de la deuxième étape est de mesurer la dépendance entre chaque mot composant les collocations prépositionnelles construites. Pour cela nous allons nous appuyer sur une des mesures couramment utilisée en Fouille de Textes qui est le coefficient de Dice [Smadja *et al.,* 1996]. Le choix de cette mesure est motivé par son bon comportement que nous avons montré dans nos précédents travaux [Roche et Kodratoff, 2009]. Une telle mesure est définie par la formule suivante :

$$Dice(X, Y) = \frac{2 \times nb(X, Y)}{nb(X) + nb(Y)}$$

[Petrovic *et al.,* 2006] présentent une extension de la formule d'origine de Dice à trois éléments :

$$Dice(X, Y, Z) = \frac{3 \times nb(X, Y, Z)}{nb(X) + nb(Y) + nb(Z)}$$

Le coeur de cette mesure consiste à calculer le nombre d'occurrences de chaque mot « a » (*nb(a)*) ou collocation « a b c » (*nb(a,b,c)*). En règle générale, le nombre d'occurrences, c'est-à-dire la fréquence d'apparition des mots/collocations est calculée relativement à un corpus [Daille, 1996]. Dans notre cas, la mesure de Dice va être appliquée dans un contexte de fouille du web (web mining). Ainsi, la fréquence d'apparition *nb* correspondra au nombre de pages web contenant les mots ou les collocations. Ce nombre est retourné par des requêtes issues des moteurs de recherche (Google, Yahoo, Exalead, etc). Par exemple, *nb(fichier)* correspond au nombre de pages retourné avec le seul mot clé *fichier* et *nb(fichier, de, client)* correspond au nombre de pages retourné avec la requête *"fichier de clients"* (utilisation des guillemets pour rechercher une expression exacte). Les valeurs obtenues avec la mesure de Dice appliquée avec les deux requêtes « *fichier de clients* » et « *logiciel de ciel* » sont données ci-dessous.

$$Dice(fichier, de, clients) = \frac{3 \times 999.000}{37.200.000 + 6.350.000.000 + 208.000.000} = 0{,}000454$$

$$Dice(logiciel, de, ciel) = \frac{3 \times 89.800}{35.000.000 + 6.350.000.000 + 35.400.000} = 0{,}0000419$$

Ce résultat montre que le score le plus faible dans des proportions importantes (facteur dix) est donné par « *logiciel de ciel* ». Ainsi, notre mesure peut prédire que le candidat « *logiciel ciel* » de type Nom-Nom a statistiquement plus de chance d'être une EN comparativement à « *fichier clients* ». Ceci est tout à fait pertinent car cette EN fait référence à un logiciel de gestion et de comptabilité, ce qui correspond au type d'EN appelé *Document* [Daille *et al.,* 2000, Paik *et al.*, 1994].

Les mesures Web donnent une indication de popularité des mots/collocations tout à fait intéressante lorsque des données issues d'un domaine plus ou moins général sont traitées. Par ailleurs, l'avantage de ces connaissances « externes » au corpus (c.-à-d.Web) tient au fait que nous sommes moins sensibles à la taille des données traitées (c.-à-d. corpus). En effet, cette taille et donc la fréquence d'apparition des mots/collocations doit être assez significative lorsque des méthodes statistiques sont appliquées. Avec nos approches de type « Fouille du Web », nous n'avons pas de telles contraintes liées à la fréquence d'apparition des éléments dans les corpus eux-mêmes.

Etape 3 – Sélection

Les candidats à la collocation de type Nom-Nom qui obtiennent de faibles scores représentent des éléments peu enclins à la variation. Dans notre approche, de tels candidats seront considérés comme des EN. La section suivante évaluera la proportion de candidats sélectionnés qui représentent réellement des EN. Dans notre approche, nous allons introduire un paramètre *S* qui représente un seuil de sélection. Par exemple, avec un seuil *S=10*, les dix candidats ayant les scores les plus faibles seront sélectionnés comme EN potentielles. Les résultats selon différentes valeurs de *S* seront discutés dans la section « Expérimentations » de cet article.

Quid de notre approche en anglais ?

La terminologie nominale de type Nom-Nom possède des formes variantes différentes en anglais. Ainsi, les variantes fréquentes d'un candidat à la collocation de type Nom-Nom (par exemple, « knowledge discovery ») sont constituées d'une préposition associée à une permutation entre les noms (par exemple, « discovery of knowledge »). L'ensemble de ces règles pour caractériser les collocations variantes sont détaillées dans [Jacquemin, 1997].

Après avoir décrit, notre approche d'identification des EN à partir de candidats à la collocation, la section suivante présente les résultats expérimentaux obtenus sur des données réelles.

4. Expérimentations

Les corpus

Le premier corpus traité est composé de 1144 Curriculum Vitae (noté **CV**) fournis par la société *VediorBis* (120.000 mots). Une des particularités de ce corpus tient au fait qu'il est composé de phrases très courtes avec de nombreuses énumérations. Les travaux à partir de ce corpus consistaient à déterminer les concepts les plus significatifs pour le domaine [Roche et Kodratoff, 2006]. D'autres travaux sur ce même corpus avaient pour but de classifier les CVs, c'est-à-dire classer les Curriculum Vitae en deux catégories : CVs de cadres et de non cadres [Clech et Zighed, 2003].

Le second corpus de spécialité étudié (noté **RH**) est composé d'un ensemble de textes également écrits en français qui sont issus du domaine des Ressources Humaines (société *PerformanSe : http://www.performanse.fr/*). Les textes correspondent à des commentaires de tests de psychologie de 378 individus (600.000 mots). Les textes sont écrits par un seul auteur qui emploie un vocabulaire spécifique.

Extraction des candidats à la collocation

Le principe d'élagage des candidats à la collocation consiste à considérer seulement les candidats présents un nombre de fois minimum dans le corpus. L'élagage permet, dans la majeure partie des cas, d'exclure les candidats trop rares qui sont souvent peu représentatifs du domaine [Roche et Kodratoff, 2006]. Ainsi, classiquement un élagage à 3 est effectué [Jacquemin, 1997; Thanopoulos *et al.*, 2002]. Le tableau ci-dessous présente le nombre de candidats à la collocation obtenu avant et après élagage à 3.

	Avant élagage		*Après élagage*	
Corpus	**RH**	**CV**	**RH**	**CV**
Nom-Nom	98	1781	11	162
Nom-Préposition-Nom	4703	3634	1268	307
Adjectif-Nom	1260	1291	478	103
Nom-Adjectif	5768	3455	1628	448

Nombre de candidats à la collocation obtenus avant et après élagage.

Suivant les domaines de spécialité écrits dans une même langue, les résultats peuvent différer de manière importante. Par exemple, sur le corpus de CVs après élagage, le nombre de candidats à la collocation de type Nom-Nom (162) est beaucoup plus important que celui du corpus des Ressources Humaines (11) également écrit en français. Le corpus des Ressources Humaines a pourtant une taille cinq fois plus importante que le corpus de CVs. Ceci est dû au fait que les CVs sont écrits de manière condensée en employant un vocabulaire très spécifique : « *emploi solidarité* », « *action communication* », « *fichier client* », « *service achat* », etc. De tels candidats pourraient être assimilés à des collocations de type Nom-Préposition-Nom : « *emploi de solidarité* », « *action de communication* », « *fichier des clients* », « *service des achats* », etc. Dans la section suivante, nous allons nous appuyer sur les candidats à la collocation Nom-Nom du corpus de CVs. Nous allons filtrer les EN à partir de ces candidats en utilisant la méthode statistique présentée dans cet article.

Filtrage des Entités Nommées

Le but de cette section est d'estimer si les candidats à la collocation sélectionnés par notre approche représentent des EN réellement pertinentes. Dans ce cadre, nous nous sommes appuyés sur 70 candidats à la collocation de type Nom-Nom les plus fréquents qui sont estimés pertinents (en tant que terme ou EN). Ces candidats ont été évalués manuellement (18 EN ont été identifiées sur les 70 candidats).

Ces candidats ont alors été classés par la mesure de Dice décrite dans la section « Filtrage des Entités Nommées ». Nous avons donc évalué la qualité des candidats sélectionnés en utilisant différentes valeurs du seuil *S* (sélection des *S* candidats ayant les valeurs de Dice les plus faibles). L'objectif est alors d'évaluer si les candidats sélectionnés par notre système correspondent à des EN pertinentes.

Notons que les mesures appliquées (mesures de Dice) avec les candidats ont nécessité l'exécution automatique de 210 requêtes avec le moteur de recherche Exalead (http://www.exalead.com/) qui utilise des ressources en français assez riches. Nous avons alors effectué 70 requêtes pour les numérateurs et 140 requêtes pour les deux noms propres aux dénominateurs (les requêtes pour les prépositions ont été appliquées une seule fois pour l'ensemble des calculs).

Mesures d'évaluation du filtrage des Entités Nommées

De multiples critères d'évaluation sont disponibles dans le domaine de la fouille de textes. Nous citerons et appliquerons dans nos travaux les deux critères

d'évaluation couramment usités que sont la *précision* et le *rappel*. La précision mesure le nombre d'EN pertinentes sélectionnées par rapport à l'ensemble des EN candidates retournées par un système. Le rappel mesure quant à lui le nombre d'EN pertinentes sélectionnées par rapport au nombre total d'EN pertinentes. Une précision de 100% signifie que toutes les EN extraites par le système sont correctes et un rappel de 100% signifie que toutes les EN correctes sont extraites.

Pour résumer, les mesures de précision et de rappel sont calculées de la manière suivante :

$$Précision = \frac{nb\,EN\ candidates\ pertinentes}{nb\,EN\ candidates}$$

$$Rappel = \frac{nb\,EN\ candidates\ pertinentes}{nb\,EN\ pertinentes}$$

Par ailleurs, une mesure très utilisée qui combine la précision et le rappel est la pondération nommée F-mesure :

$$F-mesure = \frac{2 \times précision \times rappel}{précision + rappel}$$

Résultats du filtrage des Entités Nommées

Nous allons mesurer les résultats selon différents seuils (*S*) sur la base de ces trois critères. Les résultats sont présentés dans le tableau ci-dessous.

Seuil de Sélection (S)	*Précision*	*Rappel*	*F-mesure*
10	**0.60**	0.33	0.43
20	0.45	0.50	0.47
30	0.37	0.61	0.46
40	0.35	0.78	0.48
50	0.36	**1**	**0.53**
60	0.30	1	0.46
70	0.26	1	0.41

Précision, Rappel et F-mesure selon différentes valeurs de S – Corpus de CVs.

Les résultats montrent que la meilleure valeur de F-mesure est obtenue lorsque nous considérons les 50 premiers candidats (*S=50*). Ceci s'explique par le rappel maximum (de valeur 1) car toutes les EN se situent parmi les 50 premiers candidats.

Les résultats du tableau montrent également que les premiers candidats sélectionnés sont assez souvent des EN avec notamment une précision de 60% pour les dix premiers candidats retournés (*S=10*). Ces derniers sont : *lotus note, ciel paie, agent recenseur, chauffeur livreur, go sport, rayon fruit, accueil client, france télécom, paris nord, front page*. Six candidats sont effectivement des EN (société, lieu, logiciel). Remarquons que deux candidats non pertinents ont été sélectionnés par notre approche (*agent recenseur, chauffeur livreur*) car, dans certains cas, notre méthode fondée sur la construction de termes variants de type prépositionnel n'est pas adaptée. En effet, les collocations construites (*agent de recenseur, chauffeur de livreur)* sont erronées. La mesure de Dice a donc retourné un score très faible pour ces collocations qui présentent peu de dépendance entre les trois mots les formant. Notre approche a alors naturellement déterminé ces collocations comme des EN potentiellement intéressantes. Dans ce cas, il aurait été plus pertinent de nous appuyer sur des règles de variation utilisant des conjonctions de coordination (*agent et recenseur, chauffeur et livreur*). Dans de prochains travaux, nous ajouterons de telles règles afin d'améliorer les résultats de précision.

Notons enfin qu'un classement aléatoire retourne une précision de 25% avec *S=10*. Ceci confirme donc que notre méthode, avec laquelle nous obtenons une précision de 60% dans les mêmes conditions, est tout à fait pertinente.

5. Conclusion et Perspectives

Cet article présente une méthode de fouille de textes permettant (1) d'extraire des candidats à la collocation, (2) de déterminer des Entités Nommées (EN) à partir de cette liste de candidats. La méthode de filtrage des EN s'appuie uniquement sur une approche statistique. Celle-ci utilise la mesure de Dice en exploitant les résultats de requêtes issues d'un moteur de recherche. Les EN étant *a priori* peu « stables », nous construisons des candidats variants et vérifions leur popularité via les moteurs de recherche. Si les candidats variants construits sont peu pertinents (c.-à-d. valeur faible de la mesure statistique), ils sont potentiellement considérés comme des EN.

Précisons que ces méthodes ne prétendent pas filtrer de manière exhaustive les EN mais permettent d'obtenir des résultats intéressants à présenter aux experts pour une phase de validation. Dans ce cas, nous devons, en général, privilégier une valeur élevée de précision. Dans le cas du traitement automatique de quantité importante de données (par exemple, pour des tâches d'Extraction d'Information ou de Recherche d'Information), il est souvent nécessaire de privilégier une valeur élevée de rappel même si les méthodes retournent du bruit.

Dans nos futurs travaux, nous envisageons d'enrichir les règles de recherche de variantes car cet article s'appuie sur des méthodes de base simples. Ce point reste donc crucial à améliorer afin de couvrir la grande majorité des variations linguistiquement pertinentes qui seront validées par les approches statistiques décrites dans ce document.

Enfin, nous combinerons ces approches uniquement statistiques pour prendre en compte certaines spécificités lexicales des mots, en particulier la présence de majuscules lorsque cela se révèle possible.

6. Bibliographie

Azé J., Heitz T., Mela A., Mezaour A.D., Peinl P., Roche M. *Présentation de DEFT'06 (DEfi Fouille de Textes)*. Dans les actes de l'atelier DEFT'06, SDN'06 (Semaine du Document Numérique), 2006

Bourigault D., Jacquemin C. *Term Extraction + Term Clustering: An Integrated Platform for Computer-Aided Terminology*. Proceedings of the European Chapter of the Association for Computational Linguistics (EACL'99), p.15-22, 1999.

Brill E. *Some Advances in Transformation-Based Part of Speech Tagging*. Proceedings of AAAI, Vol. 1, p. 722-727, 1994.

Clas A. *Collocations et langues de spécialité*. Meta, Vol 39, No 4, p. 576-580, 1994

Clech J, Zighed D.A. *Data Mining et Analyse des CV : Une Expérience et des Perspectives*. Actes de la conférences EGC, p.189-200, 2003

Daille B. *Study and Implementation of Combined Techniques for Automatic Extraction of Terminology*. The Balancing Act: Combining Symbolic and Statistical Approaches to Language, MIT Press. p.49-66, 1996

Daille B., Fourour N., Morin E. *Catégorisation des noms propres : une étude en corpus*. Cahiers de Grammaire, Volume 25, p.115-129, 2000.

Fort K., Ehrmann M., Nazarenko A. *Vers une méthodologie d'annotation des entités nommées en corpus*. Actes de TALN 2009 (Traitement Automatique des Langues Naturelles), 2009

Heid U. *Towards a corpus-based dictionary of German noun-verb collocations*. Proceedings of the Euralex International Congress, p. 301-312, 1998

Jacquemin C. *Variation terminologique : Reconnaissance et acquisition automatiques de termes et de leurs variantes en corpus*. Mémoire d'Habilitation à Diriger des Recherches en informatique fondamentale, Université de Nantes, 1997.

Laurens M. *La description des collocations et leur traitement dans les dictionnaires*, In Romaneske, Vol 4, 1999

Melcuk I.A., Arbatchewsky-Jumarie N., Elnitsky L., Lessard A. *Dictionnaire explicatif et combinatoire du français contemporain*. Vol 1, 2, 3, 4, Presses de l'Université de Montréal, 1984, 1988, 1992, 1999

Paik W., Liddy E.D., Yu E., McKenna, M. Categorizing and Standardizing Proper Nouns for Efficient Information Retrieval. In B. Boguraev, & J. Pustejovsly (eds), Corpus Processing for Lexical Acquisition, MIT Press, chap. 4., 1994

Petrovic S., Snajder J., Dalbelo-Basic B., Kolar M. *Comparison of collocation extraction measures for document indexing*. Proceedings of Information Technology Interfaces (ITI), p.451-456, 2006.

Roche M., Kodratoff Y. *Pruning Terminology Extracted from a Specialized Corpus for CV Ontology Acquisition*. Proceedings of onToContent'06 workshop (Ontology content and evaluation in Enterprise) - OTM'06, Springer-Verlag, LNCS, p.1107-1116, 2006

Roche M, Kodratoff Y. *Text and Web Mining Approaches in Order to Build Specialized Ontologies*. Journal of Digital Information (JoDI), Vol 10, No 4, 2009

Smadja F., McKeown K. R., Hatzivassiloglou V. *Translating collocations for bilingual lexicons : A statistical approach*. Computational Linguistics, vol. 22, No 1, p. 1-38, 1996

Thanopoulos A., Fakotakis N., Kokkianakis G. *Comparative Evaluation of Collocation Extraction Metrics*. Proceedings of 3rd International Conference on Language Resources and Evaluation (LREC'02), p.620-625, 2002.

A propos des auteurs

Équipe TAL - LIRMM
UMR 5506, CNRS, Univ. Montpellier 2
34392 Montpellier Cedex 5 - France
mroche@lirmm.fr

http://www.lirmm.fr/~mroche

Ontologies des risques financiers

Continuité d'activité, gestion de crise, protection des *infrastructures critiques* financières

J. - Yves Gresser

Résumé : Depuis la parution en 1995 du premier livre blanc sur « la sécurité des systèmes d'information dans la banque », le sujet des risques individuels ou systémiques dans la finance est toujours d'actualité.

Ces risques sont multiples. Dans cette communication, nous nous attachons à la représentation des risques liés aux infrastructures, à partir de deux exemples : - La structuration du lexique de la « continuité d'activité » de la Banque de France ; - les ontologies de risques financiers développées dans le cadre du projet européen PARSIFAL[1].

Il n'existe pas à ce jour d'ontologie générale de la sécurité mais des développements spécifiques dans certains domaines. Il existe encore moins d'ontologies relatives au domaine bancaire ou financier. La spécialisation des connaissances entre finance et sécurité est un obstacle supplémentaire à la bonne appréciation des situations comme au traitement approprié des menaces, lorsqu'elles s'annoncent ou se matérialisent. Nous avons été les premiers à procéder au rapprochement formel des connaissances dans les deux domaines.

Nous avons développé les modèles dans plusieurs directions qui trouvent leur écho dans les crises ou problèmes actuels : vols d'identités, maîtrise des marchés financiers et tout simplement continuité d'activité dans un monde de plus en plus dépendant des réseaux et de l'informatique. Un premier résultat tangible fut l'émergence de la notion de cycle de vie d'une crise et sa cartographie.

