Une ontologie du Cinéma pour évaluer les applications du Web Sémantique

Camille Pradel, Nathalie Hernandez, Mouna Kamel, Bernard Rothenburger

IRIT, Université de Toulouse le Mirail, Département de Mathématiques-Informatique, 5 allées Antonio Machado, F-31058 Toulouse Cedex 9
{pradel,hernandez,kamel,rothenbu}
@irit.fr

Résumé : Ce papier fait tout d'abord le constat du manque de visibilité et d'accessibilité des ressources exploitables pour l'évaluation objective des applications du Web Sémantique, en particulier en langue française. Après un état de l'art énumérant les principales techniques permettant l'évaluation d'ontologies, nous proposons nos propres ressources constituées d'une ontologie sur le thème du cinéma, d'un corpus de textes et d'un corpus de requêtes utilisateurs en rapport à ce même thème. L'objectif de ce papier est d'encourager la communauté à s'approprier ces ressources, à les utiliser pour l'évaluation d'applications et à aider à leur développement.

Mots-clés : Ontologie, évaluation, Web Sémantique

1 Introduction

Ce papier fait suite à un constat que nous avons fait lors de la conférence IC 2011. Sur les 22 papiers acceptés, 11 portent sur des aspects en lien avec le web sémantique. Sept d'entre-eux traitent de la construction, de l'enrichissement ou du peuplement d'ontologies (extraction de termes, concepts, relations ou instances), deux sur les moteurs de recherche sémantique, un sur le stockage de données RDF et un sur l'alignement de données sur le Web de données liées. En analysant les évaluations mises en place dans ces travaux, nous avons constaté que, dans tous les cas sauf un, les jeux de données utilisés pour évaluer l'approche sont constitués par les auteurs ou par des experts impliqués dans le même projet que les auteurs. Par ailleurs, un très petit nombre de ces ressources sont actuellement

accessibles. Ceci nous paraît dommage, car constituer un jeu de données est une lourde tâche qui pourrait être mutualisée par la communauté. De plus, il est important de pouvoir tester une approche indépendamment du contexte dans lequel elle a été conçue et ainsi de pouvoir la comparer aux autres approches proposées dans le domaine.

Notre proposition est de développer manuellement une ontologie de domaine de sens commun, ayant (i) un niveau de complexité représentatif des possibilités d'expression des langages de description (en termes de représentation des connaissances et de raisonnement), (ii) ayant la capacité de pouvoir être régulièrement enrichie et peuplée, et (iii) pouvant être évaluée dans le cadre de différentes applications. Nous avons choisi le domaine du cinéma qui vérifie tous ces critères et qui, de plus, a l'avantage de proposer différents corpus facilement accessibles via le web. Nous souhaitons utiliser cette ontologie dans des applications telles que la construction, l'enrichissement ou le peuplement d'ontologies et pour la traduction de requêtes exprimées en langage naturel vers le langage SPARQL. Nous souhaitons également que cette ontologie puisse être utilisé dans la cadre de travaux portant sur l'alignement d'ontologies. Notre ontologie pourrait par exemple être alignée avec des ontologies existant dans le domaine (http ://www.movieontology.org/documentation/ ou bien la partie de l'ontologie DBPedia dédiée au cinéma). Dans ce papier, notre objectif est de présenter l'ontologie à la communauté pour qu'elle puisse également être réutilisée dans d'autres perspectives. Notre ontologie est disponible à l'url : `http://ontologies.alwaysdata.net/cinema`.

2 État de l'art des techniques d'évaluation d'ontologies

L'usage croissant d'ontologies formelles dans les applications informatiques pose le problème de l'estimation de la qualité de ces ontologies et de leur adéquation au problème qu'elles contribuent à résoudre. Cette problématique couramment dénommée *évaluation d'ontologies* recouvre des aspects multiples. Nous décrivons ci-dessous les propositions les plus connues dans ce domaine. Nous situons ces propositions sur une échelle allant des plus quantitatives vers les plus qualitatives. Nous appelons *quantitatives* des évaluations qui tendent à produire une valeur numérique, *qualitatives* les propositions qui tendent à produire un ensemble de propriétés qui reflètent le plus ou moins grand respect de critères de qualité, et *propositions d'évaluation de cohérence logique ou conceptuelle* celles dont le but est d'identifier des défauts dans les ontologies. Parmi les proposi-

tions mises en œuvres, on peut distinguer différentes approches dont nous énonçons ci-après les principes généraux.

