

Tâche, domaine et application : influences sur le processus de modélisation de connaissances

Axel Reymonet^{1,2}, Nathalie Aussenac-Gilles¹, Jérôme Thomas²

¹ Équipe Conception de Systèmes Coopératifs, Institut de Recherche en Informatique de Toulouse, Université Paul Sabatier, 118 Route de Narbonne, F-31062 TOULOUSE CEDEX 9
reymonet@irit.fr, aussenac@irit.fr

² Division Technologie ACTIA,
25 Chemin de Pouvoirville, BP 4215, 31432 TOULOUSE CEDEX 4
axel.reymonet@actia.fr, jerome.thomas@actia.fr

Résumé Un nombre croissant d'outils de gestion de documents et de connaissances a désormais recours à des ressources terminologiques et/ou ontologiques (RTO) pour répondre à leurs besoins applicatifs. Nous montrons que le processus de modélisation de telles ressources passe par la prise en compte de la nature du domaine, de la tâche et de l'application visés. Pour cela, nous nous appuyons sur une étude de cas de construction de RTO à partir de textes dans le domaine du diagnostic automobile.

Mots-clés : Construction et maintenance de modèles de connaissances, Ontologie, Recherche d'information

d'ontologies à l'application dans laquelle ces ressources sont utilisées (2.1). Nous nous intéressons ensuite aux diverses façons d'envisager la notion de rôle, point de jonction entre modèle du domaine et modèle de raisonnement (2.2). Dans une deuxième partie, nous situons notre étude dans son contexte applicatif, à savoir la gestion par ontologie d'un module de recherche d'information et la mise en collaboration de plusieurs modes de raisonnement pour une aide au diagnostic automobile. Enfin, dans une troisième partie, nous analysons, via notre cas d'étude, l'influence que peuvent avoir sur les étapes de construction de l'ontologie la tâche et le domaine à modéliser, ainsi que l'application de destination.

1 Introduction

Depuis la démocratisation de l'outil informatique et l'apparition d'Internet, la proportion de documents disponibles sous forme électronique a considérablement augmenté et la pratique terminologique a évolué en conséquence. De nombreuses applications utilisent des ressources terminologiques et/ou ontologiques pour traiter des archives textuelles sous ce nouveau format : aide à la rédaction, mémoire d'entreprise, indexation automatique, extraction d'informations ... Dans un souci d'efficacité, plusieurs méthodes ont été développées afin de construire des ressources génériques et réutilisables. Toutefois, comme le montre le groupe TIA dans ses travaux, cette volonté de réutilisation des ressources terminologiques nuit à leur bonne utilisation pour un usage particulier.

Notre article se situe dans la ligne de pensée du groupe TIA et il se veut plus particulièrement un prolongement de (2; 7). Il souligne la nécessité d'une analyse préliminaire de l'application pour laquelle une ontologie est élaborée afin d'assurer la bonne adéquation entre la ressource construite et les besoins applicatifs.

Dans une première partie, nous évaluons d'abord l'importance donnée par différentes méthodes de construction

2 Finalité des ontologies : questions méthodologiques

Les travaux méthodologiques sur les ontologies visent à organiser et systématiser leur construction. Nous faisons le point ici sur la place faite dans ces méthodes à la finalité de l'ontologie, c'est-à-dire à la manière dont elle va être utilisée dans une ou plusieurs applications cibles.

Plus finement, l'application cible suppose qu'une tâche soit réalisée par le couple utilisateur-application. Dans le cas où un modèle de cette tâche est connu, on retrouve la problématique classique en ingénierie des connaissances des liens entre modèle du domaine (ou ontologie) et modèle du raisonnement (et de la tâche). À l'articulation entre ces deux niveaux se trouve la notion de rôle. Nous rappelons quelques travaux qui ont traité cette notion du point de vue des ontologies afin de souligner la diversité des choix de modélisation possibles pour assurer l'adéquation entre ontologie et raisonnement.