Nous avons commencé à partager cette approche innovante et ses résultats avec les différentes parties prenantes : financiers, spécialistes de la sécurité et stratèges politiques européens.

[1] www.parsifal-project.eu, en.wikipedia.org/wiki/PARSIFAL_Project_EU

Mots-clés : risque, crise, finance, paiement, règlement, infrastructure critique, continuité d'activité, lexique, sémantique, ontologie, modèle.

Risk, crisis, payment, settlement, critical infrastructure, business continuity, control engineering, glossary, semantic, ontology, model.

1. Introduction : ontologies, vocabulaires, lexiques structurés

Il n'est pas aujourd'hui de sujet en émergence qui n'exige une entente préalable sur un vocabulaire commun. C'est d'autant plus vrai dans un contexte international.

Ces vocabulaires sont présentés couramment sous forme de « lexiques » (dans le sens de répertoires), plus ou moins volumineux, où les termes sont rangés dans l'ordre alphabétique (ou tout autre ordre fondé sur la graphie). Quelle que soit la langue utilisée, l'inconvénient de ce type de présentation est de séparer des termes dont les sens sont liés. De tels répertoires sont utilisables comme référentiels, faciles à utiliser pour des mots déjà lus ou entendus. Ils ne peuvent acquérir une valeur pédagogique ou descriptive qu'après structuration. Une telle structuration peut reposer sur un fractionnement en domaines ou sous-domaines. Ce fractionnement est souvent fait de manière intuitive.

La Banque de France (BDF) a produit, en 2008, un « *Lexique structuré de la* Continuité d'activité » ordonné en 6 chapitres : Principes directeurs, Stratégie, Planification, Plan de reprise, Gestion de crise, Maintien en condition opérationnelle et optimisation. Les entrées y étaient listées et définies dans chacun des chapitres, sans que leurs relations y soient toujours explicitées.

La modélisation sémantique a permis d'en faire le tour et d'améliorer sensiblement la lisibilité du *Lexique*. Elle a surtout permis de préciser ou de faire apparaître certains concepts.

2. Premiers pas : continuité d'activité, gestion de crise.

Dans ce qui suit nous en traiterons rapidement deux exemples : la *stratégie* et la *gestion de crise*.

2.1. Vous avez dit stratégie ?

Dans le document original, le contenu d'une stratégie de continuité d'activité est définit/décrit comme suit :

> *« Bonne connaissance de l'entreprise et de ses* ***activités critiques,*** *du contexte réglementaire; du rôle économique de l'entreprise (risque que l'on fait courir) ; détermination des rôles critiques (key tasks) des données vitales (vital records) ; des vulnérabilités, exposition, zone à haut risque, point critique.. (exposure, high-risk areas, single point of failure…), menaces (vocabulaire de l'analyse des menaces*) (natural and human threats), dépendances (dependency) et scénarios plausibles (scenario) ; choix des activités à privilégier (priorization); orientations sur les moyens à retenir ; mise en place de l'organisation. »*

Ce type de définition/description est pour le moins dense. Il n'est pas très explicite sur les fondamentaux de la maîtrise des risques, ni sur les buts recherchés ou les acteurs concernés.

Un premier modèle sémantique a fait émerger ces éléments, et leurs relations fondamentales. Nous l'avons complété des éléments provenant de problématiques voisines issues d'autres secteurs. Par la suite, nous l'avons fait évoluer en le présentant à certains auteurs du *Lexique* (figure 1).

Ce modèle nous a permis :

- d'expliciter des idées essentielles, et de dégager des caractéristiques communes,
- de nommer et de commencer à caractériser les relations principales.

Ces caractérisations sont indispensables pour fonder l'intercompréhension entre les acteurs.

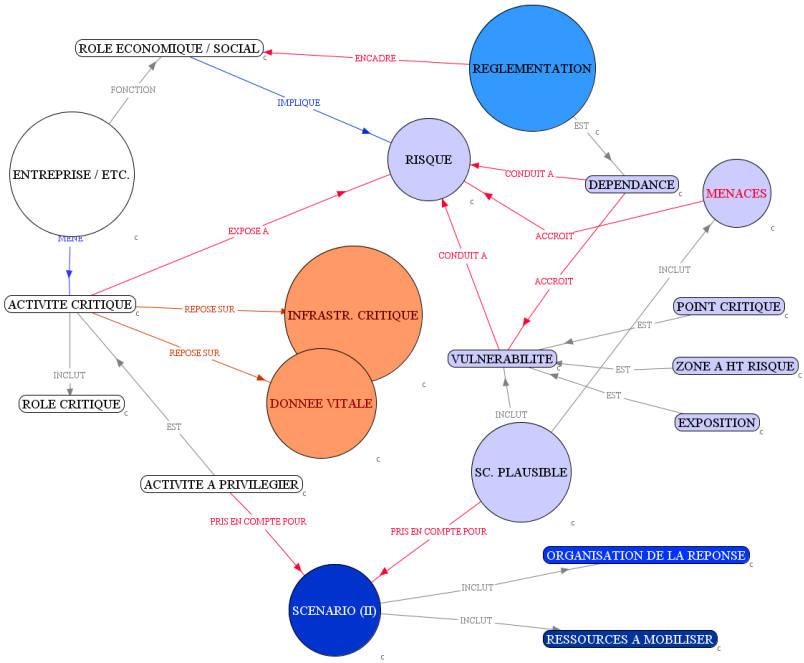


Figure 1 – Concepts[2] d'une stratégie de continuité d'activité

2.2. Vade-mecum de crise

Imaginez un haut-responsable qui entend un signal d'alarme. Sa « mallette de crise » est, bien sûr, sur sa table. Pour savoir quoi-faire, il n'a qu'à l'ouvrir, en sortir le lexique de 30 ou 60 pages qui est au-dessus de la pile de documents et d'objets qu'elle contient, ouvrir ce lexique... Où est la table des matières ?.. A... Ab... Al... Il est temps de refermer la mallette et de se sauver... Vers quel point de rendez-vous ? Par où ?

Ce serait tellement simple d'avoir un simple carton sur lequel serait clairement représenté quoi faire et où dans les prochaines heures, journées, semaines. Non pas la crise en cent mots mais en une page.

Dans le *Lexique*, la gestion de crise faisait l'objet d'une section particulière. Dans cette section, l'emploi du terme « jalon » revenait à plusieurs reprises à côté d'expressions comme « phasage » et « actions ».

[2] Il s'agit de « concepts » en ontologies.

Quelles étaient les relations entre ces désignations ? Leur emploi était-il adapté au contexte ? Comment les relier à des termes plus spécifiques comme « pré-crise », « survenance du sinistre », « paroxysme » ?

C'est ainsi qu'a émergé la notion de <cycle de (vie d'une) crise> à laquelle les « jalons » déjà désignés pouvaient se rattacher. Il nous a suffit de les compléter pour avoir une vision synthétique de ce cycle en six étapes.

Les autres notions étaient soient communes à toutes les étapes, soient rattachables à une étape spécifique. Leur articulation est devenue plus claire.

La figure 2 est une version simplifiée du diagramme final. L'actualité nous a permis de valider le diagramme complet sur des crises réelles (affaire de la Société générale, crise des crédits de 4e niveau « subprime »). Il fait maintenant partie, nous dit-on, de la mallette de crise des dirigeants de la BDF.

Dans sa version simplifiée ou complète, ce diagramme nous semble de portée universelle, applicable aux catastrophes naturelles comme à d'autres crises comme la marée noire de Louisiane ou le blocus de la bande de Gaza.

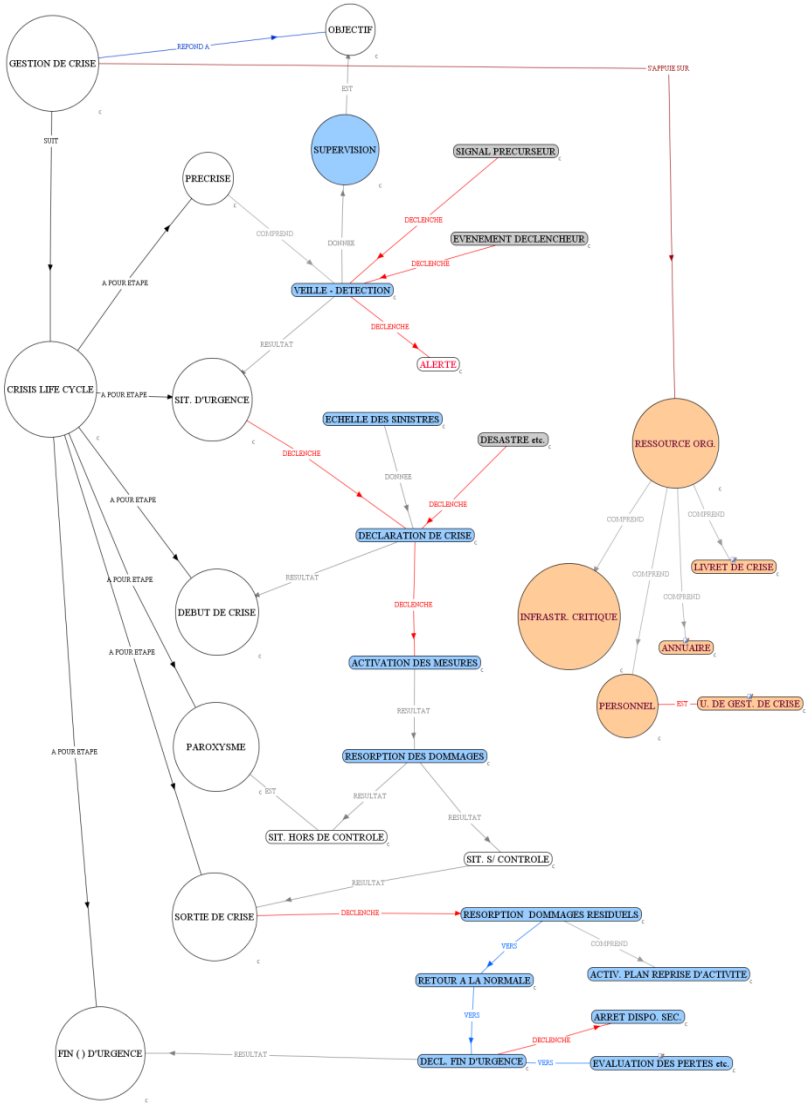


Fig. 2 – La crise en une page (diagramme simplifié)

3. Systématisation de la démarche

Les travaux menés sur les termes de la continuité d'activité ont constitué le point de départ d'un travail plus approfondi sur la modélisation des risques financiers et la protection des « infrastructures critiques ». Celui-ci a fait l'objet

d'un rapport définitif remis à la Commission européenne en janvier 2010 dans le cadre du projet PARSIFAL (Gresser 2010).

Dans cet article, nous en retraçons les grandes lignes et développons deux aspects :

- la formulation d'une ontologie commune à la sécurité et au domaine financier,
- la modélisation de l'ingénierie du contrôle, l'un des trois sous-domaines que nous avons approfondis.

Les services financiers font appel à une multiplicité grandissante d'intervenants dont les interconnexions sont de plus en plus complexes.

Un des objectifs principaux du projet était d'identifier les meilleures pratiques relatives à la protection des infrastructures critiques financières, la manière de les diffuser auprès des responsables futurs ou actuels de ces infrastructures, opérateurs ou prestataires de services.

Ceci demandait davantage de clarté et de visibilité des méthodes utilisées, en bref : le recours à des moyens permettant aux parties prenantes de :

- avoir facilement accès aux savoirs existants ;
- capitaliser sur ces savoirs en mémorisant conjointement leurs propres connaissances[3], ceci d'une manière concise et parlante.

3.1. Vers une ontologie commune des risques et du domaine financier

Les ontologies formelles devraient constituer la colonne vertébrale de la Toile sémantique. Dans notre domaine, le W3C a créé en 2007, un groupe d'incubation (XG) pour une ontologie de maîtrise des désastres (Ianella *et al.* 2009).

Après un désastre ou un cataclysme un échange rapide d'informations entre les multiples intervenants est crucial. Seul un référentiel ontologique commun peut assurer l'interopérabilité entre équipes et systèmes dont les rôles sont liés,

[3] Voir Rute Costa, Raquel Silva, De la typologie à l'ontologie des textes, in Actes de la conférence TOTh 2008, p.5

complémentaires et tous indispensables pour minimiser les dégâts et sauver des vies humaines.

Un tel référentiel n'existait pas en 2007. L'EIIF- XG a clos ses travaux en 2009, en publiant une ébauche fondée sur un nombre restreint de concepts capables d'assurer l'interopérabilité des ontologies spécialisées.

Une enquête menée en 2008 dans le cadre de l'Université de la Manche (Blanco *et al.* 2009) sur les ontologies du domaine sécuritaire confirmait l'inexistence d'un modèle général capable d'évolution.

Il existe encore moins d'ontologies dans le domaine financier (Vanderlinden *et al.* 2009, IFIP 2003) et, jusqu'à nos travaux, aucun modèle général combinant sécurité et finance.

Plutôt que de chercher à capitaliser sur des modèles, somme toute, assez sommaires, nous avons préféré continuer à exploiter la littérature de fond (vocabulaires, dictionnaires, « green or blue books » etc.) des autorités financières. Dans cette optique, le projet de « Lexique relatif aux systèmes de paiement, de compensation et de règlement » de la Banque centrale européenne (BCE) – dans (Gresser 2010), s'est révélé d'un apport fondamental, non par sa structure mais par la liste des termes qu'ils comportaient. Nous avons complété cette liste de termes issus de différents organismes de normalisation, de tutelle ou de groupes spécialisés (ISO, CEN, Forum tripartite, TWIST, etc.).

On peut classer l'ensemble de ces termes en deux groupes : - ceux spécifiques à la finance (ou au commerce) comme paiement, *livraison-règlement de valeurs mobilières, produits dérivés*; - ceux de portée plus générale comme *système, règle, processus, infrastructure.*

La figure 3 schématise les concepts de haut niveau et leurs principales relations. La représentation graphique traduit leur regroupement en sept ensembles (*règle, rôle, service, infrastructure, système, processus, évènement*).

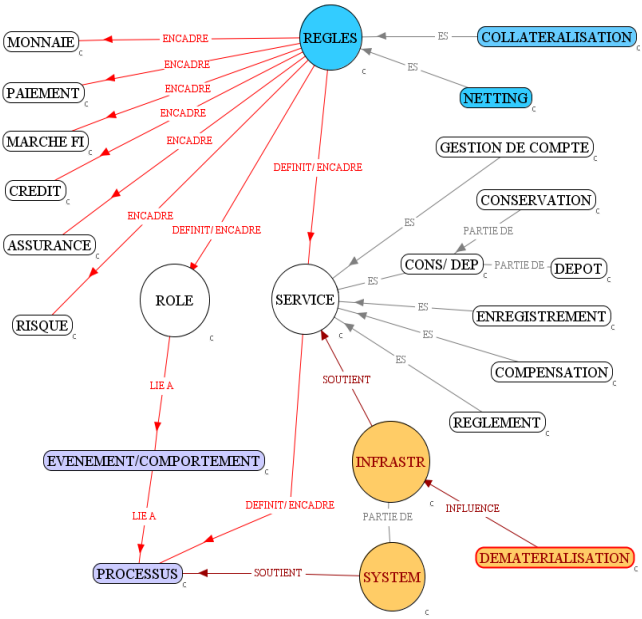


Figure 3 – Concepts fondamentaux du secteur financier

Cette représentation n'est pas exhaustive. Ainsi un *évènement* ou un *comportement* peut déclencher un *processus* tout comme constituer le résultat d'un autre processus. De même le concept d'*établissement* n'apparaît pas. Les acteurs y sont plutôt désignés par leur *rôle*.

La figure 4 couvre les marchés financiers. Elle peut être utilisée de deux façons : en tant que validation du diagramme précédent, ou en tant que représentation d'un domaine plus restreint.

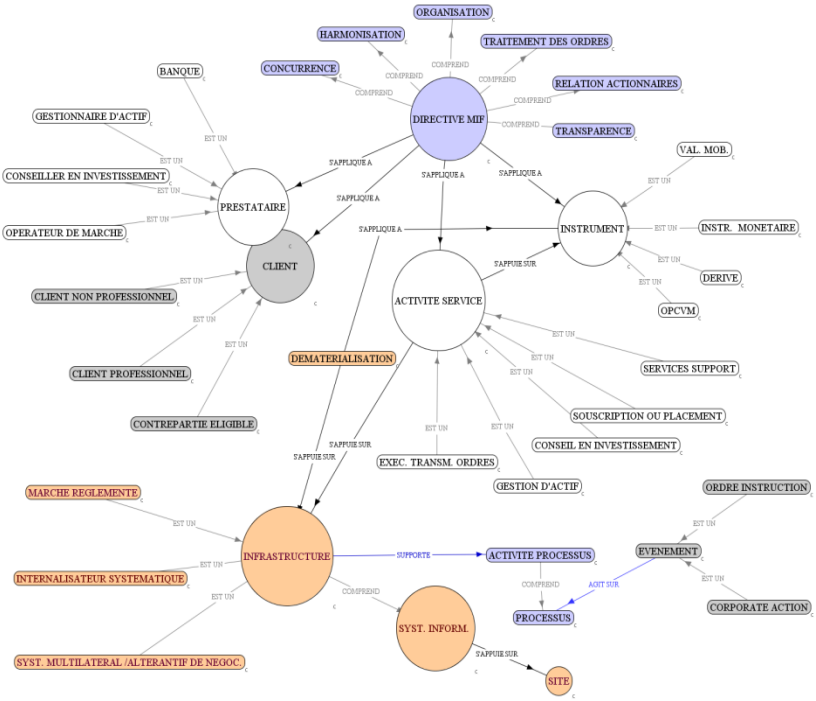


Figure 4 – Marchés financiers, concepts de haut niveau

Venons-en au domaine sécuritaire. La figure 5 est dérivée du schéma des concepts de haut niveau des « Critères communs ». Nous l'avons complété de considérations relatives à la continuité d'activité et à la gestion de crise. Certains termes propres à la protection des infrastructures critiques (CIIP) commencent à y apparaître, de même que des termes que l'EIIF-XG tend à considérer comme des « méta-concepts », comme :

- *acteur (Player ou actor)/partie prenante* ;
- *actif* ou *bien* (*Asset*) qui peut désigner à ce niveau aussi bien une *infrastructure* qu'un *produit* ou un *service* utilisant cette infrastructure, ou tout autre actif ou bien détenu par les parties prenantes.