Un certain nombre de propositions quantitatives pour l'évaluation d'ontologies sont basées sur des critères structurels. Par exemple, (Ning & Shihan, 2006) proposent de prendre en compte six critères : la quantité de concepts pour estimer si la « granularité » de l'ontologie est adaptée, la densité des relations par rapport aux concepts pour estimer si les connexions entre concepts reflètent la réalité modélisée, l'équilibre des hiérarchies, la connectivité des sous-graphes de concepts, la quantité de concepts-clés. Dans un ordre d'idée différent, dans (Alani & Brewster, 2006), les auteurs proposent un moyen de classement de pertinence des ontologies trouvées (ranking) lorsqu'une requête est soumise à une collection d'ontologies. Il s'agit donc d'estimer structurellement la manière dont chacune des ontologies prend en compte les concepts contenus dans la requête. Quatre critères sont pris en compte : la couverture de l'ontologie par les concepts de la requête, le niveau de détail des descriptions autour de ces concepts, la proximité entre ces concepts, la centralité ou le caractère pivot de ces concepts. On peut aussi trouver dans (Gangemi *et al.*, 2005) une liste d'une trentaine de mesures caractérisant la structure d'une ontologie.

L'approche consistant à comparer une ontologie avec une ontologie de référence permet de valider la méthode de construction d'ontologies. Ce n'est qu'indirectement qu'elle permet d'évaluer des ontologies : lorsque l'on a démontré qu'une méthode de construction d'ontologies donne de bons résultats pour un certain type d'ontologies, on peut considérer que les ontologies de ce type créées par cette méthode seront de qualité. La comparaison consiste à évaluer la similarité entre les deux ontologies en prenant en compte le niveau lexical et le niveau de structure conceptuelle (Maedche & Staab, 2002), ou le niveau de partitionnement (clustering) calculé par rapport à un partitionnement de référence : Rand Index (Brank *et al.*, 2006).

Concernant les évaluations par rapport à un corpus du domaine, il s'agit ici d'évaluer la qualité d'une ontologie en mesurant comment le contenu de cette ontologie correspond aux connaissances issues d'un corpus représentatif du domaine modélisé. Dans cette approche les travaux de C. Brewster (Brewster *et al.*, 2004) font autorité. Le but est de trouver l'ontologie la plus adéquate parmi un ensemble d'ontologies possibles. On peut avoir une première estimation en repérant le nombre de concepts qui sont atteints par des termes issus du corpus de référence. Pour améliorer cette première estimation les auteurs proposent de produire des regroupements

de textes du corpus annotés par les concepts, puis de prendre comme indice de qualité pour l'ontologie, le fait que les termes regroupés sont aussi proches dans l'ontologie. Un modèle probabiliste rend compte de cette approche.

Dans les approches basées sur le comportement d'applications il ne s'agit pas de mesurer la qualité d'une ontologie en soi mais de la valider par rapport à la tâche ou l'application qu'elle est sensée supporter. L'ontologie est d'autant mieux adaptée que l'exécution de la tâche pour laquelle elle a été construite produit des résultats qui se rapprochent de résultats de référence. Il existe une approche « duale » : étant donnés une ontologie et un ensemble de données, on peut s'en servir pour comparer différents moyens d'assurer une certaine tâche. (Maynard *et al.*, 2006) évalue ainsi des applications d'extraction d'information en vue du peuplement d'ontologies en fixant une ontologie et un corpus de textes annotés.

On peut aussi identifier des méthodes de choix basés sur des critères généraux. La méthode OntoMetric (Lozano-Tello & Gómez-Pérez, 2004) permet d'effectuer le choix d'une ontologie parmi un ensemble possible en caractérisant les objectifs de l'application visée et les ontologies disponibles. La méthode est basée sur une taxinomie de 160 caractéristiques prenant en compte plusieurs aspects : le contenu des ontologies, le langage utilisé pour ces ontologies, la méthodologie utilisée pour les développer, leurs environnements de développement et leurs coûts d'utilisation. Une fois les caractéristiques identifiées la méthode utilise l'AHP (Analytic Hierarchy Process), une méthode d'aide à la décision multicritère afin de classer les ontologies par ordre de pertinence. Plus récemment, (Poveda-Villalon & Suárez-Figueroa, 2012) ont proposé le système OOPS qui permet de vérifier en ligne la qualité de l'ontologie à partir d'un catalogue de 30 erreurs courantes faites lors de la conception de l'ontologie.