2.1 Méthodes de construction d'ontologies : influence de l'application

Historiquement, la construction d'ontologies fait l'objet de recherches méthodologiques dans un double souci. Une première préoccupation est d'éviter de construire des ontologies ad hoc, non réutilisables, et qui soient de simples modèles de domaine. Une deuxième volonté vise à passer de démarches " artisanales ", essentiellement manuelles, dont la durée et le coût sont difficiles à estimer, à des approches plus systématiques, outillées et mieux maîtrisées.

Les recherches visant le premier objectif s'appuient sur les fondements théoriques et philosophiques de ce qu'est une ontologie. Ainsi, la méthode Archonte (3) propose des principes de structuration ontologique basés sur des critères différentiels, OntoClean (15) définit des méta-propriétés pour vérifier la définition et l'organisation hiérarchique de concepts, ou encore OntoSpec (17) permet d'appliquer ces méta-propriétés au fur et à mesure de la spécification de concepts. La deuxième orientation, plus pragmatique, a donné lieu à des propositions méthodologiques mettant l'accent sur la réutilisation comme Methontology (14), sur des guides pratiques (13) ou, depuis 10 ans environ, sur l'analyse systématique de textes à l'aide de logiciels de traitement automatique des langues, comme Terminae et les méthodes répertoriées dans (19). Dans un courant méthodologique comme dans l'autre, les méthodes proposées se veulent génériques, applicables dans tous les contextes et les domaines, quelle que soit l'application dans laquelle s'intègre l'ontologie construite.

La plupart des méthodes considèrent que l'ontologie est un modèle consensuel promis à plusieurs types d'applications et qui ne doit en rien être influencé par l'usage qui en sera fait. Au contraire, dans la méthode Archonte, B. Bachimont insiste sur le fait que les ontologies vraiment utiles et utilisables dans des applications ne peuvent être que des ontologies régionales, dans lesquelles le choix et l'organisation des concepts tiennent compte de leur utilisation dans l'application. Selon cette analyse, reprise dans les travaux du groupe TIA et dans Terminae, les ontologies n'ont pas de portée universelle mais ce sont des représentations de définitions consensuelles des concepts retenus au sein d'un domaine en vue d'un objectif précis (10).

Or, les analyses de (2) et (7) montrent que l'application ciblée détermine bien plus que le contenu de la ressource construite : elle engage des choix méthodologiques à tous les stades de la modélisation. L'impact de l'application cible sur le processus de modélisation a été examiné à travers les points suivants : profil de l'analyste, construction du corpus, choix et manière d'utiliser les outils de TAL, choix et utilisation des outils de modélisation. Il en ressort la nécessité de faire évoluer les propositions méthodologiques pour indiquer précisément quand et comment l'application ciblée oriente des choix. Pour cela, des études précises doivent être menées, pour différents types d'application. Ainsi, la plate-

forme et la méthode IndDoc proposées par (20) sont une adaptation particulière au cas où la ressource sert d'index à un document numérique.

Cet article se veut un prolongement des travaux présentés dans (2; 7). Nous souhaitons pousser plus loin les réflexions qui ont été menées précédemment à partir de différentes expériences de construction de ressources termino-ontologiques (RTO). Notre cas d'étude comporte une dimension particulière : l'ontologie doit favoriser la collaboration de plusieurs méthodes de raisonnement.

2.2 Ontologies et modèles de raisonnement : place des rôles

Dès les premiers travaux sur les ontologies en ingénierie des connaissances, la question de l'articulation entre raisonnement et domaine s'est posée en s'appuyant sur deux notions centrales : celle d'engagement ontologique d'une part, celle de rôle d'autre part. L'engagement ontologique renvoie soit à la définition formelle des concepts, une fois qu'a été fixée leur interprétation au niveau linguistique (3), soit aux contraintes et prérequis imposés sur la nature et la structuration des éléments du domaine par les méthodes de résolution (16).