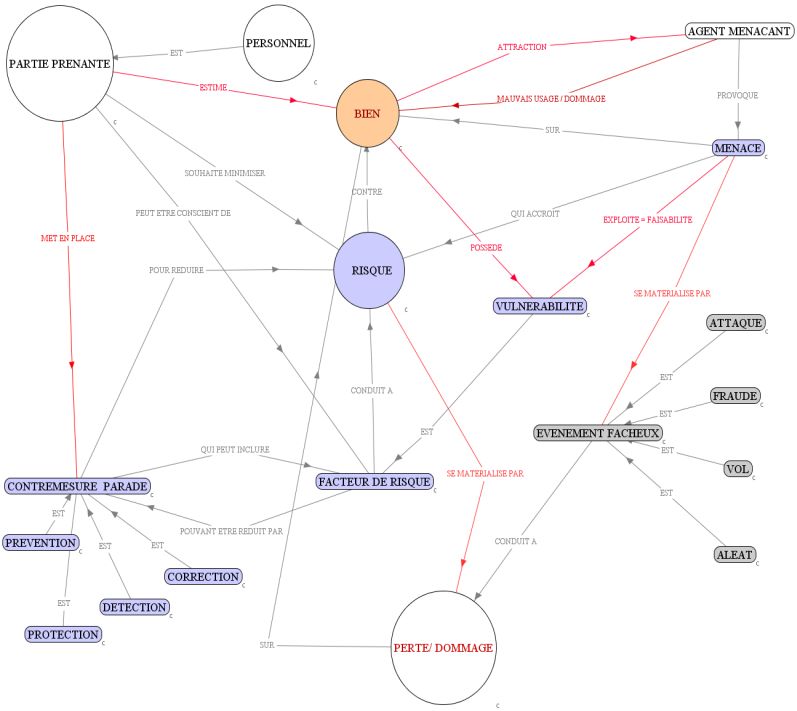


Figure 5 – Gestion du risque, concepts de haut niveaux

Comprendre les différences entre risque/menace/sinistre est essentiel. Celles-ci peuvent s'exprimer notamment à travers l'opposition potentiel/effectif : *risques* et *menaces* sont potentiels; *évènements* et *comportements* (au sens de *behaviour* dans, par exemple, UML) sont effectifs. Notre figure exprime cette opposition que les schémas classiques ignorent.

Pour fusionner les ontologies financières et sécuritaires, notre première idée fut de nous concentrer sur les concepts fortement reliés à la CIIP comme *risque infrastructure*, *système* et certains processus choisis, en utilisant les « manques » réciproques.

Mais cela ne suffisait pas. Il fallait prendre en compte certaines particularités du secteur financier. Ce secteur est très fortement spécialisé et encadré. Cet encadrement s'exprime à des niveaux divers, *stratégie*, *conduite* ou surveillance (*supervision*) des *opérations*, et se traduit par des principes directeurs, des manuels de *procédures* (Rule Books), des textes juridiques de portée locale ou internationale.

La difficulté pour l'analyse est qu'une *règle* ou *procédure* peut s'appliquer simultanément au domaine financier et au domaine technique. Il en est ainsi des règles devant assurer la *liquidité* des marchés financiers à travers les *systèmes de règlement* électronique, fonctionnant en temps réel.

Il ne s'agit pas d'un problème purement notionnel. Le risque est important pour un responsable financier d'accorder une trop grande confiance à certaines infrastructures techniques. Ce que l'on qualifie généralement de *risque de contrôle* ou de *risque de gouvernance*.

En tout état de cause une liste commune de haut niveau doit inclure certaines des *règles* les plus importantes et nommer les *acteurs* principaux.

La figure 6 est dérivée de la figure 5. Les relations entre les concepts ont été tirées des définitions disponibles ou ajoutées et validées par des experts. Le diagramme a été conçu de manière à minimiser le nombre de relations. Tous les concepts n'ont pas été pris en compte. Certains sont regroupés sous un seul concept comme ceux relatifs au *personnel critique*.

Cette figure 6 peut apparaître comme trop générale. La figure 7 représente l'application de notre démarche aux systèmes de règlement.

Comme les précédentes, elle est fondée sur sept ensembles : *règle, acteur, service, infrastructure, système, processus, risque*. Y figurent des éléments spécifiques aux systèmes de règlement. La notion de règle y couvre des éléments complexes et variés qui vont du *prudentiel* à l'*opérationnel*, cela pour couvrir les droits ou obligations des différentes parties prenantes.

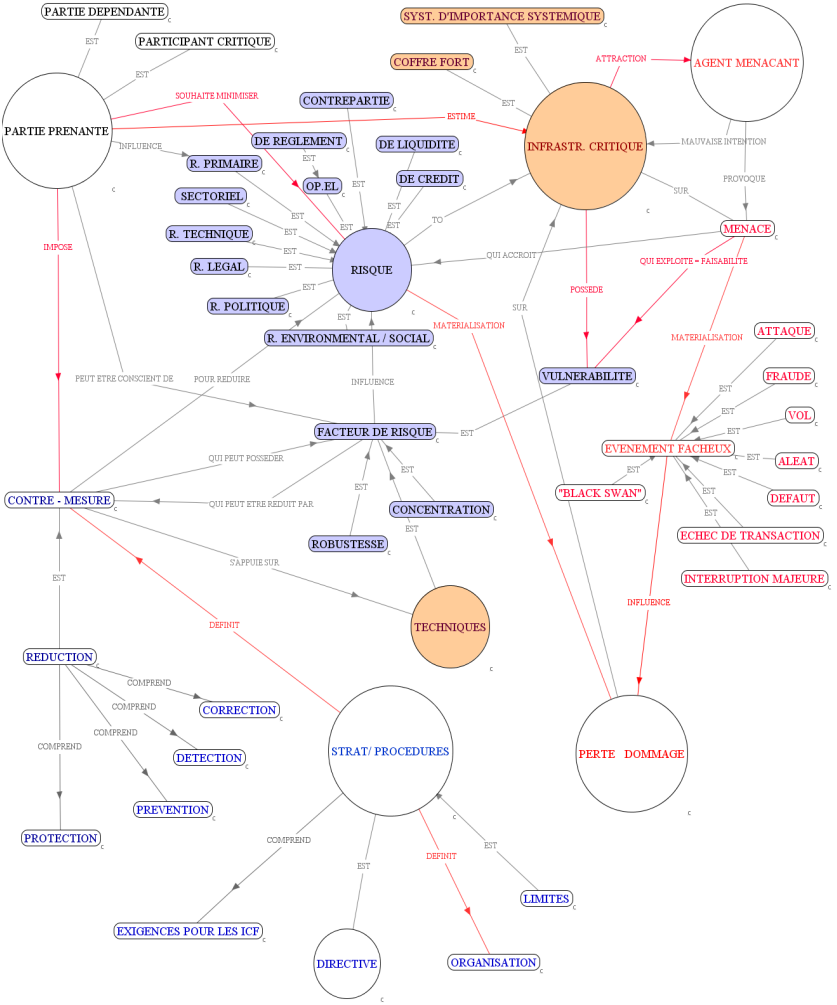


Figure 6 – Risques financiers & techniques, concepts de haut niveau

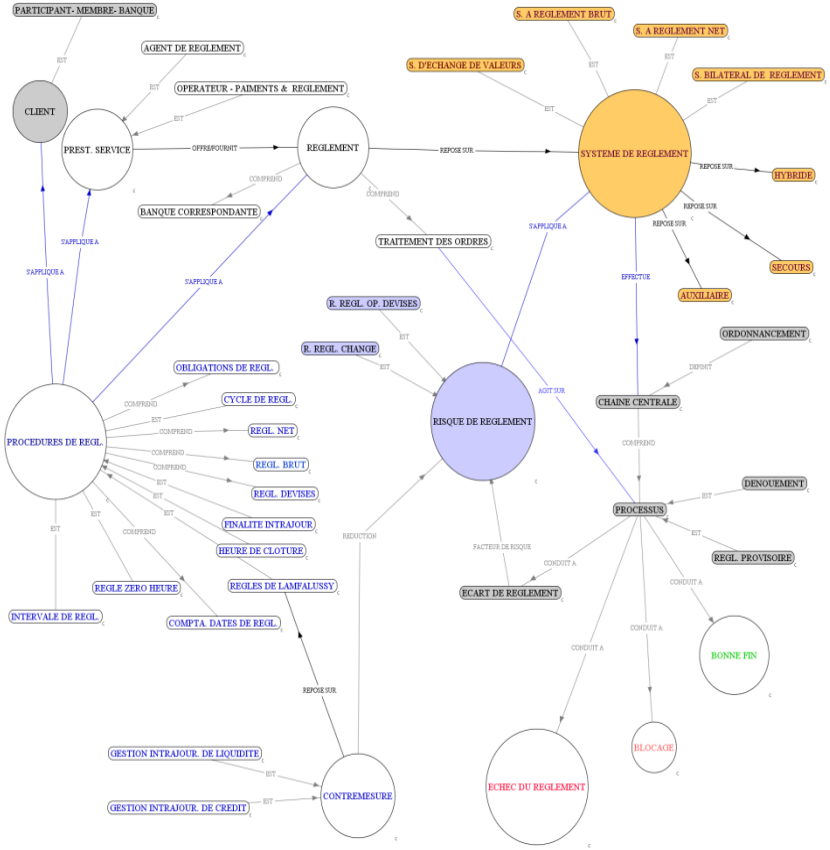


Figure 7 – Risques et règlements, concepts de haut niveau[4]

3.2. L'ingénierie du contrôle5

Le « Rapport au Premier ministre concernant les « Enseignements à tirer des événements récemment intervenus à la Société Générale » de février 2008[6] confirmait l'opinion générale selon laquelle les infrastructures informatiques sont encore largement perméables à la fraude.

---

[4] Dans ce diagramme nous avons privilégié l'aspect opérationnel. Tous les acteurs ne sont pas représentés et certaines situations sont décrites de manière raccourcie.

[5] L'anglais *control* signifie à la fois *commande*, *pilotage* et *contrôle*. C'est bien de cela qu'il s'agit dans notre contexte.

[6] www.ladocumentationfrancaise.fr/rapports-publics/084000062/

En juillet 2008, le Secrétariat général de la Commission bancaire émettait un blâme en direction de la Société générale (SG) après s'être penchée « sur la qualité du système de contrôle des opérations et procédures internes, en particulier la maîtrise des risques opérationnels[7] ».

Avec un petit groupe de banquiers du « 2e cercle », nous nous sommes attachés à décoder les documents que la SG elle-même a rendus publics[8]. Nous en avons retenu les aspects suivants :

- l'occultation des opérations ;
- la discontinuité des contrôles ;
- l'intérêt d'une vision de bout en bout et l'auditabilité des procédures, notamment celles liées à la sécurité. L'absence d'une vision globale de départ peut avoir un impact sur la pérennité des mesures mises en œuvre dans l'urgence, bien sûr, mais surtout sur l'efficacité de ces mesures ;
- les contrôles hiérarchiques, la gestion des délégations et des habilitations fait partie des attendus de la CB. Un excès de confiance en l'informatique pour la gouvernance, l'identification ou l'authentification, et l'audit, entraîne un abandon des prérogatives managériales aux informaticiens.

En bref, le défi est double : pouvoir réagir très vite à un instant donné, à un événement attendu ou inattendu, et savoir s'adapter au fur et à mesure des inventions du marché. Dans ce cas, l'enjeu véritable est celui du renforcement des dispositifs de contrôle des opérations de marché dans les banques ainsi qu'entre les banques et leurs partenaires.

La figure 8 reprend la figure 6 pour ce genre de contexte. C'est une interprétation sémantique du court rapport dont nous avons extrait le texte ci-dessus. Il est important de noter qu'il s'agit d'un modèle de haut niveau et que l'analyse d'un cas particulier nécessite davantage de détails.

---

[7]www.banque-france.fr/fr/supervi/telechar/supervi_banc/200807-decisions-juridictionnelles.pdf

[8] Ces rapports ne sont plus accessibles en ligne mais ils ont pu être téléchargés en leur temps.

La figure 9 est une extension de la figure 8 à partir des notions dégagées lors d'une série d'exercices de créativité organisés avec les différentes parties prenantes : *financiers, prestataires de services et organismes de tutelle* ou *de supervision*.

Elle comporte :

- des précisions sur les *attaques* ou *attaquants potentiels* ;
- les *contre-mesures* envisageables notamment celles relatives à la gestion des *identités numériques*, la nécessité d'expliciter les règles d'habilitation et de délégation ;
- certaines évolutions technologiques comme la *virtualisation* (l'informatique en nuage) ou le *nomadisme*.

Par contre, aucune relation nouvelle n'y apparaît.

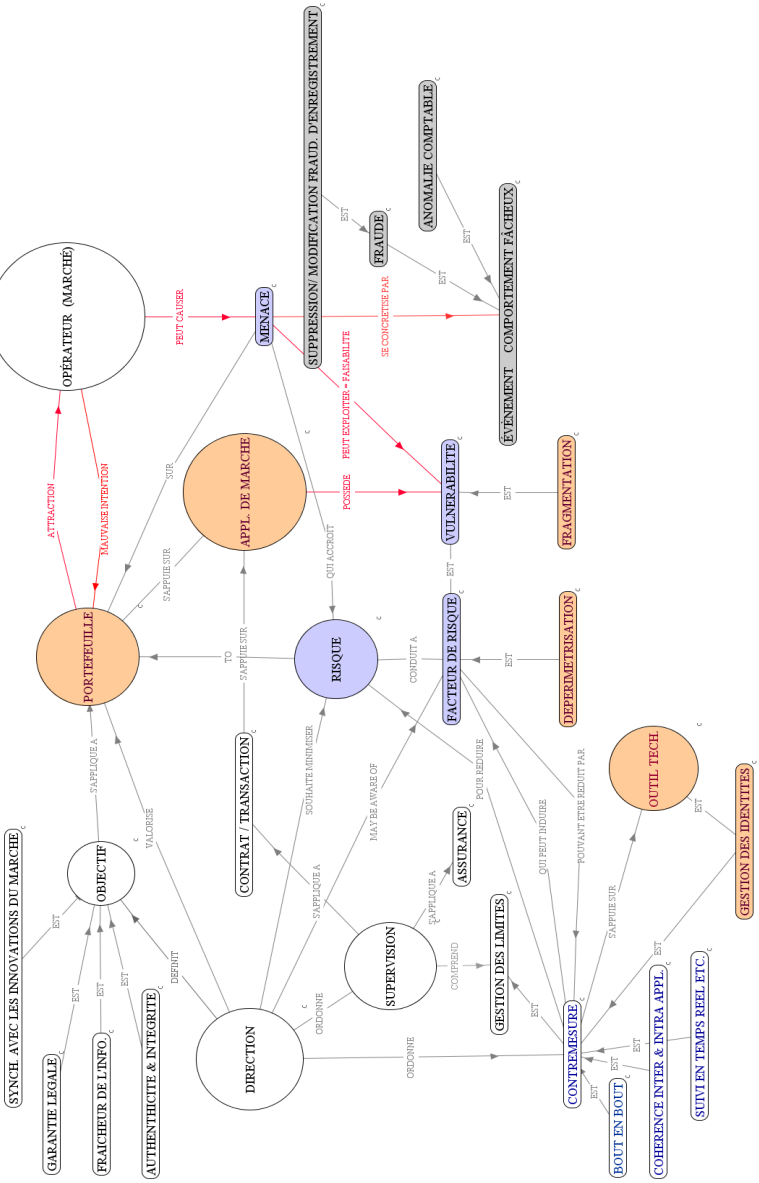


Figure 8 – Concepts de l'ingénierie du contrôle, appliqués aux opérations sur les marchés financiers

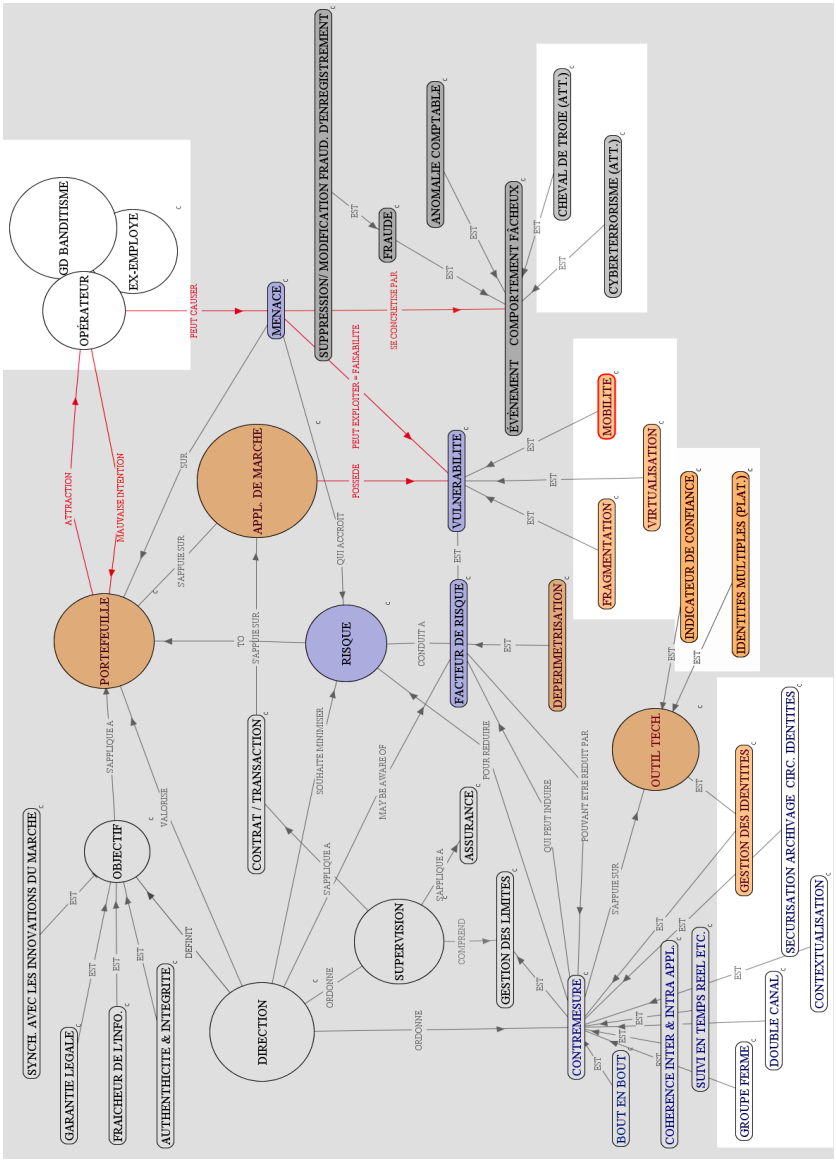


Figure 9 – Concepts de l'ingénierie du contrôle, appliqués aux opérations sur les marchés financiers – Stade final (simplifié)

3.3. Autres développements

Ces ontologies de haut niveau sont les prémisses de développements en surface et en profondeur. Dans le cadre du projet PARSIFAL, nous pensons avoir démontré leur capacité d'évolution (fig. 6 vers fig. 7, fig. 8 vers fig.9).