Une autre approche pour la mesure d'évaluation consiste à estimer la cohérence logique. Etant donnée une ontologie formelle on ne veut pas pouvoir en déduire logiquement une assertion qui soit en contradiction avec une autre assertion de l'ontologie. De nombreux outils (FACT, Pellet, Racer...) ont été développés pour assurer cette « validation » logique d'une ontologie. Ils peuvent être mis en œuvre sur des outils courant comme Protégé.

Un autre type d'approche consiste à évaluer le comportement des utilisateurs des ontologies. Proches d'idées du Web 2.0, les ontologies sont ici évaluées à partir du comportement « social » des utilisateurs. Dans un premier cas, la qualité des ontologies est évaluée à partir de la confiance (trust)

et du classement que les utilisateurs des ontologies ont produits lorsqu'ils les ont utilisées (Kalfoglou & Hu, 2006). On trouve là des idées proches de l'algorithme PageRank pour le classement de pages Web. Dans un second cas,(Lewen *et al.*, 2006) se servent d'annotations ajoutées par les utilisateurs pour caractériser les ontologies. On a là une démarche identique à celle qui consiste à exploiter les « tags » d'utilisateurs tels que ceux qui sont apparus dans des outils du Web 2.0 (Delicious ou Flickr).

L'évaluation de la qualité d'une ontologie par un expert du domaine est une démarche classique. Elle n'est pas toujours aisée à mettre en œuvre. Elle peut être facilitée par des outillages de visualisation adaptés. Les travaux présentés dans (Hernandez, 2005) présentent une interface de visualisation (ELEGIE) qui va dans ce sens. La méthodologie OntoClean présentée dans (Welty & Guarino, 2001) est basée sur l'identification de caractéristiques (ou meta-propriétés) fondamentales des concepts pour identifier des défauts dans une ontologie. Ces caractéristiques, au nombre de quatre sont issues de la philosophie : rigidité, unicité, identité et dépendance. A partir de ces quatre méta-propriétés OntoClean définit des règles permettant de nettoyer une ontologie quelconque. Par exemple une classe qui est 'unique' ne peut être sous-classe d'une classe anti-unique ou une classe rigide et une classe anti-rigide sont forcément disjointes.

3 Description de l'ontologie

Cette section décrit les principes mis en œuvre pour la construction de l'ontologie, ainsi que les trois composantes principales de l'ontologie, à savoir la représentation des rôles joués par les acteurs d'un film, le doublage d'un film et les événements associés aux films.

3.1 Exigences à satisfaire

Nous avons défini plusieurs exigences auxquelles doit répondre notre ontologie.

Tout d'abord, nous souhaitons développer une ontologie de domaine de sens commun dans le but (i) quelle soit appréhendable par les membres de la communauté sans nécessité la mobilisation d'experts et (ii) qu'il existe des sources de données sur le domaine de façon à ce que ces sources puissent servir de jeux de données en complément de l'ontologie

Nous souhaitons que l'ontologie soit en français dans la perspective de permettre l'évaluation d'applications visant un public francophone et de mettre en avant les travaux portant sur le français.

La structure de l'ontologie doit également faciliter son peuplement et son évolution en utilisant les possibilités d'expression des langages de description (en termes de représentation des connaissances et de raisonnement). Une autre des qualités recherchées est que la structure de l'ontologie soit construite à la main et issue d'un consensus relevant d'une démarche ontologique.

3.2 Principes mis en œuvres

Pour cette première version de l'ontologie du cinéma, nous avons choisi de nous limiter aux aspects les plus factuels du domaine à savoir les types de films, leurs contributeurs, les événements, les genres, les courants et les techniques cinématographiques. En ce qui concerne le contenu des films, nous ne représentons que les rôles interprétés et nous ne détaillons pas l'intrigue. Pour l'instant nous n'avons pas représenté les connaissances plus subjectives telles que les opinions que peut avoir une personne sur une oeuvre.

L'ontologie a été construite par 3 personnes ayant de l'expérience dans la construction d'ontologies.

Durant la conception de cette ontologie, nous nous sommes efforcés d'appliquer les techniques couramment recommandées dans la communauté. Nous avons notamment implémenté des patrons de conception d'ontologies. Le plus important de ces patrons est le patron de conception *normalization* introduit dans (Rector, 2003) ; cette méthode permet de construire des ontologies modulaires et réutilisables en définissant les classes par les propriétés que doivent vérifier leurs instances ; le concepteur de l'ontologie n'a ainsi pas à se soucier des propriétés de subsomption : la taxonomie est inférée automatiquement par le moteur d'inférences (une classe A subsume une classe B si les instances de B présentent au moins toutes les propriétés des instances de A).