La notion de rôle a été l'objet de débats car elle prend des sens différents (21). Dans un modèle CommonKADS (22), les rôles sont les paramètres des opérateurs d'une méthode de résolution de problème. Ils renvoient à la manière dont les concepts du domaine interviennent dans le raisonnement. Cette distinction permet de regrouper des entités du domaine tantôt en fonction du raisonnement (rôles) tantôt selon les classes conceptuelles du domaine (concepts), décrites en fonction des besoins de ce raisonnement. De ce fait, le rôle n'est pas considéré comme faisant partie du modèle du domaine (16). La matérialisation des liens entre rôles et concepts peut être fixe ou dynamique (21) (15), c'est-à-dire actualisée au cours de la résolution d'un problème.

On aurait pu croire que la notion d'ontologie allait aider à mieux situer la frontière entre domaine et raisonnement. Or la place des rôles dans les ontologies a été d'emblée très discutée (16). Suivant les méthodes, les rôles sont représentés comme des concepts ayant un statut particulier, ou bien en dehors de l'ontologie, comme dans CommonKADS. Dans une tradition plus ontologique, les rôles sont représentés dans l'ontologie et sont reliés à d'autres concepts qui " jouent " ce rôle. Par exemple, dans DOLCE, un rôle est exprimé à l'aide de deux types de concepts mis en relation : un concept de type " Agentive Physical Object " (ex : être humain) ou " Agentive Social Object " (ex : Enseignant, Étudiant) qui sont des endurants (des concepts stables dans le temps) et un autre endurant non agentif. Dans OntoSpec, la propriété qui relie ces deux types de concepts, " joue le rôle ", est une propriété anti-rigide (non stable) et dépendante pour traduire le fait qu'un concept ne joue un rôle que

dans un certain contexte.

Or même le point de vue des concepteurs de KADS a évolué. Aujourd'hui, J. Breuker place des rôles dans l'ontologie LRI-Core. Pour lui, les rôles sont des classes qui rendent compte de contraintes sur certains concepts à un instant donné (9). S'intéresser aux rôles suppose une certaine dynamique des connaissances d'un domaine, puisque ces contraintes peuvent ou non s'appliquer sur un concept donné. Cette dynamique peut être temporelle ou liée à des processus ou encore à des raisonnements.

3 Contexte de l'étude

Notre étude se situe dans le cadre du projet MODE qui associe la section Recherche de la société ACTIA et les laboratoires LAAS et IRIT. Ce projet a pour ambition d'amener plusieurs modes de raisonnement à collaborer dans un même outil logiciel afin d'aider un garagiste à repérer la (les) fonction(s) défaillante(s) d'un véhicule en panne, trouver la cause du problème et le résoudre.

Nous nous plaçons dans le paradigme de (19), selon lequel une ontologie est définie comme une hiérarchie de concepts reliés par des relations à laquelle vient s'ajouter un ensemble lexical non ordonné, les termes.

3.1 Le projet MODE

Parmi les différents modules du projet MODE, l'application qui nous intéresse en premier lieu, le module RXP, consiste à rechercher dans une base d'expériences des fiches de réparation pertinentes pour une panne sur un véhicule donné : à partir de la description en langue naturelle des symptômes, l'outil doit retrouver un ensemble de fiches traitant d'un problème similaire sur le même type de véhicule. Un moteur de recherche basé sur une indexation classique par mots-clés aurait pu suffire mais, en l'absence d'une modélisation des connaissances du diagnostic automobile, il aurait été impossible de raisonner sur les objets du domaine (e.g. la présence d'un symptôme peut entraîner celle d'un autre). De plus, les travaux de (5) sur l'indexation conceptuelle avec une ressource générique (WordNet) concluent sur la nécessité de combiner celle-ci avec une indexation classique pour obtenir un gain de précision significatif. Le projet MODE fournit un cadre pour étudier l'indexation conceptuelle dans le cas d'une ressource centrée sur une tâche et un domaine. Ces différentes motivations nous ont donc poussé à construire une RTO pour modéliser le domaine du diagnostic automobile. Cette ressource, du fait de sa spécificité, sera sans doute difficilement réutilisable. Toutefois, nous considérons ce critère de réutilisabilité comme moins essentiel que la bonne adéquation avec l'application (7).