Pour compléter le travail sur les ontologies, nous les avons accompagnées, de manière exploratoire et non exhaustive, de deux vocabulaires. Le premier, un vocabulaire autour de cinq concepts de haut niveau auxquels nous avons ajouté celui d'identité (numérique) (fig.10). Le deuxième, un vocabulaire qui regroupe les expressions de notre travail et celles utilisées par la profession bancaire.

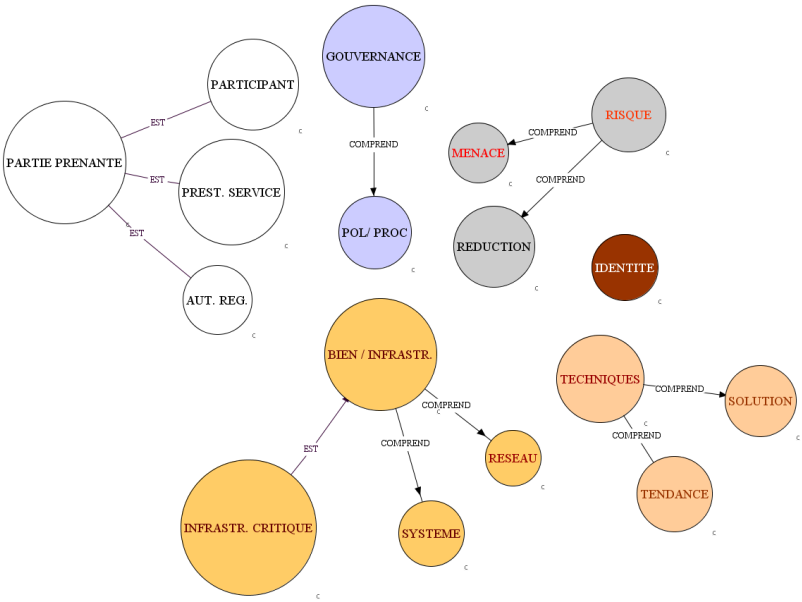


Figure 10- Lexique de PARSIFAL, catégories de haut niveau

4. Conclusion

Le rapport précité du projet PARSIFAL traite de manière identique deux autres grands sujets : le *partage des informations sensibles* et la *continuité d'activité* dans un contexte géographique et institutionnel étendu.

Le premier objectif des modèles était le partage de connaissances entre les différents intervenants internes et externes du projet. Il a été atteint même si une telle approche est encore peu courante. Les modèles ont, bien sûr, besoin d'être

présentés et commentés. À ce jour, ceux qui y ont été exposés, ont réagi très favorablement à : - leur valeur éducative ou tout simplement opérationnelle (fig.2) ; - la modélisation plus ou moins détaillée relative à la gestion des risques ou de situations diverses allant d'un simple incident à une catastrophe.

Ces représentations sont utiles sinon indispensables au développement de méthodes quantitatives d'évaluation des risques et de l'efficacité des mesures proposées. Mais cela va plus loin : dans un contexte où les menaces sont mondiales et où la réponse des États reste fragmentée, la création d'une ontologie ou d'une famille d'ontologies communes est une condition essentielle à la coordination des actions internationales.

Nous avons aussi montré, via la structuration des vocabulaires, qu'une démarche ontologique est un outil précieux pour simplifier la présentation de savoirs ou de savoir-faire complexes, tout en assurant l'exhaustivité et l'évolutivité des modèles sous-jacents.

En bref, ces représentations sont des instruments clés de la maîtrise des risques. Dans le projet PARSIFAL, elles nous ont servi à l'identification d'axes de progrès ou de recherches. Elles constituent aussi, en elles-mêmes, un sujet de recherches.

Remerciements

Ceux-ci vont à mes ex-collègues des professions financières (banque et assurance), en particulier A. Dequier (Dequier 2008, 2004), à mes collègues de l'équipe et aux *parties prenantes* de PARSIFAL, qui ont fourni termes ou listes de base et participé à la validation de certains schémas, à la Commission européene qui a soutenu une partie de ces recherches, et à Dardo De Vecchi pour la relecture attentive de cette communication.

Bibliographie

Adar, E., *European Framework for CIIP Risk Management: challenges, initiatives, policies,* in CR 3e réunion CEPS CIIP Task Force, janvier 2010, p.6-9

Banque de France (BDF), *dictionnaire structuré de la continuité d'activité*, 2008, document interne

Blanco, C., Lasheras, J., Valencia-García, R., Fernández-Medina, F., Toval, A. et Piattini, M., Ontologías de Seguridad: *Revisión sistemática y comparativa, Departamento de Tecnologías y Sistemas de Información*, Universidad de Castilla-La Mancha, juillet 2008, 74 p.

Club de la continuité d'activité (CCA), *Lexique structuré de la continuité d'activité*, Livre blanc n°1, 2009, 59 p.

Dequier A. (BDF), *Un témoignage issu d'une expérience professionnelle à la Banque de France et deux suggestions*, Actes TOTh 2008, p. 259-271

Dequier A. (Secrétariat général de la Commission bancaire SGCB), Pierre Poulain (BDF), *Pour un vocabulaire de crise raisonné et partagé, Gestion des crises, Risque & Prudentiel*, in Banque, septembre 2004

Gresser, J.-Y., *Draft Ontology Of Financial Risks & Dependencies Within & Outside The Financial Sector*, Vol. 1 & 2, D2.1, projet PARSIFAL, janvier 2010, 161 p. disponible sur le site www.parsifal-project.eu[9]

Iannella, R. NICTA, Berg-Cross, G. EMI Semantic Technology, Curzon, R. IBM, de Silva, C. Lanka Software Foundation, Di Maio, P. University of Strathclyde, Cutter Consortium, Sotoodeh, M. University of British Columbia, Olsson, O. SICS, Vetere, G. IBM Italia, *Emergency Information Interoperability Frameworks*, W3C Incubator Group Report, août 2009 disponible sur www.w3.org/2005/Incubator/eiif/XGR-Framework-20090806/

International Financial Information Publishing Ltd (IFIP), *Semantic Web 'Financial Exchange Framework' (FEF)* Ontology, 2003, initialement accessible à www.financial-format.com/fef.htm

Vanderlinden, E., Allemang D., Finance: *demonstration of risk factors under control, Breaking the conspiracy for ignorance*, Press Release, août 2009 disponible sur *www.fadyart.com/pressreleases.html*

De Vecchi, D. (Université Paris 7 – EILA Euromed-Marseille, Ecole de management), *Pragmaterminologie : une terminologie de l'entreprise en évolution*, *http://www.realiter.net/spip.php?article1763*

A propos des auteurs

Jean-Yves Gresser
GELM entreprises, Paris
jgresser@gelm-entreprises.com

[9] Contient une bibliographie détaillée.

Vers une ontologie pour le domaine de l'analyse de sécurité des systèmes de transport automatisés

Lassaâd Mejri, Habib Hadj-mabrouk, Patrice Caulier

Résumé : La résolution de problèmes dynamiques confronte l'opérateur humain à de nombreuses tâches à forte composante de prise d'information et de décision, sous contrainte temporelle. Si elle requiert des savoirs et des savoir-faire fondamentaux, la résolution de problèmes dynamiques nécessite aussi l'exploitation d'une expertise, à la fois statique et dynamique, trop peu souvent formalisée et capitalisée à des fins, notamment, d'assistance ou de supervision.

Une expérience de développement de systèmes d'aide à la résolution de problèmes dynamiques dans plusieurs domaines, notamment dans les domaines de l'analyse sécuritaire (Mejri, 1995), (Mejri *et al.*, 1998) , de la fiabilité du logiciel (Hadj mabrouk *et al*, 2000), du web sémantique (Mejri *et al.*, 2008), de l'information multimodale (Mejri *et al.*, 2008) ou du génie logiciel (Kessentini et al. ,2009) nous a amené à spécifier et modéliser un cadre générique de formalisation de l'expertise spatio-temporelle de résolution de problèmes dynamiques baptisé « scénario de résolution de problème » (Mejri *et al.*, 2005). En effet, le problème de l'analyse de sécurité des systèmes de transport automatisés est un problème typique où les experts du domaine proposent des situations d'insécurité potentielles conjointement par des descriptions statique et dynamique.

Mots-clés : Mise en forme, représentation et capitalisation de connaissances, scenario d'insécurité, réutilisation, apprentissage et classification, analyse de sécurité, systèmes de transport automatisés.

1. Introduction

L'analyse de sécurité des systèmes de transport automatisés (exemple : le Métro sans conducteur : Véhicule Automatique Léger) et des automatismes embarqués est une tâche cruciale qui doit se faire avant la mise en site définitive d'un nouveau système de transport terrestre automatique. Dans cette tâche plusieurs intervenants contribuent ensemble pour s'assurer de la conformité du système (matériel et logiciel) aux exigences de sécurité. Nos travaux antérieurs sur ce problème ont mis à contribution plusieurs domaines de recherche complémentaires. Ces domaines sont liés à :

a) l'acquisition de connaissances,
b) la représentation de la connaissance,
c) et enfin le domaine de la résolution de problème connu sous le nom "problem solving" en littérature. Les méthodes de raisonnement sur la connaissance (l'apprentissage automatique et les systèmes à base de règles) sont très utiles pour dériver de nouvelles connaissances à partir de celles obtenues lors de la phase d'acquisition (Kodratoff et Diday, 1991), (Michalski et Kodratoff, 1993 : 1-27).

Dans ce qui suit, nous allons parcourir l'application de ces domaines sur le problème d'analyse de sécurité pour déboucher ensuite sur la proposition d'une ontologie pour ce domaine d'analyse des systèmes de transport automatisés. Un modèle de scénario de résolution de problème à caractère générique qui devrait utiliser des connaissances ontologiques a été proposé dans (Mejri et al, 2009). Ce modèle initié dans (Mejri et Caulier, 2005) bien que applicable au domaine de l'analyse de sécurité pourrait être appliqué à d'autres domaines d'étude comme le web sémantique (Mejri et al, 2008) ou l'information multimodale personnalisée (Mejri et Ben fraj, 2008).

2. Résultats de l'acquisition des connaissances de sécurité

Nous résumons ici les principaux résultats de cette phase. L'acquisition des connaissances est un domaine de recherche à part entière lié au champ d'Intelligence Artificielle. L'analyse de la sécurité au niveau des systèmes de transport automatisés est une tâche assez complexe qui requiert l'emploi efficace de méthodes appropriées d'acquisition pour rendre compte des différents acteurs impliqués dans le processus d'analyse de sécurité ainsi que de l'ensemble varié de connaissances utilisées.

2.1. Les acteurs du processus d'analyse de sécurité

La tâche d'analyse de sécurité du système de transport est assignée aux **experts de certification**. Ces experts du domaine du transport ont pour activité essentielle de rendre compte du degré de sécurité que procure le système de transport à ses passagers et sont amenés à étudier en

profondeur les documents et les dossiers du constructeur du système de transport dans leurs volets de sécurité (Mejri, 1995). Ces experts se basent alors sur leur expérience et leur savoir faire sur des systèmes analogues déjà promulgués et certifiés. Le **constructeur** du système de transport quant à lui, propose un ensemble de documents dans lesquelles il prouve la complétude de son étude en mentionnant un ensemble de fonctions de sécurité installées dans le système en vue de pallier aux différents incidents probables de sécurité. **L'exploitant** représente l'autorité qui va mettre en exploitation le système de transport (exemple : SNCF / RATP/…etc.). L'exploitant attend du système de transport un degré de confiance acceptable au niveau sécuritaire.

2.2. Le dossier de sécurité

Il s'agit simplement d'un document fourni par le constructeur du système de transport. Dans ce document, ce dernier essaye de défendre son système qu'il propose sur le plan de la sécurité. Notamment, l'étude du dossier de sécurité du constructeur est souvent appuyée par la proposition, par les experts, de scénarios d'insécurité pour lesquels le constructeur doit prévoir des fonctions de sécurité bien adaptées à l'incident ou à l'accident susceptible d'arriver.

1.1. Les scénarios d'insécurité

Un scénario d'insécurité ou encore un scénario contraire à la sécurité représente souvent une situation d'insécurité au niveau du système de transport automatisé. Cette insécurité, jugée potentielle devrait être résolue par la proposition d'une solution à adopter pour pallier au risque qu'elle présente (exemple : collision entre deux rames de métro ou déraillement de la voie, etc). Un scénario d'insécurité est un concours de circonstances spécifiques pouvant mener à un danger sur le plan de la sécurité. Il est décrit conjointement par des attributs de situation qui donnent une idée sur l'insécurité et des attributs de solution à adopter pour anéantir ou réduire le risque d'insécurité.

3. Représentation de la connaissance en sécurité

Les travaux antérieurs ont ciblé surtout sur la représentation statique d'un scénario. Nos travaux actuels visent l'intégration de plusieurs modes de représentation d'un scénario d'insécurité (description textuelle, description graphique et description dynamique). Vu la complexité et la

variété des connaissances employées en matière d'analyse de sécurité (connaissances d'ordre statique et dynamique, connaissances expertes et épisodiques, implicites et explicites…etc.), il a été judicieux de bien les représenter en vue d'une exploitation efficace. Ces connaissances sont représentées surtout par des scénarios d'insécurité. Un scénario d'insécurité est généralement décrit par des formes complémentaires de description (Hadj mabrouk et al 1994), (Mejri et al 1995) :

1.2. Description textuelle

Il s'agit simplement d'un texte illustratif qui explique le déroulement du scénario c'est-à-dire ses tenants et aboutissants. A titre d'exemple, nous donnons la description textuelle fournie par les experts pour le scénario N°34 et dont le nom est "***Echec d'effacement d'un élément secouru après dés initialisation***" :

Description :

1) Elément A est devenu muet (ne dialogue plus avec le PA : Pilote Automatique) et il est pris en charge par le Pilote automatique du tronçon n nommé Pan qui est en cours de lancer une initialisation par parcours de l'élément B en CM "conduite manuelle".
2) L'élément B accoste l'élément A et se met en CMS "Conduite Manuelle Secouru".
3) Alarme de désinitialisation.
4) Solution : Il faut vider la section de tout élément avant de procéder à une initialisation.

1.3. Description graphique

Elle consiste en un synoptique qui représente très schématiquement la situation indésirable. Elle illustre surtout les automatismes impliqués dans l'insécurité. Le synoptique tente de donner une idée très sommaire du déroulement d'une séquence non sécuritaire.

Elle illustre surtout les automatismes impliqués dans l'insécurité. Le synoptique tente de donner une idée très sommaire du déroulement d'une séquence non sécuritaire.

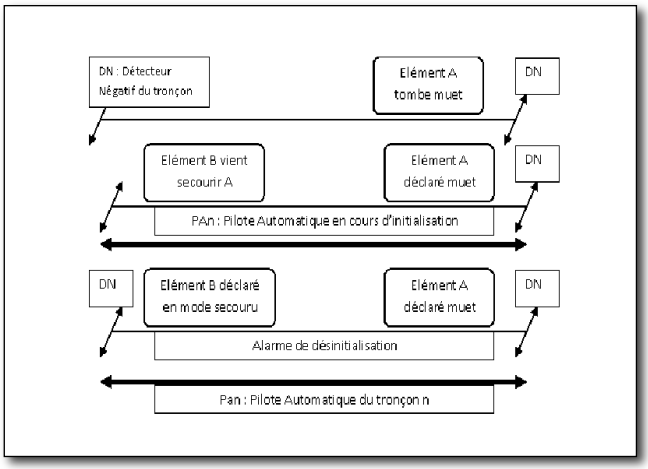


Figure 1 : Synoptique graphique du scénario d'insécurité N°34.

.

1.4. Description statique

C'est un ensemble de paramètres descriptifs du scénario sous forme d'une fiche <Attribut / Valeurs>. On recense alors à ce niveau plusieurs paramètres caractéristiques qui rendent compte de la description du risque / des fonctions de sécurité / des acteurs du scénario / de la zone géographique où se déroule le scénario / des pannes touchant une partie du système / des solutions adoptées pour anéantir le risque (voir figure2, figure 3).

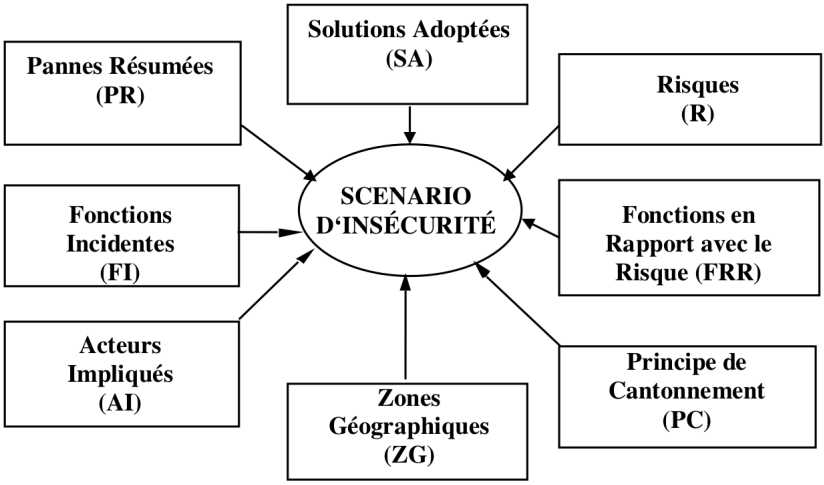


Figure 2 : Les paramètres de la description statique d'un scénario d'insécurité.

Les Détails concernant les différents paramètres de description d'un scénario d'insécurité et leurs valeurs possibles sont donnés en (Tableau 1) suivant.

Liste d'attributs	Liste de valeurs possibles	Concepts clés (*)
Principe de Cantonnement	Canton Fixe	*
	Canton Mobile	
Risque	Collision	*
	Déraillement	
	Chute dans le véhicule	
	Chute sur la voie	
	Électrocution	
Fonctions en Rapport avec le	Gestion de conduite automatique	

Risque		
	Localisation des Trains	
	Contrôle d'Entrée /Sortie	
	Autorisation CI/HT	
	Commutation de redondance	
	Initialisation	
	Conduite Manuelle	
	Gestion des alarmes	
	Accostage	
	Traction/accostage	*
Zones Géographiques	Terminus	
	Station	
	Ligne	*
	Zone Injection Rames	
	Limite de Tronçon	
Acteurs Impliqués	Nombre de Rames	2
	Opérateur au PCC	*
	Opérateur Itinérant	
	PA Avec redondance	
	PA Sans redondance	*
Fonctions Incidentes	Gestion des Itinéraires	
	Contrôle de Trafic	
	Consignes (consistance)	*
	Communications	
Pannes	PR52 Stationnement de	*

Résumées	train sur la voie	
	PR9 Pénétration dans un canton occupé	*
Solutions Adoptées	SA51 : Contrôler le Courant de traction lors d'un freinage d'urgence. Ouvrir les Disjoncteurs si nécessaire.	

Tableau 1 : Les détails des différents paramètres descriptifs d'un scénario d'insécurité.