Parmi les autres bonnes pratiques utilisées, citons la définition pour chaque propriété d'objet d'une propriété inverse facilitant ainsi les manipulations et l'alignement de notre ontologie avec des ontologies de référence (comme FOAF présenté dans (Brickley, 2007)). Afin de faciliter l'intégration de notre ontologie au Web Sémantique et d'encourager son utilisation dans les applications développées par la communauté, nous l'avons alignée sur des ontologies reconnues déjà largement utilisées sur le web de données. Ainsi, la classe `Artiste` définie dans l'ontologie du cinéma est une sous-classe de la classe `Person` définie dans l'ontologie FOAF, et la classe `Competition` est une sous-classe de la classe `Event` définie

dans l'ontologie `EVENT` [1].

Enfin, l'ontologie ayant été décrite en OWL2, nous avons tiré parti des possibilités de ce langage en termes d'expressivité (OWL 2 new features, 2009). Nous avons notamment défini des chaînes de propriétés permettant d'inférer de nouvelles relations sans avoir recours à un langage dédié à l'expression de règles. Par exemple, la relation `acteurDoublePar` est inférée lorsqu'une personne `incarne` un role et que ce role `estDoubléPar` une personne.

3.3 Représentation des rôles

Détaillons maintenant certains aspects de notre ontologie. Notre ontologie permet de représenter les personnes impliquées dans la réalisation d'un film ainsi que la nature de cette implication. Par exemple, pour ce qui concerne un acteur, nous avons défini la relation ternaire qui lie l'*acteur* au *rôle* qu'il interprète dans un *film*. Cette relation est représentée par deux relations binaires illustrées dans la figure 1. Dans cet exemple, l'instance « MO Brian » correspond au rôle du père de famille interprété par « Brad Pitt » dans le film « The tree of life ».

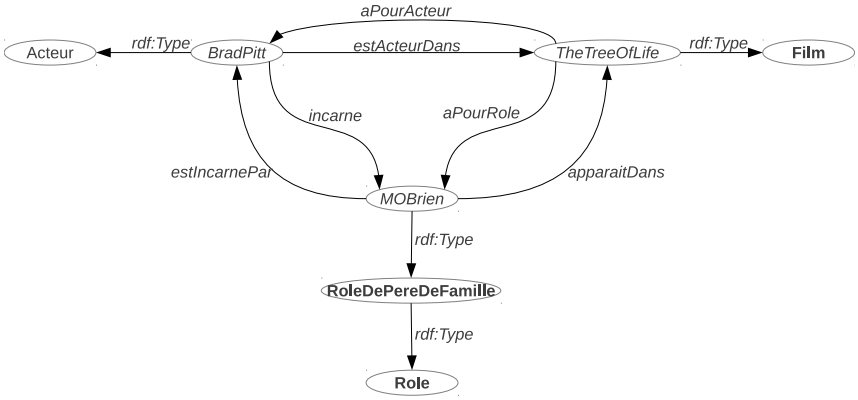

FIGURE 1 – Film/Acteur/Rôle

1. http ://motools.sourceforge.net/event/event.html

3.4 Représentation des versions

Nous représentons également les différentes versions d'un film à savoir la version originale et les versions distribuées dans les différentes zones du monde. Une version localisée a les mêmes propriétés que la version originale (acteurs, réalisateur, courant cinématographique, lieu de tournage, etc.), mais possède également des propriétés qui lui sont propres comme la zone de distribution, la langue, les personnes ayant doublé les acteurs, etc. Dans l'exemple de la figure 2, « LArbreDeLaVie » est une version du Film « TheTreeOfLife » distribuée au Quebec. Pour cette version du film, « AlainZouvi » est le doubleur de « Brad Pitt » pour le role « MOBrien ».