En plus de son utilisation par RXP, la RTO doit faciliter la collaboration avec et entre les autres méthodes de

raisonnement mises en oeuvre au sein de MODE, à savoir la reconnaissance de formes (RdF), le diagnostic embarqué (DDP) et le raisonnement à base de modèles (MBR). En effet, celles-ci manipulent en commun certains objets du domaine. Par exemple, les modules DDP et RdF ont pour but de fournir au module MBR une liste de fonctions suspectes (vis-à-vis de la panne) afin que celui-ci puisse proposer une séquence optimale de tests à effectuer par le garagiste dans le cadre du diagnostic. Un formalisme commun (en l'occurrence la RTO) est donc souhaitable afin que les différents modes de raisonnement puissent communiquer (fig. 1). Cette utilisation des ressources ontologiques entraîne des contraintes supplémentaires sur leur structure.

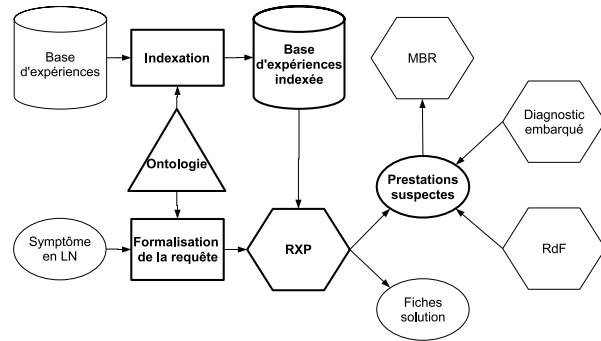


FIG. 1 – La place de l'ontologie dans MODE

3.2 Description de la base d'expériences

Les documents appartenant à la base d'expériences ont la particularité de provenir de deux sources différentes, à savoir un constructeur automobile français d'une part, et un réseau de services pour réparateurs indépendants d'autre part. Ces fiches sont dans les deux cas des synthèses d'experts en diagnostic automobile à partir d'analyses de cas réels rencontrés en garage. Elles sont formatées selon une structure fort similaire avec 4 champs principaux (fig. 2) :

- le type de véhicule concerné (" concerne ")
- le symptôme client, indicatif de panne (" constatation ")
- le diagnostic du problème (" diagnostic ")
- la réparation associée (" remède après-vente ")

Nous disposons actuellement d'environ 800 fiches constructeur (700 triplets (*symptôme, diagnostic, remède*) distincts¹) et 4700 fiches multi-marques (1400 triplets distincts²). Cette base est amenée à évoluer régulièrement afin de prendre en compte et de diffuser les nouveaux problèmes que certains garagistes pourraient rencontrer. D'un point de vue linguistique, les fiches se distinguent notamment par la

¹Certains triplets sont dupliqués lorsque le problème survient sur plusieurs types de véhicules, d'où la différence.

²Le nombre de duplications est dans ce cas très important car la base fait souvent référence à des problèmes génériques affectant plusieurs modèles.

<p>CONCERNE Citronault Pipo</p> <p>CONSTATATION En roulant le moteur manque de puissance en accélération et en vitesse de pointe.</p> <p>DIAGNOSTIC Pas de défaut mémorisé dans le calculateur d'injection. Après avoir contrôlé que le débit d'air dans le débitmètre est trop faible, vérifier la connexion du débitmètre.</p> <p>REMEDE APRES-VENTE Démontage et remontage du connecteur du débitmètre. Procédure habituelle.</p>
--

FIG. 2 – Exemple de fiche de réparation

concision de la rédaction avec une proportion importante de phrases nominales.