Il est décrit conjointement par des attributs de situation qui donnent une idée sur l'insécurité et des attributs de solution à adopter pour anéantir ou réduire le risque d'insécurité.

1.5. Description dynamique

Cette forme de description donne une vue sur la dynamique de déroulement séquentiel dans le temps et dans l'espace du scénario d'insécurité. Cette représentation fait appel aux Réseaux de Pétri : RdP (figure 3) comme formalisme de représentation dans lequel les places représentent des états, les transitions représentent les possibilités d'évolution d'un état à un autre. Les Réseaux de Pétri possèdent des propriétés intéressantes pour la validation et la simulation d'une modélisation (Murata, 1989 :541-542) :

- Visuel, bon pour la compréhension.
- Preuves formelles de propriétés nécessaires pour un système (scénarios) : Choix d'architecture et de principes.
- Évolution du système :
  - Dynamique (conditions, durée).
  - Structures de contrôle : Parallélisme, synchronisation…
  - Répond aux besoins de comptage (ressources).
  - Extension à un réseau de Pétri de haut niveau avec des réseaux colorés et hiérarchisés.

Le Réseau de Pétri utilisé est non marqué et il est alors complété par un Tableau de Séquencement de Marquage qui permet de retracer

l'évolution depuis un marquage initial jusqu'à un marquage jugé critique sur le plan de la sécurité.

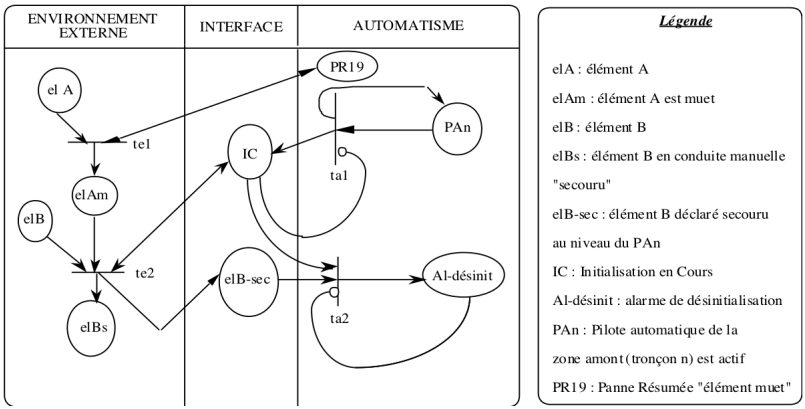


Figure 3 : Exemple de Réseau de Pétri d'un scénario d'insécurité N°34.

La lecture du réseau de pétri complétée par le tableau de marquage permet de mieux comprendre ce scénario d'insécurité :

- Un Elément A (el A) est devenu muet : el Am (ne dialogue plus avec le PA n : Pilote Automatique gérant le tronçon numéro n) suite à l'occurrence d'une Panne Résumée notée PR19 qui signifie que la rame est en panne au niveau du dialogue avec le PA.
- Il est pris en charge par le Pilote automatique du tronçon n nommé PA n qui est en cours de lancer une initialisation par parcours de l'élément B en CM "conduite manuelle".
- L'élément B (el B) accoste l'élément A (el A) et se met en CMS "Conduite Manuelle Secourue" noté el Bs et au niveau de l'interface avec l'automatisme un signal est perçu noté el B-sec.
- Une alarme de désinitialisation notée Al-désinit est alors amorcée.

Les petits ronds dans le réseau représentent la logique négative c'est-à-dire l'absence d'un jeton dans une place en amont de la transition, ce qui aurait pour effet d'avoir un jeton en aval de la transition. A titre d'exemple, dans la transition ta2, une alarme de désinitialisation se déclenche alors qu'elle était inactive avant ta2. Elle est donc engendrée par l'évènement IC

(Initialisation en cours) et le mode secouru dans lequel s'est mis l'élément B (elBsec).

La Solution à adopter est qu'il faut vider la section de tout élément avant de procéder à une initialisation.

4. Le système ACASYA

Ce problème d'analyse de sécurité a été déjà traité avec une première approche lors de nos travaux de thèse (Mejri et al, 1993), (Mejri et al, 1994), (Mejri, 1995). Le travail accompli avait abouti d'abord à la réalisation d'une maquette de faisabilité (Mejri et al, 1995) et à la mise au point d'un système d'acquisition et de classification des scénarios d'accidents baptisé ACASYA (Hadj mabrouk et Mejri, 1998). Les premiers résultats d'évaluation de nos travaux de recherche appliqués au domaine de l'analyse de sécurité (Mejri, 1995), nous ont permis de chercher ainsi à identifier un modèle d'acquisition de connaissances plus générique (Mejri et Hadj mabrouk, 2000) et une démarche plus générale pour la résolution de problèmes (Mejri & Caulier, 2005).

Dans l'objectif d'aider les experts dans leur activité d'évaluation de scénarios d'insécurité, ACASYA s'articule autour de deux étapes essentielles complémentaires :

- **Classer** d'abord automatiquement le scénario d'insécurité dans une classe prédéfinie à l'aide **d'un algorithme de classement**. Ceci afin de cibler sur un cadre de référence qui est la classe de scénarios d'insécurité au lieu de toute la base de scénarios.
- **Evaluer** ensuite le scénario d'insécurité proposé par le constructeur en référence à la classe d'appartenance trouvée dans l'étape précédente.

4.1. Classement des situations d'insécurité

Pour ce faire, une base de scénarios d'insécurité historiques a été élaborée. Cette base regroupe tous les scénarios plausibles et pertinents collectés auprès des experts et se rapportant à des systèmes de transport automatisés analogues. Nous avons alors réussi à collecter jusqu'à 70 scénarios d'insécurité qui ont été répertoriés en 11 classes (Mejri et al

,1995). Les classes ont été prédéfinies par les experts (Tableau 2). Le principe de regroupement des scénarios en classes reflète la démarche intuitive des experts du domaine. Nous n'avons pas réussi à identifier de manière explicite cette démarche mais nous avons réussi à la confirmer ensuite par l'algorithme d'apprentissage que nous avons mis en place. Cet algorithme a pu retenir les descripteurs clés sélectionnés par les experts comme descripteurs pertinents d'une classe.

Classes	**Scénarios d'insécurité**
C1 : Commutation de Redondance	S1, S8, S29, S30, S47
C2 : Séquence d'initialisation	S4, S5, S6, S21, S22, S23, S34, S66
C3 : Localisation des Trains	S3, S7, S17, S19, S25, S38, S45, S53, S63, S65
C4 : Gestion du freinage d'urgence	S9, S10, S20, S24, S50, S64
C5 : Accostage	S13, S14, S15, S32, S33, S35, S40
C6 : Gestion du Sens de marche	S2, S12, S16, S18, S55
C7 : Contrôle d'Entrée / Sortie	S11, S31, S36, S37, S39, S44
C8 : Suivi de l'ordre des trains	S26, S41, S42, S54
C9 : Conduite Manuelle	S27, S28, S43, S49, S51, S60
C10 : Contrôle/Commande des aiguilles	S52, S67
C11 : Contrôle de vitesse	S68, S69, S70
Autres : classes d'autres risques que la collision	S56, S57, S58, S59, S61, S62, S46, S48

Tableau 2 : Répartition des scénarios du Risque Collision en Classes.

Un **algorithme d'apprentissage automatique** par recherche de procédures de classification et développé par nos soins, a été appliqué pour déterminer les descriptions des classes en opérant sur l'ensemble des descriptions statiques des scénarios. Chaque classe de scénarios a été ainsi caractérisée par une description en termes des descripteurs les plus pertinents de la classe. Un descripteur n'est autre qu'un paramètre de description du scénario auquel est attachée une fréquence d'apparition dans la classe de scénarios. Le module de classification (figure 4) intègre deux algorithmes : le premier est un algorithme d'apprentissage de descriptions de classes de scénarios d'insécurité historiques, le second classifie automatiquement les scénarios d'insécurité proposés par les constructeurs. L'algorithme de classification calcule des fréquences d'apparition des paires <attributs/valeurs>.

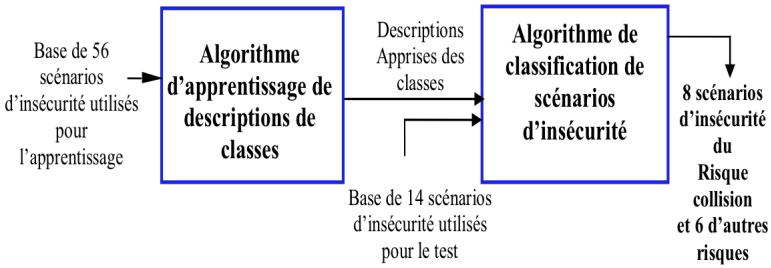


Figure 4: Bilan du module de classification.

- La base d'exemples d'apprentissage comporte une proportion de 4/5 (56 scénarios) de la base d'exemples totale (70 scénarios) ;
- La base d'exemples de tests comporte une proportion de 1/5 (14 scénarios) de la base d'exemples totale.

Les résultats sont positifs : 8 scénarios sont classés dans le risque de "collision" (frontale et/ou par rattrapage), les 6 scénarios restants sont relatifs à d'autres types de risque. Comme le montre l'histogramme (figure 5), les scénarios d'insécurité sont répartis entre 11 classes du risque collision et une classe notée "Autres" relative aux différents autres risques d'où la nécessité d'enrichir et de consolider la BHS par d'autres scénarios d'accident. Nous avons choisi volontairement de focaliser d'abord nos

efforts sur le risque de collision jugé le plus significatif par les experts du domaine.

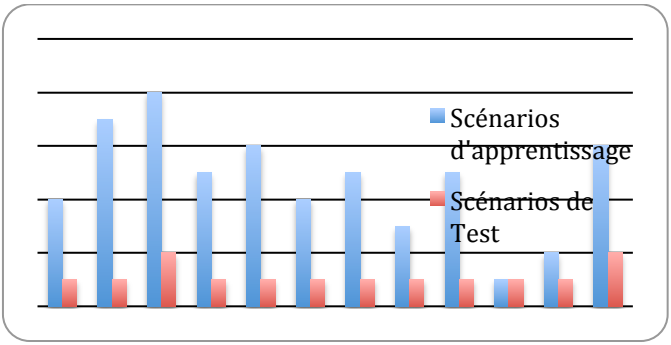


Figure 5: Histogramme de la répartition des scénarios d'insécurité en classes.

4.2. Evaluation des situations d'insécurité

Notre démarche avait consisté à évaluer un scénario d'insécurité après la phase de classification. Ceci avait pour but de restreindre l'espace d'évaluation uniquement à la classe d'appartenance du scénario détectée précédemment. L'évaluation consiste à tester la complétude et la cohérence d'un scénario d'insécurité. La complétude informe sur le fait que le scénario d'accident évalué regroupe tous les descripteurs nécessaires à une bonne définition. La cohérence renseigne sur l'intégrité du scénario d'insécurité évalué et sur le fait qu'il ne comporte pas de descripteurs inutiles ou superflus. L'évaluation débouche sur la détection des causes de l'insécurité et la proposition de solutions à adopter pour pallier à cette insécurité. En effet, un raisonnement causal est utile dans un diagnostic ou dans une évaluation (Mattson, 1997). Pour ce faire, un algorithme d'apprentissage de règles a été appliqué. Nous avons choisi l'algorithme CHARADE (Ganascia, 1987) pour induire un ensemble de règles d'évaluation à partir de la classe de scénarios d'insécurité repérée dans l'étape précédente par le classificateur. L'intérêt du système CHARADE est qu'il ne produit pas une base de règles isolées, c'est-à-dire des règles de classification de type :

Si conjonction de descripteurs Alors Classe d'appartenance.

Mais il permet carrément de produire par inférence inductive, un système de règles de complétion (Ganascia, 1990, 1991) de type :

Si conjonction1 de descripteurs Alors Conjonction2 de descripteurs.

Voici un exemple d'une règle induite par le système CHARADE :

If *Acteurs_impliqués = Opérateur_itinérant,*
Fonctions_Incidentes = Consignes
Acteurs_impliqués = Opérateur_en_Pcc.

Then *Pannes_ Résumé = PR11 (élément invisible sur une zone de conduite automatique intégrale),*
Acteurs_impliqués = PA_avec_redondance,
Fonctions_en_Rapport_avec_le_Risque =
Localisation_Trains, Zones_géographiques = Terminus[1]

Avec Coefficient de Certitude est égal à 1.

L'évaluation des situations d'insécurité se déroule alors en deux phases complémentaires :

a- Utiliser un moteur d'inférence d'un système expert dont la base de faits contient tous les descripteurs hors PR (hors Pannes Résumées) et dont la base de règles contient le (Système de règles concluant sur des causes) induit par le système d'apprentissage CHARADE. Cette phase permettrait en chainage avant de déduire toutes les pannes résumées causes de l'insécurité.

b- Utiliser le même moteur d'inférence avec la base de faits contenant cette fois ci tous les descripteurs hors SA (hors Solutions Adoptées) et dont la base de règles contient le (Système de règles concluant sur des remèdes) induit par le système d'apprentissage CHARADE. Cette phase permettrait en chainage avant de déduire toutes les solutions à adopter pour pallier à l'insécurité.

5. Vers une ontologie du domaine d'analyse de sécurité des Systèmes de Transport Automatisés

La démarche de résolution envisagée devrait vérifier les propriétés suivantes :

- Permettre la prise en compte **d'un modèle riche en connaissances pour la représentation de problèmes**
- Permettre **la capitalisation de l'ensemble des connaissances** du domaine. La démarche envisagée devrait permettre l'acquisition des connaissances épisodiques, en vue de les archiver d'abord et ensuite de dériver des connaissances plus synthétiques.
- Permettre **la réutilisation d'une résolution** antérieure de problème (ici réutiliser une analyse de sécurité déjà faite sur des systèmes de transport analogues).

D'où il serait très intéressant de proposer **une ontologie pour l'analyse de sécurité** et appuyée d'un modèle approprié de représentation et de spécification de problème. Le modèle de résolution de problème a fait déjà l'objet d'une publication dans la revue RTS (Mejri & al, 2009). Nous décrivons ici l'ontologie du domaine.

Une ontologie de domaine est une sorte de conceptualisation des différents aspects relatifs à la terminologie et le vocabulaire associé à un domaine d'expertise particulier. Parmi les objectifs d'une ontologie c'est d'uniformiser le langage et de mettre en place un cadre de référence pour favoriser une communication plus rigoureuse entre les différents acteurs du domaine. Dans notre cas, nous envisageons :

- Mettre en avant un cadre de référence terminologique pour le domaine de l'analyse de sécurité des systèmes de transport automatisés permettant de pallier aux problèmes liés à la multi expertise (présence de plusieurs experts) ;
- Permettre plus qu'une taxinomie de détecter des liens (Chandrasekaran, 1983) et des relations sémantiques entre les différents concepts du domaine. En effet, par exemple, la description statique et dynamique des scénarios d'accidents

semblent très **hétérogènes et n'ont au premier regard aucun lien** mais à travers le parcours de l'ontologie nous pouvons réaliser qu'il existe bien des liens.

- Favoriser ainsi, l'évaluation de la consistance et de la complétude des scénarios d'insécurité. La présence dans la description dynamique d'un scénario d'un concept particulier permettrait de vérifier la prise en compte dans sa description statique d'un certain nombre d'autres concepts liés et qu'on retrouve facilement à travers l'ontologie.
- Favoriser la génération de scénarios d'accidents. En effet, connaissant la description dynamique d'un scénario d'insécurité on pourrait générer facilement sa description statique et moins facilement inversement.

Actuellement, nous sommes dans une phase de construction de l'ontologie relativement au domaine de l'analyse de sécurité. Ce travail laborieux nécessite aussi une étude terminologique qui fixe les concepts et spécifie et rend plus explicite les termes et le vocabulaire employés en matière d'analyse de sécurité des systèmes ferroviaires. Nous donnons ici juste un état d'avancement de cette ontologie terminologique (ou onto-terminologie) à travers des extraits. Le travail n'est pas encore terminé. Il mérite à être plus poussé. Nous nous sommes limités volontairement au risque de collision jugé représentatif du domaine. Nous envisageons ensuite étendre l'ontologie à d'autres types de risques d'accidents.

La figure 6 suivante montre un premier extrait de l'ontologie du domaine permettant de mettre en place les liens entre la description statique et dynamique. En effet un risque d'insécurité (exemple la collision) qui est un attribut de la description statique de scénario est décrit dynamiquement par :

- Un environnement externe de l'insécurité. Il représente les éléments (rames de métro) impliqués dans l'accident ou l'incident ainsi que la zone géographique où se déroule le scénario.
- Des automatismes qui impliquent soit des fonctions de sécurité en rapport avec le risque ou des fonctions d'ordre secondaire ou aussi des acteurs humains ou automatiques. Les automatismes représentent un concept de la description dynamique alors que ses

fils sont donnés dans la description statique d'où l'intérêt de l'ontologie pour remédier à ces confusions dans le langage expert.

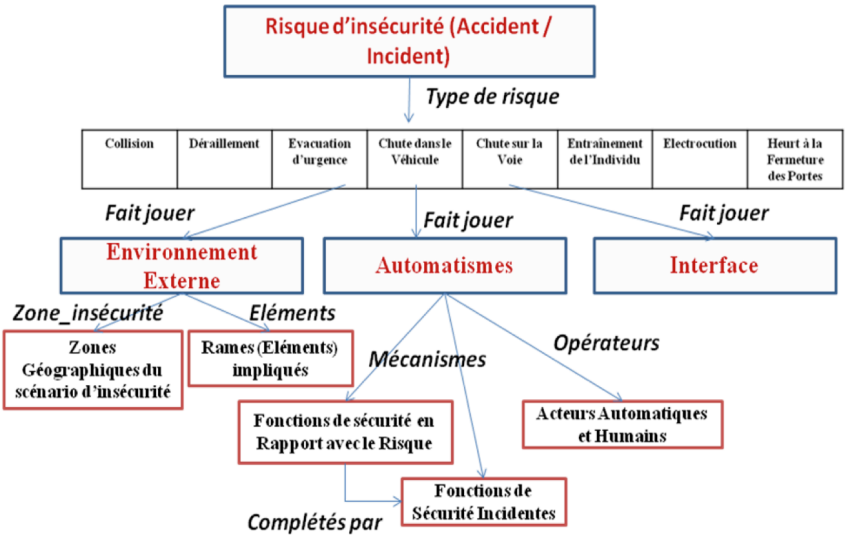


Figure 6 : Extrait N°1 de l'ontologie (Lien Statique/Dynamique).

Les figures 7 et 8 sont le fruit d'une étude approfondie sur les descriptions de la base de scénarios historiques en vue d'attacher chaque fonction de sécurité à l'ensemble des attributs des automatismes décrits dans les descriptions dynamiques. En effet, une fonction de sécurité n'est autre qu'une protection prévue et qui a été violée par le scénario d'accident et devrait se manifester par des automatismes. Les Figures 9 et 10 montrent d'une part, les liens entre les Acteurs Impliqués dans l'insécurité (attribut statique) et les automatismes (attribut dynamique) et d'autre part les liens entre Les zones géographiques du scénario et l'environnement externe du scénario d'insécurité.