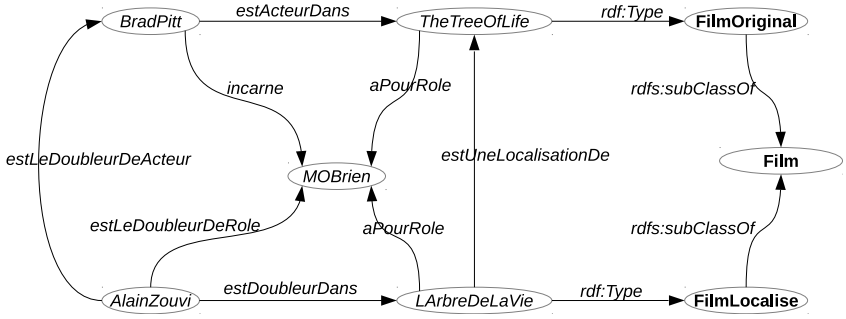

FIGURE 2 – Film Localisé/Doubleur

3.5 Représentation des événements associés aux films

Des événements cinématographiques tels que les festivals, les cérémonies, les votes par voie de presse, etc. sont organisés régulièrement (généralement la périodicité est annuelle) dans le but de récompenser les meilleurs films de l'année et les meilleurs artistes participant à ces films. Les concours se déroulent en deux étapes. La première étape consiste à sélectionner les films et artistes qui seront présentés au concours : il s'agit alors des films et artistes nominés. La seconde étape a pour but de désigner les vainqueurs dans chaque catégorie : meilleur film, meilleur scénario, meilleur acteur, meilleure actrice, meilleure bande originale, meilleure prise de son, etc. Les récompenses sont délivrées lors d'un des événements cinématographiques cités ci-dessus, par un jury, chaque jury étant présidé par une personne célèbre dans le monde du cinéma (réalisateur, acteur,

producteur, etc.). Les récompenses peuvent être spécifiques ou non à un concours, et concernent aussi bien les différents artistes ayant contribué à la réalisation du film (acteur, metteur en scène, réalisateur, scénariste, musicien, producteur, etc.), que l'œuvre dans sa globalité (du film à l'œuvre totale d'un acteur, d'un réalisateur, etc.).

Dans l'exemple de la figure 3, la personne « JeanDujardin » a obtenu le prix « PrixInterpretationMasculineFestivalDeCannes2011 » décerné lors de la compétition « FestivalDeCannes2011 », et le film « TreeOfLife » a obtenu le prix « PalmeDor2011 » lors de cette même compétition. La compétition « FestivalDeCannes » a eu pour jury « JuryFestivalDeCannes2011 » présidé par la personne « RobertDeNiro », qui est donc aussi membre de ce même jury. La définition de « chaînes de propriétés » permettent d'inférer que le film « TreeOfLife » et la personne « JeanDujardin » ont été récompensés lors de la compétition « FestivalDeCannes2011 », et que la personne « JeanDujardin » a été récompensée pour sa performance dans le film « TheArtist » (cette dernière propriété n'apparaît pas dans la figure 3 pour des raisons de lisibilité).

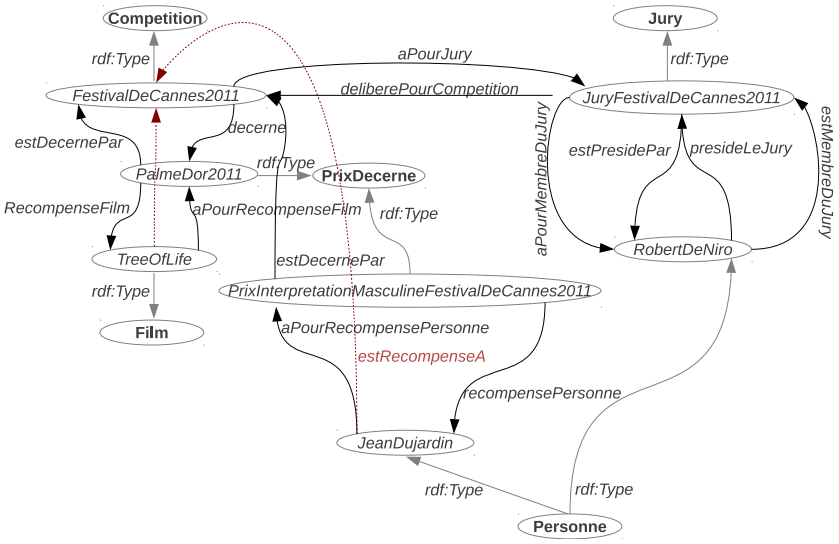

FIGURE 3 – Film/Competition

4 Proposition de jeux de données

Afin de rendre notre ontologie utilisable dans le cadre de l'évaluation d'applications portant sur le Web Sémantique, nous proposons de la compléter par plusieurs jeux de données. Les différents jeux de données sont disponibles à l'url :
`http://ontologies.alwaysdata.net/cinema/description/`.