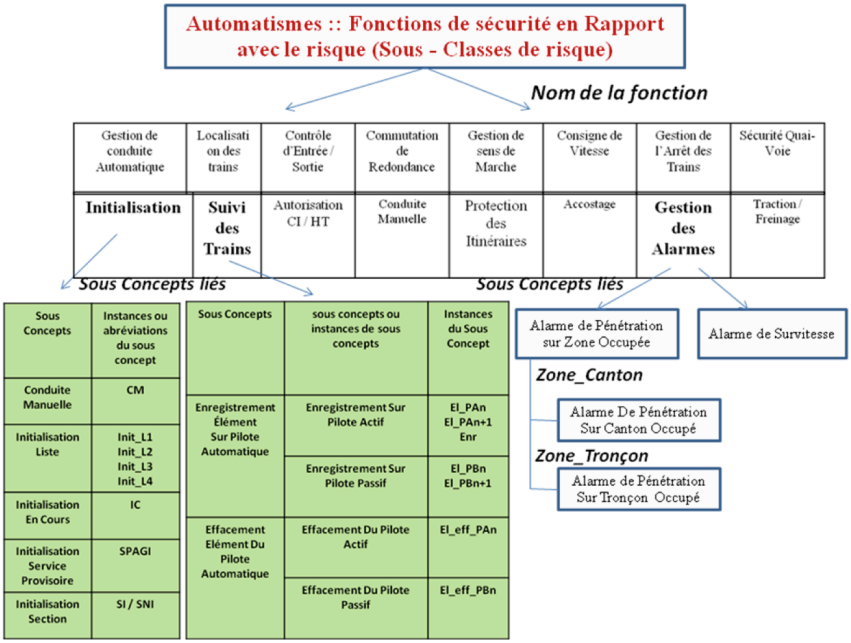


Figure 7 : Extrait N°2 de l'ontologie (liens Fonctions de sécurité / Détails automatismes).

Figure 8 : Extrait N°3 de l'ontologie (liens Fonctions de sécurité Incidentes / Détails automatismes).

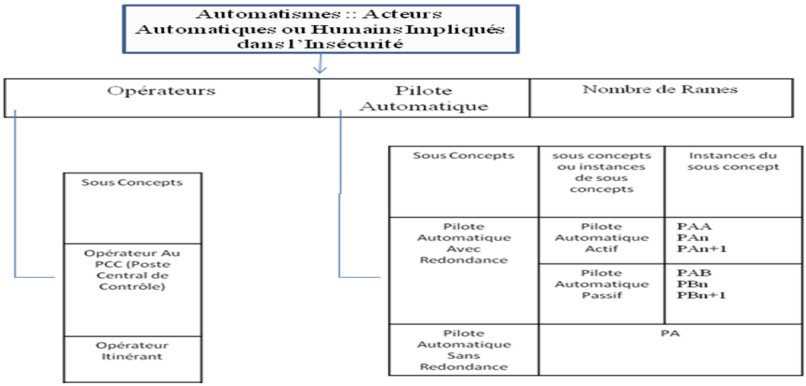


Figure 9 : Extrait N° 4 de l'ontologie (Liens Acteurs Impliqués / Automatismes)

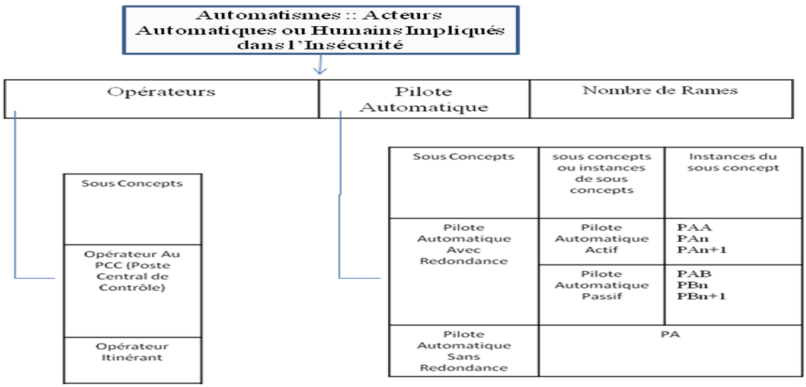


Figure 10 : Extrait N°5 de l'ontologie (liens Zones Géographiques / Détails Environnement Externe de l'insécurité).

6. Conclusion & perspectives

Ce papier s'est attaché à décrire un domaine d'étude lié à l'analyse de sécurité des systèmes de transport automatisés. Il a mis en évidence des modes de représentation de scénario d'insécurité (statique / dynamique/ textuelle / graphique). La notion d'ontologie a été introduite et son intérêt pratique a été avancé. Les principaux apports de l'ontologie du domaine sont :

- D'abord, il intègre deux vues intimement liées : une vue qui présente le **modèle de représentation statique** utilisé (Attributs Risque / Fonctions de sécurité en rapport avec le risque / Fonctions de sécurité incidentes / Acteurs Impliqués / Zones Géographiques / Pannes Résumées et Solutions Adoptées) et une vue qui présente **La représentation dynamique** (Environnement externe, les Automatismes et l'interface entre automatismes et environnement externe) ;

- Il met à contribution dans sa modélisation deux aspects complémentaires : statique et dynamique ;

- Il donne les liens entre l'aspect statique et dynamique de description. On a ainsi une vision complète du scénario accompagnée des connaissances qui permettent de l'interpréter, de l'expliquer ou même de l'argumenter s'il le faut.

Ce papier s'est aussi intéressé à présenter une démarche de représentation et d'extraction automatique des connaissances impliquées dans des scénarios. Cette démarche a pour intérêt principal de favoriser l'apprentissage par recherche de procédures de classification servant pour la résolution d'un nouveau problème ce qui assure la réutilisation. Plusieurs perspectives s'ouvrent à ce jour à nos travaux de recherche :

- Affiner notre modèle de représentation de scénario, notamment en tenant compte de la notion d'ontologie surtout dans le déroulement dynamique du scénario ;

- Affiner notre démarche de résolution de problèmes ;

- Compléter cette ontologie de domaine afin de tenir compte d'autres types de risque outre celui de la collision choisi volontairement pour la validation des travaux sur le système ACASYA.

Bibliographie

Chandrasekaran, B., (1983) : *Towards a Taxonomy of Problem Solving Types*. The AI Magazine, Winter/Spring, pp 9-17.

Ganascia, J.G., (1987) : *AGAPE et CHARADE : deux mécanismes d'apprentissage symboliques appliqués à la construction de Bases de Connaissances*. Thèse d'état, Paris sud, 27 mai 1987.

Ganascia, J.G., (1990) : *L'âme Machine : les enjeux de l'Intelligence Artificielle*. Le Seuil éd., janvier 1990.

Ganascia, J.G., (1991) : *L'apprentissage symbolique*, Encyclopédie de la communication - PUF, S/direction de Lucien SFEZ, 1991.

Hadj mabrouk, H., Darricau, M., Mejri, L., (2000) : *Contribution of Case-Based Reasoning to the Software Error Effect Analysis*. International Conference on Artificial and Computational Intelligence for Decision Control and Automation in Engineering and Industrial Applications. Tunisia, 22-24 March 2000, pp 83-89.

Hadjmabrouk, H., Mejri, L. ,Elkoursi, E.M., Bied-charreton, D., Letrung, B., Houriez, B., (1994) : *Base de scénarios d'accidents : résultats des travaux d'acquisition des connaissances pour le développement d'un système d'aide à l'examen des études de sécurité des systèmes de transport guidés*. Rapport de Convention, 03 1994.

Hadj mabrouk, H., Mejri, L., (1998) : *ACASYA: a knowledge-based system for aid in the storage, classification, assessment and generation of accident scenarios*. IEEE, Computational engineering in systems applications, Nabeul-Hammamet, Tunisie.

Kessentini, M., Bouchoucha, A., Mejri, L., Houari, S., (2009) : *Transforming sequence diagrams to colored petri nets using examples and heuristic search*. ASE 2009.

Kodratoff, Y., Diday, E., (1991) : *Induction symbolique et numérique à partir de données*. Cepadues éd., Toulouse.

Mattsson, M., Bosch, J., (1997) : *Framework Composition: Problems, Causes and Solutions*. In Proceedings of the 23rd International Conference in Technology of Object-Oriented Languages and Systems (TOOLS '97 USA).

Mejri, L., Hadj mabrouk, H., Elkoursi, M., Houriez, B. ,(1993) : *Un système expert d'aide à la génération de scénarios d'accidents basé sur l'apprentissage automatique*, ITTG 1993, Lille, France.

Mejri, L. ,Houriez, B., Millot, P., (1994) : *Automatic generation tool of accident sequences for automated transport systems security analysis*. International symposium on advanced transportation applications, Aachen, Allemagne.

Mejri, L., Houriez, B., Millot, P., (1995) : *Un système d'aide à la génération de scénarios Contraires à la sécurité*, Rapport de fin de contrat INRETS-LAMIH, Valenciennes, Janvier 1995.

Mejri, L., (1995) : *Une démarche basée sur l'apprentissage automatique pour l'aide à l'évaluation ET à la génération de scénarios d'accident. Application à l'analyse de sécurité des Systèmes de transport automatisés*, Mémoire de thèse de doctorat, Université de Valenciennes, Décembre 1995.

Mejri, L., Houriez, B., (1998) : *Apport des techniques d'intelligence artificielle pour l'aide à l'évaluation et à la génération de scénarios d'accidents*. Revue Technologies avancées du Centre de développement de technologies avancées d'Alger.

Mejri, L., Hadj mabrouk, H., (2000) : *Le concept de scénario. Application à l'analyse de sécurité des systèmes de transport guidés*. Journée Thématique Sécurité, Réseau Inter-Régional de Recherche Technologique dans les Transports Terrestres, GRRT, Villeneuve d'Ascq, 19 mai 2000. Publié dans les Actes de Journées.

Mejri, L. ,Caulier, P., (2005) : *Formalisation of a scenario concept to dynamic problem solving.*, 24 th European Annual Conference on Human Decision Making and Manual Control EAM, Athens, 17-19 October 2005.

Mejri, L., Zemzem, B., Caulier, P., **(**2008) : *Une approche de résolution de problèmes pour le web sémantique. Application à la découverte de services web*. Colloque annuel INFOL@NGUES II, Hammamet, Tunisie, juin 2008.

Mejri, L., Ben fraj, F., (2008) : *Multimodal and personalized information system : The most familiar itineraries*, IADIS International conference WWW/Internet 2008, short paper, Freiburg, Germany 13-15 October 2008.

Mejri, L., Hadj mabrouk, H., Caulier, P., (2009) : *Un modèle générique unifié de représentation et de résolution de problème pour la réutilisation de connaissances. Application à*

l'analyse de sécurité des systèmes de transport automatisés. Revue Recherche, Transport Sécurité, RTS N°103. (Accepté à paraître le mois de Février 2009)

Michalski, R. S., Kodratoff, Y., (1993) : *La recherche en apprentissage symbolique automatique : développements récents, classification des méthodes et perspectives*. In Michalski, R., Carbonell, J., Mitchell, T., and Kodratoff,Y., editors, Apprentissage symbolique, une approche de l'intelligence artificielle,volume 2, chapter 1, pages 1-27. Cepadues-Editions, Toulouse, France.

Murata, T., (1989) : *Petri Nets: Properties, Analysis, and Applications*. *Proceedings of the IEEE*, April 1989, 77(4):541–580,.

A propos des auteurs

Mejri Lassaâd

Laboratoire d'Automatique, de Mécanique et d'Informatique industrielles et Humaines. LAMIH CNRS UMR 8530

Dr en Informatique Industrielle et Humaine et Maître de Conférence à la Faculté des Sciences de Bizerte –Département Informatique

7021 JARZOUNA –BIZERTE – TUNISIE

Tél. : 216-98-362222 - Fax : 216-72-590566.-

E_mail : *mejrilassad@yahoo.fr*

Hadj-mabrouk Habib

INRETS : Institut National de Recherche sur les Transports et leur Sécurité.

HDR et chargé de recherche à l'INRETS de Paris

2 avenue du Général Malleret-Joinville, 94114 Arcueil Cedex – France.

Tél. : 01 47 40 73 52 - Fax : 01 45 47 56 06 –

Email : *hadj.mabrouk@inrets.fr*

Caulier Patrice

Laboratoire d'Automatique, de Mécanique et d'Informatique industrielles et Humaines. LAMIH CNRS UMR 8530

Maître de Conférences à l'Université de Valenciennes

Campus du Mont Houy F-59313 VALENCIENNES CEDEX 9

Email : *patrice.caulier@univ-valenciennes.fr*

DEMONSTRATIONS

Une « ontoterminologie » pour les interprètes de conférence – Un outil développé au sein de l'environnement académique

Elisa Veronesi, Franco Bertaccini

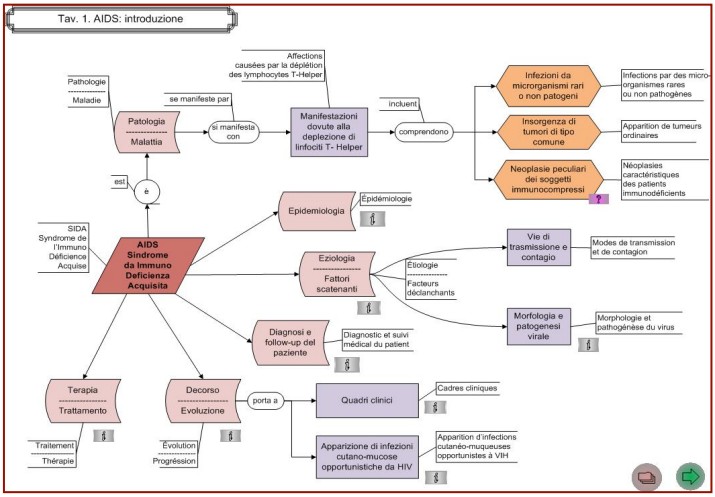
1. Introduction

Dans un monde de plus en plus mondialisé et multiculturel, l'interprétation et la traduction ont largement recours à la terminologie et à la création d'ontologies de domaine pour pouvoir développer des bases de données dynamiques, modulaires et interactives. Ce projet a donc pour but d'étudier l'usage d'arbres ontologiques complexes en tant que support à l'interprétation simultanée. Nous avons essayé de mettre la terminologie au service des interprètes en concevant une ontologie dynamique et modulaire pouvant être utilisée par les interprètes en cabine, au cours du travail. Pour ce faire, nous avons créé une conceptualisation formelle d'un domaine spécialisé développée à partir de ses termes et concepts fondamentaux. La méthodologie de sa structuration repose par conséquent sur la façon dont les termes (et les concepts auxquels ils font référence) sont reliés entre eux d'un point de vue sémantique et logique. Pour cette raison, ce support pourrait être défini comme une « ontoterminologie ».

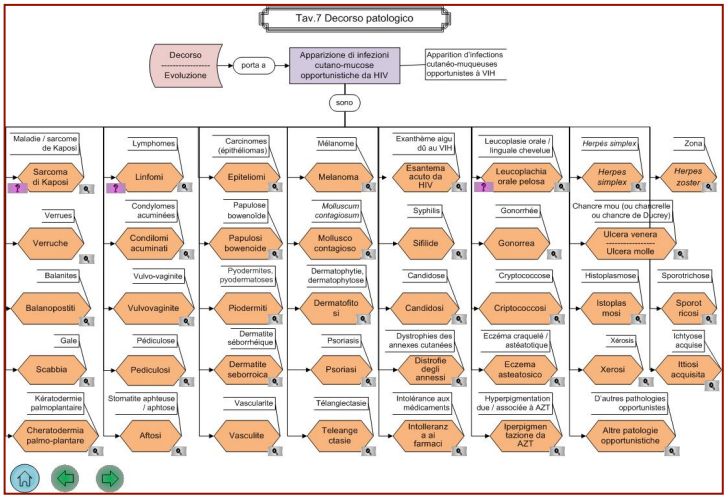
Cet outil linguistique et ontoterminologique a été développé au sein du milieu académique avec des supports informatiques qui n'étaient pas complètement adéquats et peu de ressources financières. Il nécessiterait donc d'être davantage formalisé et mis au point pour l'adapter à différents domaines d'application et être utilisé par des interprètes professionnels. Il représente toutefois la modélisation d'un hypothétique système à base de connaissances pour l'interprétation simultanée, qui nous permet d'en postuler le développement et la réalisation par des moyens appropriés. La raison en est que nous avons pu démontrer, par le truchement d'un essai empirique, que ce type d'outil serait tout à fait utile et d'appoint aux interprètes non seulement au cours de la phase de préparation au travail, mais aussi lors de la traduction simultanée en cabine.

2. L'outil ontoterminologique : fonctionnement

Ce travail a eu pour but l'étude d'un outil linguistique gérable avec un ordinateur portable comme support à l'interprétation simultanée. L'ontologie développée est destinée spécifiquement à des interprètes de conférence et, par conséquent, la création d'arbres modulaires et très spécialisés s'est imposée. Nous avons donc conceptualisé un domaine particulier et restreint par le truchement d'une ontologie régionale et structurée *ad hoc*. On s'est attachés en particulier à modéliser une ontologie régionale schématisant la terminologie relative au domaine des pathologies cutanéo-muqueuses se présentant au cours de l'infection par le VIH. Pour que les arbres reconstruisent les connaissances du domaine de façon très claire et exhaustive, le critère de structuration principal a porté sur le primat des concepts et des relations sémantico-logiques les reliant. En ce moment on ne dispose pas de logiciels ni de langages numériques à même de traiter des relations sémantiques plus complexes que les relations standards de subsomption (*is_a*) et de méronymie (*part_of*). Par conséquent, nous avons créé notre modèle manuellement, en utilisant le logiciel Microsoft Visio 2007, pour pouvoir représenter également les relations sémantiques transversales comme celles de cause-effet ou de temps et, par conséquent, toute irrégularité de la structure arborescente. En outre, nous avons utilisé l'option *Tool tip*, qui permet de faire apparaître un commentaire sur un mot au passage de la flèche de la souris, pour que l'interprète puisse visualiser les équivalents des termes le plus rapidement possible, mais seulement s'il en a besoin.



Page d'accueil de l'ontoterminologie, de type « top level » par rapport à la spécificité des concepts inclus (les équivalents apparaissent tous, pour mieux montrer la structure des arbres et le fonctionnement du support).



La feuille synthétisant l'évolution pathologique des maladies cutanéo-muqueuses dues au VIH. Le bouton gris avec une loupe permet d'accéder directement à la fiche relative à chaque pathologie recensée.