4.1 Données de peuplement

Nous avons manuellement peuplé l'ontologie avec les faits issus du festival de Cannes 2011, de la cérémonie des César 2012 et du festival du court métrage de Clermont-Ferrand 2012. Pour ces trois compétitions, nous avons représenté les films nominés et primés ainsi que toutes les relations qui les impliquent directement dans l'ontologie.

4.2 Documents

Plusieurs tâches nécessitent l'utilisation de documents. Le domaine du cinéma étant largement décrit, il existe une grande quantité de documents relatifs à ce sujet. Par exemple, le portail du cinéma de wikipedia [2] offre une vue d'ensemble cohérente et structurée du domaine. D'autres ressources telles que cinefiches [3] ou nord-cinema [4] peuvent également être utilisées.

4.3 Requêtes

Nous voulons également mettre à disposition des requêtes pour permettre l'évaluation de système de questions-réponses. Nous avions recueilli 160 requêtes distinctes en lien avec le domaine modélisé auprès de 24 personnes, chaque requête étant composée d'un ensemble de mots-clés et d'une phrase en langage naturel exprimant le besoin en information.

5 Conclusion

Nous avons développé une première version de l'ontologie du cinéma. Notre ontologie répond aux exigences exposées en 3.1. De plus, nous l'avons

2. http ://fr.wikipedia.org/wiki/Portail :Cin%C3%A9ma
3. http ://www.cinefiches.com/
4. http ://www.nord-cinema.com/fiches/

évaluée et corrigée à l'aide de la méthode OOPS ! présentée en 2. Enfin, elle est accessible en ligne, accompagnée de différents jeux de données pouvant servir de base à l'évaluation d'applications du Web Sémantique français.

Références

ALANI H. & BREWSTER C. (2006). Metrics for ranking ontologies.

BRANK J., MLADENIĆ D. & GROBELNIK M. (2006). Gold standard based ontology evaluation using instance assignment.

BREWSTER C., ALANI H., DASMAHAPATRA S. & WILKS Y. (2004). Data driven ontology evaluation.

BRICKLEY D. (2007). Foaf vocabulary specification 0.9. *http ://xmlns. com/foaf/spec/20070524. html*.

GANGEMI A., CATENACCI C., CIARAMITA M., LEHMANN J., GIL R., BOLICI F. & STRIGNANO O. (2005). *Ontology evaluation and validation*. Rapport interne, Citeseer.

HERNANDEZ N. (2005). Ontologies de domaine pour la modélisation du contexte en recherche d'information.

KALFOGLOU Y. & HU B. (2006). Issues with evaluating and using publicly available ontologies.

LEWEN H., SUPEKAR K., NOY N. & MUSEN M. (2006). Topic-specific trust and open rating systems : An approach for ontology evaluation. In *Proceedings of WWW'06 4th International EON Workshop Evaluating Ontologies for the Web*.

LOZANO-TELLO A. & GÓMEZ-PÉREZ A. (2004). Ontometric : A method to choose the appropriate ontology. *Journal of Database Management*, **2**(15), 1–18.

MAEDCHE A. & STAAB S. (2002). Measuring similarity between ontologies. *Knowledge Engineering and Knowledge Management : Ontologies and the Semantic Web*, p. 15–21.

MAYNARD D., PETERS W. & LI Y. (2006). Metrics for evaluation of ontology-based information extraction.

NING H. & SHIHAN D. (2006). Structure-based ontology evaluation. In *Proceedings of the IEEE International Conference on e-Business Engineering*, p. 132–137 : IEEE Computer Society.

OWL 2 NEW FEATURES (2009). Owl 2 web ontology language new features and rationale. Web site. http ://www.w3.org/TR/2009/REC-owl2-new-features-20091027/.

POVEDA-VILLALON M. & SUÁREZ-FIGUEROA M. (2012). Oops !–ontology pitfalls scanner !

RECTOR A. (2003). Modularisation of domain ontologies implemented in description logics and related formalisms including owl. In *Proceedings of the 2nd*

international conference on Knowledge capture, p. 121–128 : ACM.

WELTY C. & GUARINO N. (2001). Supporting ontological analysis of taxonomic relationships. *Data and Knowledge Engineering*, **39**(1), 51–74.