Pour rendre notre ontoterminologie bien navigable et facile à consulter, nous avons également inséré des liens intertextuels qui permettent à l'usager de se déplacer à l'intérieur des modules de la conceptualisation. Le bouton « Home » permet de revenir à la page d'accueil ; le bouton gris avec un « i » (« informations ») permet d'accéder directement à la conceptualisation spécifique à une instance particulière ; le bouton rose avec un point d'interrogation permet d'accéder à la fiche terminologique créée spécifiquement pour un terme plus intéressant et/ou plus difficile. Enfin, sur toutes les feuilles sont présentes les flèches « suivant » et « retour », pour pouvoir mieux se déplacer parmi les graphiques.

L'ontologie se compose en outre de plusieurs feuilles de travail et de différentes sections. Chaque page contient une partie de l'arbre ontologique du domaine. L'arbre se développe en suivant la démarche typique d'un discours scientifique et, de même, présente les instances et les concepts relatifs dans l'ordre de parution le plus probable d'un point de vue logique et conceptuel – ordre proposé aux experts et validé par ceux-ci. D'un point de vue graphique, l'ordre logique et chronologique prévu pour l'apparition des concepts est mis en exergue en utilisant toujours une forme et une couleur spécifique pour chaque niveau de la conceptualisation. De même, les fiches ontoterminologiques sont structurées et positionnées de façon à représenter le plus exactement possible l'évolution d'un discours de médecine très spécifique et dédié aux experts du secteur. L'ontologie comme nous l'avons structurée et développée n'est donc pas seulement une représentation conceptuelle graphique d'un domaine spécifique, mais elle suit aussi de près la démarche de la conférence à interpréter, pour faire en sorte que l'arbre ontologique soit un véritable support pour le travail de l'interprète.

3. A propos des auteurs

Veronesi Elisa

Adresse postale : via N. Buzzi 124, 48124 Piangipane / Ravenna (Italy);

Adresse électronique : everonesi@gmail.com

Bertaccini Franco

Adresse postale : SSLMIT - Scuola Superiore di Lingue Moderne per Interpreti e Traduttori C.so Repubblica, 136 Forlì (Italy) ;

Adresse électronique : franco.bertaccini@unibo.it

ITM, une infrastructure sémantique pour la maintenance du thésaurus multilingue Eurovoc

Thomas Francart, Charles Teissèdre

1. Introduction

Intelligent Topic Manager (ITM) est un outil propriétaire dédié à la gestion de connaissances. Dans le cas d'étude retenu pour la démonstration, l'outil sert de support logiciel pour maintenir et faire évoluer Eurovoc[1], un thésaurus multilingue (disponible en 25 langues), couvrant les différents domaines d'activité de l'Union Européenne (politique, géographie, agriculture, etc.). Ce thésaurus est utilisé par les services documentaires des administrations européennes et nationales et permet d'indexer des documents en les arrimant à un vocabulaire destiné à servir de norme. Sa constitution et sa maintenance soulèvent des difficultés, qui tiennent autant à son évolution constante, qu'à l'édition d'un vocabulaire commun à différentes langues, ce qui renvoie au problème de la traduction, dont ici le paradigme retenu (qui ne va pas de soi) est que les concepts sont partagés par tous et que les termes qui les désignent sont spécifiques à chaque langue. Les fonctionnalités d'ITM utilisées dans ce contexte peuvent être réparties en quatre grandes catégories, chacune répondant à des besoins spécifiques pour la maintenance du thésaurus dans un contexte collaboratif et partagé.

2. Un moteur de stockage "sémantique"

ITM s'appuie sur un moteur de stockage de données dites "sémantiques", qui recouvrent une superposition de modèles de représentation des connaissances. Dans le cadre de la gestion d'un thésaurus, on distingue : (1) le vocabulaire proprement dit (les concepts, les termes qui leur sont associés, les synonymes, les relations hiérarchiques ou transversales entre concepts, etc.) ; (2) le modèle du thésaurus, qui s'apparente ici à une transposition étendue de la norme SKOS[2], où les notions de termes et de concepts sont distinguées, afin d'une part de permettre de définir des relations structurelles au niveau des concepts et

d'autre part de définir des relations sémantiques entre les termes attachés aux concepts (traduction, abréviation, terme préférentiel pour l'affichage, synonyme, etc.) ; (3) le modèle qui décrit les opérateurs permettant de construire formellement le modèle du thésaurus (classes, types d'attributs, cardinalités, etc.) ; (4) le modèle de réseau sémantique sur lequel reposent les objets de modélisation : ce modèle propriétaire est un modèle formel proche, mais étendu, des Topics Maps[3], qui manipule des topics, des associations, des rôles, des attributs et des métadonnées.

Pour la maintenance d'Eurovoc, un moteur d'inférence est utilisé pour valider des contraintes d'intégrité sur le thésaurus – contraintes paramétrables qui vérifient que les connaissances produites se conforment à différentes règles éditoriales (par exemple, que la traduction d'un terme a bien une langue cible différente de la langue source). Le moteur de stockage d'ITM fournit également une gestion de métadonnées administratives non-fonctionnelles, afin de gérer la traçabilité de chaque valeur d'attribut : date de création et de dernière modification, utilisateurs qui ont créé et modifié pour la dernière fois une valeur. Une piste d'audit complète associée à ces métadonnées permet de suivre l'utilisation de l'application ("qui a fait quoi quand ?"). Cette gestion implique par exemple de ne pas supprimer physiquement les informations qu'un contributeur a supprimées, mais seulement de les "marquer" comme telles, la suppression effective n'intervenant qu'à la validation par un administrateur.

3. Des fonctions dédiées à la gestion de thésaurus et d'ontologies

A un niveau plus fonctionnel, la démonstration du logiciel s'attache à illustrer les fonctionnalités dédiées à la maintenance et à l'évolution d'une base de connaissances et de sa modélisation, qui recouvre ce que la communauté du Web Sémantique désigne sous le terme d'"ontologie", soit, pour Eurovoc, le modèle du thésaurus, sa structure et les éléments qu'elle permet d'articuler : les notions de hiérarchie arborescente ou de réseaux de concepts, les notions de termes associés, de termes équivalents, de définition, de synonymes, etc. ITM propose différentes interfaces dédiées à la maintenance d'un thésaurus : navigation hiérarchique dans des arborescences de concepts, interfaces de recherche croisant plusieurs critères, possibilité d'éditer, d'ajouter, de supprimer, de fusionner, de déprécier des termes, ou encore la possibilité d'associer plusieurs concepts pères à un concept donné (poly-hiérarchie).

De manière générale, l'ergonomie des interfaces s'efforce de rendre transparent le formalisme des connaissances manipulées en mettant en avant les fonctionnalités utiles aux utilisateurs finals (dans le cadre d'Eurovoc, des terminologues et des traducteurs), qui n'ont ainsi pas à se soucier des modèles de représentation sous-jacents.

4. Un environnement collaboratif

ITM distingue plusieurs profils d'utilisateurs qui correspondent à différents droits. En particulier, pour Eurovoc, un profil d'utilisateur spécifique a été créé : celui de "contributeur". Un contributeur effectue des modifications dans le cadre des tâches qui lui sont affectées par des administrateurs - modifications qui doivent alors être validées par ces derniers pour intégrer la version stable du vocabulaire. L'administration des droits joue ainsi essentiellement sur différentes déclinaisons de trois paramètres : l'étendue de la visibilité sur les connaissances, les droits d'édition des connaissances visibles, et les accès aux fonctionnalités d'administration (boutons visibles ou masqués).

ITM dispose de fonctionnalités de validation des modifications des contributeurs qui s'appuient sur les capacités de traçabilité de son moteur de stockage. Cela se traduit par des écrans de revue et de validation des modifications par les administrateurs. Cela permet de conserver, à tout instant, une version du thésaurus stable, et de n'y inclure progressivement que des modifications validées. Ce workflow permet aux éditeurs d'Eurovoc de contrôler les versions successives du thésaurus publié.

5. Des capacités d'intégration et de synchronisation des données

ITM propose différents outils pour faciliter son intégration dans des systèmes d'informations plus vastes (portails intranet, chaînes de traitements de l'information), en exposant notamment une API ainsi que des web services pour manipuler ses fonctionnalités (création, interrogation, génération de rapports, etc.). Il permet une synchronisation des données avec d'autres agents logiciels notamment par des outils dédiés à l'importation et l'exportation des connaissances dans différents formats (RDF, OWL, SKOS, Excel, XML, XTM). Pour Eurovoc, un export du thésaurus en SKOS est réalisé vers un portail Web de diffusion.

Enfin, l'application permet de générer des tableaux Excel, selon des modèles paramétrables, destinés à servir de support pour la revue des comités qui assurent le suivi des évolutions du vocabulaire.

L'utilisation d'un outil basé sur la sémantisation des données, ainsi qu'un export en SKOS, vont permettre l'interconnexion avec d'autres vocabulaires (Agrovoc, GEMET, etc.). La création de correspondances entre ces vocabulaires et leur publication en RDF seront un élément important dans le paysage des données sémantiques interreliées sur le web, les "linked data".

A propos des auteurs

Francart Thomas
Mondeca
Directeur technique
3, cité Nollez 75 018 Paris
thomas.francart@mondeca.com
http://mondeca.wordpress.com

Teissèdre Charles
Mondeca
Ingénieur de recherche
3, cité Nollez 75 018 Paris
charles.teissedre@mondeca.com

Approche onomasiologique de la phraséologie transdisciplinaire des écrits scientifiques : la recherche sémantique dans les textes dans le cadre du projet Scientext

Falaise Achille, Tutin Agnès

Résumé : L'accès à la phraséologie, en particulier pour les applications en langue étrangère et seconde, se fait rarement à partir de corpus. Dans le cadre du projet ANR Scientext, nous avons élaboré un mode d'accès à la phraséologie transdisciplinaire des écrits scientifiques par un mode onomasiologique. Des requêtes prédéfinies portant sur la question linguistique du positionnement et du raisonnement ont été élaborées à partir de schémas syntaxiques et sémantiques, par exemple sur l'expression de l'opinion ou de l'évaluation. Ces grammaires sont ensuite appliquées à un large corpus d'écrits scientifiques balisé au plan structurel et au plan syntaxique (analyse de dépendance). L'utilisateur peut ainsi extraire, selon ses besoins, une phraséologie adaptée à une requête sémantique.

1. Introduction

Dans cette démonstration, nous souhaitons présenter les modes d'accès aux informations lexicales et phraséologiques élaborés dans le cadre du projet ANR Scientext 2007-2010 "Corpus et outils de la recherche en sciences humaines et sociales" que nous pilotons au LIDILEM (U-Grenoble 3).

Dans ce cadre, un site web (*www.u-grenoble3.fr/lidilem/scientext*), qui permet un accès aux écrits scientifiques, a été élaboré. Nous souhaitons mettre l'accent dans cette démonstration sur l'accès sémantique à la phraséologie des écrits scientifiques qui constitue une originalité de notre projet (Tutin, à paraître).

2. Trois modes d'accès aux textes

2.1 Le corpus

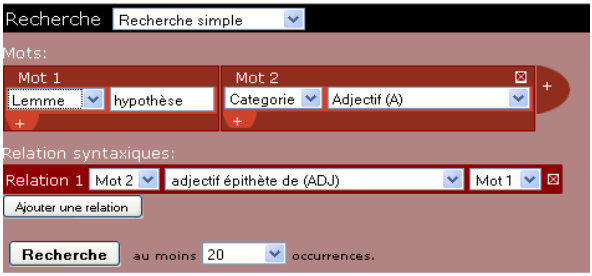
Dans le cadre de ce projet, un large corpus d'écrits scientifiques, a été constitué[1]. A ce jour, il contient 4,3 millions de mots dans des disciplines variées (linguistique, psychologie, sciences de l'éducation, traitement automatique du langage, médecine, biologie, mécanique, électronique) et dans des sous-genres variés (articles et communications écrites, thèses de doctorat, mémoires d'habilitation à diriger des recherches).

Le corpus a été annoté structurellement, en suivant les recommandations de la TEI Lite, et analysé syntaxiquement à l'aide de l'analyseur de dépendances Syntex, développé par Didier Bourigault (2007).

2.2 Les modes d'accès aux textes

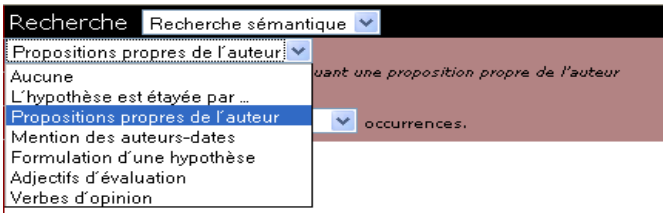
Une fois le corpus sélectionné selon les disciplines, les genres textuels et les parties textuelles désirés, l'utilisateur peut accéder au contenu du texte par trois types de recherche :

- **Un mode simple et guidé** avec des ascenseurs permet à l'utilisateur de sélectionner des formes, lemmes et/ou catégories, ainsi que les relations syntaxiques désirées. La figure ci-dessous permet ainsi d'extraire les occurrences de *hypothèse* avec un adjectif épithète.

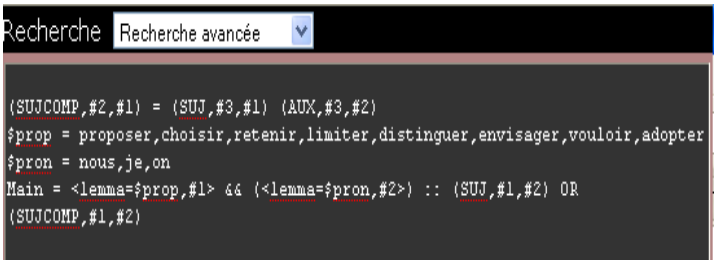


[1] Une large partie de ce corpus sera disponible pour la communauté de recherche.

- **Un mode sémantique** permet d'accéder à des occurrences en corpus, à partir de grammaires prédéfinies. Les grammaires sont définies à l'aide d'un langage de requête, ConcQuest, défini par Olivier Kraif, et étendu par nous.



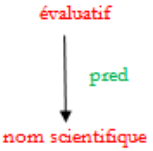
- **Un mode complexe** permet d'accéder à des occurrences en corpus, à partir de grammaires, utilisant les dépendances syntaxiques, les relations linéaires et des variables.



3. Le mode sémantique

Le mode sémantique suit des schémas sémantico-rhétoriques, qui sont ensuite traduits dans le langage de requête, à l'aide de variables et de dépendances syntaxiques.

A titre d'exemple, voici le schéma simple utilisé pour l'évaluation adjectivale :



On lui fera correspondre une grammaire utilisant pour chaque notion (en rouge) un ensemble de lexèmes, alors que la relation pred (en vert) sera traduite par la relation syntaxique épithète ou attribut, comme indiqué ci-dessous.

```
//TITRE: Adjectifs d'évaluation

//INFO: Les adjectifs d'évaluation qui portent sur les noms scientifiques

(ATTRIB,#2,#1) = (SUJ,#3,#1) (ATTS,#3,#2) ;
$eval=acceptable,adéquat,aisé,ambitieux,approximatif,bon,clair,classique,cohérent,complexe,concis,confus,convaincant,correct,crucial,déterminant,difficile,discutable,encourageant,épineux,essentiel,excellent,faible,fin,flou,fondamental,important,innovant,intéressant,irréprochable,judicieux,majeur,mauvais,meilleur,important,pertinent,nouveau,original,passable,passionnant,performant,principal,prometteur,riche,rigoureux,satisfaisant,séduisant,sérieux,significatif,solide,souhaitable,stimulant,vague,valable
$theo=
analyse,approche,article,caractéristique,choix,communication,concept,contribution,critère,élément,étude,exemple,facteur,fonction,idée,méthode,modèle,notion,objectif,phénomène,problème,projet,proposition,qualité,question,réflexion,résultat,solution,test,théorie,travail

Main = <lemma=$eval,#1> && <lemma=$theo,#2> :: (ATTRIB,#1,#2) OR (ADJ,#1,#2);
```

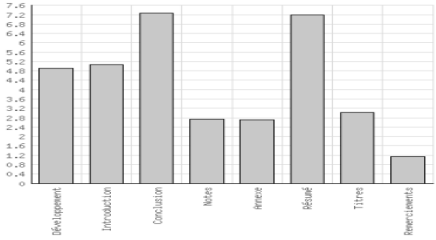
Une fois la requête lancée, il sera possible d'obtenir un affichage des occurrences dans un concordancier, d'exclure des réponses non pertinentes (toujours possibles dans une extraction automatique) et d'obtenir des extractions statistiques.

Au total 2068 occurrences ont été trouvées.

Liste des lemmes

Lemme	occurrences
meilleur résultat	53
proposition principal	44
objectif principal	44
bon résultat	36
problème majeur	29
principal caractéristique	28
bon qualité	27
principal résultat	25
modèle classique	23
élément essentiel	22
meilleur qualité	21
problème complexe	21
analyse fin	21
facteur important	20
résultat intéressant	19
caractéristique principal	19
phénomène cohérent	19
résultat satisfaisant	18

Partie textuelle	Nombre absolu d'occurrences		Nombre de mots total		Nombre relatif d'occurrences
Développement	1789	/	3645711	=	4.91 ‰
Introduction	106	/	209266	=	5.07 ‰
Conclusion	62	/	85200	=	7.28 ‰
Notes	42	/	153355	=	2.74 ‰
Annexe	32	/	118198	=	2.71 ‰
Résumé	18	/	25061	=	7.18 ‰
Titres	17	/	56124	=	3.03 ‰
Remerciements	2	/	17551	=	1.14 ‰



4. Bibliographie

Bourigault Didier (2007). *Un analyseur syntaxique opérationnel : SYNTEX*. Mémoire d'Habilitation à Diriger les Recherches, Toulouse.

Kraif Olivier (2008). Comment allier la puissance du TAL et la simplicité d'utilisation ? l'exemple du concordancier bilingue ConcQuest, *Actes des 9ème Journées d'analyse statistique des données textuelles, JADT 2008*. Lyon: Presses universitaires de Lyon: 625-634.

Tutin Agnès (2010). Showing phraseology in context: an onomasiological access to lexico-grammatical patterns in corpora of French scientific writings, *Proceedings of eLexicography in the 21st century: new challenges, new applications, 22-24 october 2009, Louvain la Neuve.*

A propos des auteurs

Falaise Achille

Laboratoire GETALP, LIG

Post-doctorant en informatique

Thèmes de recherche : TAL, traduction automatique, écrits électroniques, IHM

GETALP-LIG

385 rue de la Bibliothèque - BP 53

38041 Grenoble Cedex 9

achille.falaise@imag.fr

http://www-clips.imag.fr/geta/User/achille.falaise/

Tutin Agnès

LIDILEM, Université Grenoble3-Stendhal

Maître de conférence (HDR) en linguistique
Thèmes de recherche : linguistique de corpus, phraséologie, écrits scientifiques, TAL

UFR des Sciences du Langage
BP 25 - 38040 Grenoble cedex 9

agnes.tutin@u-grenoble3.fr

www.u-grenoble3.fr/tutin

Ontoterminologie : méthode et mise en œuvre

Marie Calberg-Challot, Christophe Tricot

1. Contexte

Les organisations manipulent de plus en plus d'informations. Comment gérer toutes ces informations ? Comment faire face à autant de données ? Comment maîtriser un espace informationnel ?

L'association de l'ontologie et de la terminologie a abouti à l'introduction d'une nouvelle notion, celle de l'ontoterminologie (Roche 2008). En explicitant et en prenant en compte les deux dimensions conceptuelle et linguistique de la terminologie (Calberg-Challot *et al.* 2010, Calberg-Challot *et al.* à paraitre), ce nouveau paradigme offre de nouvelles perspectives pour l'opérationnalisation des terminologies, que ce soit à des fins de recherche d'information, d'aide à la traduction, de cartographie sémantique, de gestion documentaire ou de capitalisation des connaissances.

Nous commencerons par une présentation d'Onomia dont le cœur de métier est l'accompagnement, la construction de référentiels (ontologie, terminologie, ontoterminologie, dictionnaire, thésaurus…) et la cartographie des savoirs pour aider ses clients à gérer et à mieux appréhender leurs connaissances.

Nous présenterons ensuite l'outil iMap d'Onomia pour la construction de l'ontoterminologie, c'est-à-dire la définition du réseau conceptuel, des termes et des informations associées. Nous terminerons par la présentation des logiciels de visualisation Codex pour la définition d'encyclopédies métier et Index pour la recherche sémantique d'information.

2. La société

Onomia[1] est une entreprise innovante qui fédère les compétences d'une équipe pluridisciplinaire possédant une forte expérience. En unifiant les apports de la linguistique, de l'intelligence artificielle, des sciences cognitives et de la logique, Onomia propose une approche unique.

L'expertise des équipes d'Onomia est reconnue dans le domaine de l'ingénierie des connaissances et de la terminologie au travers de leurs interventions, publications et des formations qu'ils dispensent[2].

Onomia accompagne ses clients dans leur démarche de capitalisation et de valorisation des savoirs pour répondre à leurs problématiques de gestion des connaissances comme le développement du patrimoine intellectuel, la formation des collaborateurs, l'amélioration de la communication ou la recherche d'information.

Cette démarche de capitalisation passe par l'explicitation des savoirs, l'identification des besoins de développement en assurant la consistance de la documentation, l'optimisation de la transmission des savoirs au travers de supports pédagogiques adaptés, la prise en compte de la diversité langagière et l'adaptation du vocabulaire aux différents métiers de l'entreprise, l'identification de documents pertinents indépendamment de leur langue ou de leur source.

3. Démarche

La méthode Onomia de construction d'ontoterminologies met l'accent sur la conceptualisation du domaine et sur le rôle des experts. Elle repose également sur le fait qu'il existe une différence entre cette conceptualisation et les discours scientifiques et techniques auxquels elle peut donner lieu.

[1] http://www.onomia.fr/

[2] http://www.onomia.fr/recherche/publications/

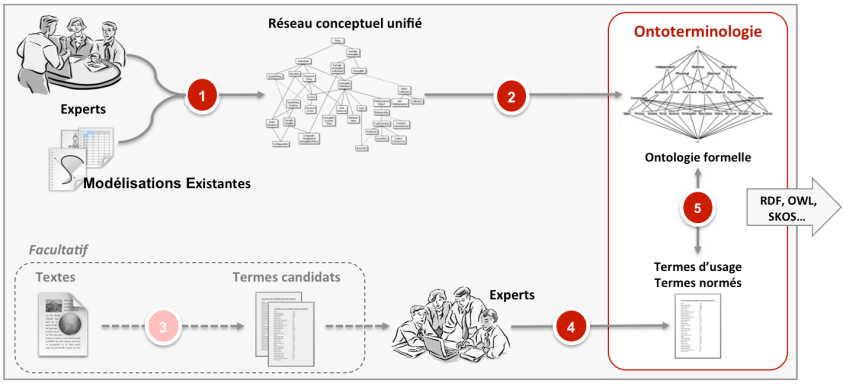


Figure 1. Méthode Onomia

Cette méthode se déroule en cinq étapes et s'appuie sur les outils Onomia de gestion et de construction d'ontoterminlogies :

La première étape est la construction du réseau conceptuel unifié. L'objectif est, avec l'aide des experts et des modélisations existantes, de construire une première conceptualisation du domaine. Cette phase est indépendante de tout langage de représentation (OWL, RDF, KIF) et s'appuie sur des principes épistémologiques. Elle repose sur les trois niveaux de langue : langue naturelle, langue de l'intellection et langages de représentation.

A la suite de la construction du réseau conceptuel est extrait une ontologie formelle sur la base de principes épistémologiques en s'appuyant principalement sur la définition par différenciation spécifique (Roche 2001). Le but est d'obtenir une ontologie consensuelle, cohérente et partageable.

Si on ne peut extraire une ontologie à partir de textes (la structure lexicale se superpose pas avec la structure conceptuelle du domaine) (Roche 2007), le traitement automatique de corpus est source de nombreuses informations linguistiques portant en particulier sur les termes d'usage.

Les experts identifient à partir de l'extraction de termes candidats les termes d'usage et les termes normés et proposent le cas échéant de nouveaux termes (« néotermes »).

La dernière phase de la démarche permet d'associer les concepts et les termes (normés et d'usage) en distinguant les définitions formelles des concepts des explications linguistiques des termes. L'ontoterminologie est exportée selon différents formats : RDF, OWL, SKOS...

Notons que, dans la mise en œuvre de la méthode, les tâches sont cycliques et interconnectées et le déroulement de la méthode se fait davantage de manière itérative. Cette méthode a été illustrée dans divers milieux industriels et a fait l'objet de plusieurs publications (Calberg-Challot et al. 2008, 2010).

4. Quelques applications

Tous les environnements logiciels Onomia reposent sur une architecture orientée services (SOA), une programmation objet et sur des formats d'échanges issus du W3C (langages XML). Les bases de connaissances sont au format RDF.

L'environnement de développement d'ontologies par différenciation spécifique repose de plus sur une architecture multi-agents.

Dans ce contexte, Onomia propose deux produits reposant sur cette architecture SOA : Codex et Index d'Onomia.

Vision globale de l'ensemble des savoirs, exploration et navigation au travers de cartes et schémas, supports pédagogiques et plateforme ergonomique et collaborative sont autant de fonctionnalités offertes par l'encyclopédie des savoirs d'Onomia, Codex.

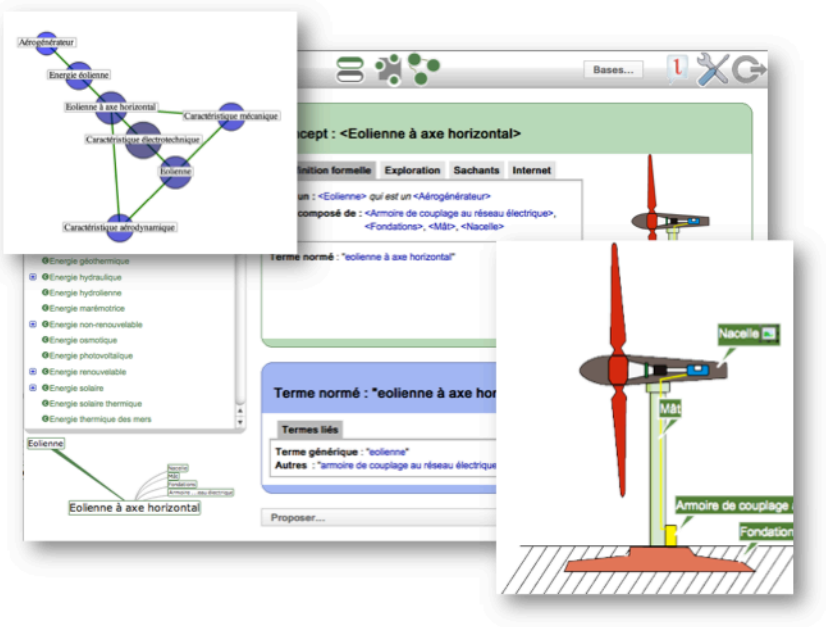


Figure 2. Codex d'Onomia : encyclopédie des savoirs

Index, quant à lui, permet une interrogation des moteurs internes et externes (méta moteur), une optimisation de la pertinence, une expression de la requête en langage naturel ou via des cartes et une cartographie des résultats.

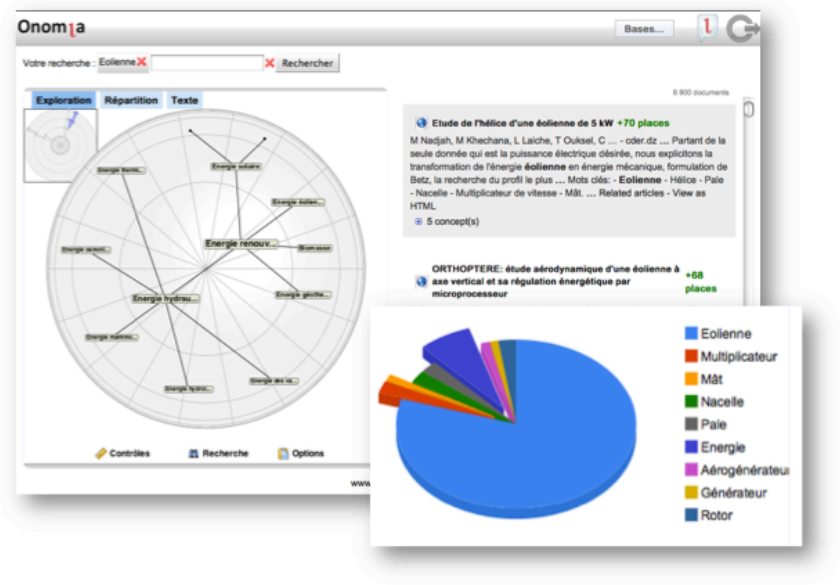


Figure 3. Index d'Onomia : recherche sémantique

Nous avons ainsi montré à travers les outils de construction et d'exploitation d'Onomia le rôle central et nécessaire pour une capitalisation pérenne des connaissances que joue l'ontoterminologie.

Références bibliographiques

Calberg-Challot Marie, Candel Danielle & Roche Christophe (2008) : "De la variation des usages au consensus terminologique : vers un dictionnaire de l'ingénierie nucléaire", *Actes de la première conférence TOTh 2007, Terminologie & Ontologie : Théories et applications*, Christophe Roche éd., Annecy, Institut Porphyre, pp. 199-141

Calberg-Challot Marie, Pierre Lerat, Christophe Roche (2010) : "Quelle place accorder aux corpus dans la construction d'une terminologie ?", *Actes de la troisième conférence TOTh 2009*, Terminologie & Ontologie : Théories et applications, Christophe Roche éd., Annecy, Institut Porphyre.

Calberg-Challot Marie, Christophe Roche, Luc Damas (à paraitre) : "Définition du terme vs. définition du concept", *Actes des 8e Journées scientifiques du réseau LTT 2009,* « Passeurs de mots, passeurs d'espoir : lexicologie, terminologie et traduction face au défi de la diversité » Lisbonne, 15-17 octobre 2009.

Roche Christophe (2001) : "The 'Specific-Difference' Principle : a Methodology for Building Consensual and Coherent Ontologies", *Actes de la conference IC-AI'2001*, Las Vegas , USA

Roche Christophe (2007) : "Dire n'est pas concevoir", IC 2007 : 18e Journées Francophones d'Ingénierie des Connaissances, Grenoble 2-6 juillet 2007

Roche Christophe (2008) : "Le terme et le concept : fondements d'une ontoterminologie", *Actes de la première conférence TOTh 2007, Terminologie & Ontologie : Théories et applications*, Christophe Roche éd., Annecy, Institut Porphyre, pp. 1-22, 2007

Libellex, plateforme de travail multilingue et référentiel terminologique d'entreprise

François Brown de Colstoun

Estelle Delpech

Libellex est une plateforme de travail multilingue pour tous les employés de l'entreprise, elle permet des rédactions et traductions exactes et sans fautes quelle que soit la langue. L'environnement est simple et intuitif, aucune formation n'est nécessaire. L'accès se fait par un navigateur via l'intranet de l'entreprise.

1. Le besoin : l'information interne

Le développement du nombre de documents créés et reçus dans une entreprise donne lieu à une problématique : il devient de plus en plus difficile de classer, stocker et surtout retrouver ces documents dans des délais et avec un effort raisonnables. Ce problème est amplifié pour les grands groupes : une multitude de collaborateurs travaillent sur des sites différents et dans des langues différentes. Au résultat la connaissance est dispersée, inutilement recréée et souvent inaccessible.

Une réponse à ce problème consiste à considérer ces documents comme de l'information non structurée et utiliser des moteurs de recherche internes. Pour cela, il faut que le moteur ait accès à toutes les aires de stockage de l'entreprise ; la connaissance récupérée n'est structurée que le temps de sa lecture, et pour le seul bénéfice de celui qui effectue la recherche.

Une autre réponse consiste à structurer l'information dans des bases documentaires ; l'indexation et les mots-clés utilisés pour cela dépendent de la compétence et du bon vouloir du rédacteur.

Ces deux solutions donnent satisfaction dans certains cas mais ont pour inconvénient d'être lourdes en fonctionnement et de ne pas modifier l'information, qui reste ainsi statique et vieillit.

2. L'offre Libellex : une mémoire collective dans l'entreprise

Pour développer une intelligence collective, il faut une mémoire collective, vivante. La plateforme Libellex est un ensemble d'outils permettant de développer, d'entretenir et d'utiliser une mémoire d'entreprise. Cette mémoire indexe automatiquement les textes, les phrases, les expressions et les mots, ainsi que leurs traductions. Ces textes peuvent prendre les formats Word, PDF, HTML, ou XML et être rédigés dans les langues suivantes : Français, Anglais, Allemand, Espagnol, Portugais, Italien ou Néerlandais.

3. La valeur de Libellex repose sur trois piliers :

Une technologie unique d'analyse linguistique de chaque document, ce qui permet leur indexation dynamique à de multiples niveaux : mot, groupe de mots, phrase, document, groupe de documents de thématique similaire, quel que soit le format ou la langue des documents.

L'interface utilisateur est la plus simple envisageable : une fenêtre de navigateur en Intranet. L'accès est immédiat et intuitif pour tous les collaborateurs de l'entreprise.

Un mode interactif collaboratif en ligne qui permet aux experts, mais aussi à tous, de contribuer à la qualification, la validation, l'indexation, la modification, la mise à jour ou la normalisation du contenu.

4. Fonctionnalités

Libellex traduction – Il s'agit d'un environnement destiné à simplifier le processus de demande de traduction et à systématiser l'utilisation de Similis, pour de meilleurs résultats. Les collaborateurs de l'entreprise postent leurs requêtes de traduction par une fenêtre de leur navigateur en intranet, reçoivent et contrôlent le résultat par le même moyen. Une fiche d'information leur indique l'économie de coût et de délai réalisée grâce à la

mémoire de traduction de l'entreprise, que les traductions soient faites en interne ou en sous-traitance. L'administrateur linguistique supervise la mémoire. Les relations avec les traducteurs sous-traitants sont maintenues, les coûts et délais sont réduits jusqu'à 75 %.

Libellex recherche d'expressions – la mémoire de traduction de l'entreprise est accessible sur l'intranet. Au sein de ce corpus, tout collaborateur de l'entreprise peut instantanément récupérer « la bonne expression » ou « la bonne traduction », par exemple quand il rédige un courrier électronique dans une langue qu'il maîtrise imparfaitement.

Libellex dictionnaire terminologique de l'entreprise – Libellex permet d'extraire et d'affiner des glossaires bilingues à partir de la mémoire de traduction, l'entreprise dispose ainsi de dictionnaires spécialisés. Inversement, des glossaires existants peuvent être insérés dans la mémoire, pour une plus grande rigueur terminologique.

Libellex détection systématique de coquilles – Libellex compare un document original validé (typiquement MS Word ou Excel) avec son bon à tirer (PDF) après maquettage. Cette détection automatisée d'erreurs typographiques s'applique aux rapports annuels, rapports financiers, manuels techniques, notices légales ou encore aux plaquettes commerciales. Elle fonctionne également pour contrôler un site web. Toute différence entre les deux documents est détectée, y compris dans les tableaux. Cette fonctionnalité peut être réalisée par nos soins en prestation de service pour nos clients qui ne désirent pas s'équiper de Libellex. Ce service est facturé en fonction de la taille du document à vérifier; nous pouvons également y adjoindre une analyse et une correction orthographique et grammaticale approfondie.

Libellex aide à la traduction d'un texte court – Libellex analyse un paragraphe à traduire et propose des éléments de traduction à partir de ce qui a été stocké en mémoire. Ces éléments peuvent être des mots, des expressions ou des phrases entières qui auront été récupérés des traductions précédentes.

5. Les bénéfices client

Libellex est une plateforme d'installation rapide. Elle peut servir un groupe restreint d'utilisateurs, comme par exemple une direction de la communication ou une direction marketing.

Libellex apporte à l'ensemble des collaborateurs d'une entreprise les mêmes services que chacun attend de sa mémoire personnelle : exhaustivité, classement et accès immédiat. De par sa nature informatique, Libellex apporte l'avantage supplémentaire de la mutualisation des connaissances issues de tous les documents de l'entreprise, et ce par-delà la barrière des langues, ce qui représente un intérêt majeur pour une société possédant des marchés, des partenariats ou des implantations à l'étranger.

Plus précisément, Libellex permet de :

Mutualiser, structurer, pérenniser et distribuer l'information textuelle dans l'entreprise, et ainsi dégeler l'accès au capital intellectuel que représente cette information, qu'elle soit en français ou dans les autres langues de l'entreprise.

Pour les domaines où c'est nécessaire, normaliser la terminologie (médecine numérique et non médecine digitale), la typographie (Lingua et Machina et non Lingua & Machina), la syntaxe d'un message (Ensemble tout devient possible) et sa traduction dans le contexte culturel de la langue cible (Yes We Can).

Dans les domaines pour lesquels la normalisation n'est pas primordiale, avoir un accès informatif à l'ensemble des documents enregistrés.

Tenir automatiquement à jour des lexiques et des glossaires validés dans tous les métiers et les gammes de produits.

Réduire les coûts internes de rédaction, de relecture et de validation des documents, que ce soit le bon-à-tirer du rapport annuel ou la traduction d'un communiqué de presse.

Augmenter la qualité de l'écrit et la fluidité de la lecture, quelle que soit la langue, quelles que soient les compétences linguistiques du rédacteur ou du lecteur.

Capitaliser l'historique de traduction de l'entreprise, simplifier la chaîne de traduction et systématiser les économies en coût et en délais sur les nouvelles commandes de traduction.

A propos des auteurs

François Brown de Colstoun

Président, Lingua et Machina

fbc@lingua-et-machina.com