4 Analyse en situation des influences

Nous comptons montrer, par une étude détaillée de ce projet, comment certains facteurs influencent le processus de construction de RTO et son contenu. Dans un premier temps, nous définissons ces facteurs, à savoir les notions de tâche, domaine et d'application, en particulier dans le contexte de notre analyse. À travers des exemples issus de notre expérience dans le domaine du diagnostic automobile, nous nous intéressons ensuite aux conséquences de ces paramètres sur le processus de modélisation d'une RTO. Pour notre étude, nous avons suivi une méthode de construction de RTO proche de Terminae (1), partant de l'analyse de textes, mais en accordant une place plus importante aux connaissances détenues par les experts du domaine.

4.1 Définition des paramètres d'influence

La notion d'**application** renvoie à la prise en compte des besoins de l'utilisateur et des objectifs visés par la RTO au sein du logiciel qui sera développé. Ce paramètre primordial influence de façon implicite certains choix faits en fonction de la tâche et du domaine. Dans notre cas d'étude, l'application correspond à la recherche d'information dans une base d'expériences d'une part, à la mise en collaboration de plusieurs méthodes de raisonnement d'autre part.

Le concept de **tâche** est à prendre au sens classique utilisé dans CommonKADS (22). La tâche définit les buts que doit réaliser le système ainsi que la ou les méthodes de résolution mises en oeuvre pour les atteindre. Pour le projet MODE, la tâche correspond au diagnostic automobile.

Ce modèle n'a pas besoin d'être développé pour décrire le raisonnement au delà de ce qui est explicité par la structure des fiches de la base d'expériences. En effet, le module RXP n'effectue pas une résolution de problème mais une recherche d'information dans ces fiches. On peut en voir un modèle simplifié sur la figure 3 : à partir d'un symptôme (révélé par un indicateur ou un comportement anormal signalé par le garagiste et recherché dans le champ " constatation " des fiches), il s'agit d'identifier un élément défaillant (mentionné dans le champ " diagnostic "). L'élément défaillant

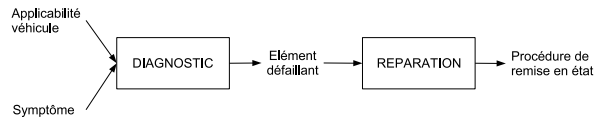


FIG. 3 – Modèle de la tâche du garagiste

est, dans le meilleur des cas, un composant (si le diagnostic a complètement abouti), au pire une prestation³. La figure 1 montre que notre application ne nécessite pas de guider le processus de réparation.

Enfin, la notion de **domaine** correspond à la sphère de connaissances que l'on cherche à modéliser. Dans le cadre de notre construction, il s'agit de certains savoirs liés à l'automobile et détenus par un garagiste. Le domaine est étroitement lié à la tâche : le modèle doit couvrir uniquement les connaissances que le garagiste utilise pour réaliser diagnostic et réparation.

4.2 Choix préliminaires

4.2.1 Méthode de construction

Comme le souligne (7), il serait idéal que la personne en charge de la modélisation détienne un certain nombre de compétences sur le domaine, en ingénierie des connaissances, en linguistique et en informatique. Pratiquement, c'est en fonction de ses propres compétences, des ressources disponibles et des besoins applicatifs que l'analyste désigné (qui pourra éventuellement faire appel à des spécialistes complémentaires) adapte sa méthode de travail pour construire une RTO.

Dans le projet MODE, la méthode utilisée se rapproche fortement de celle employée pour le projet VERRE (1) : comme les connaissances de l'analyste dans le domaine du diagnostic automobile étaient faibles, il a fallu compenser en utilisant au mieux les connaissances présentes dans une sélection de textes, les connaissances de spécialistes du diagnostic automobile, et enfin les connaissances élémentaires que nous avons en tant qu'utilisateurs de véhicules

³Une prestation véhicule est un service rendu à l'utilisateur (comme la climatisation, l'essayage de la lunette arrière ...) par le biais d'un système de composants.

pouvant tomber en panne. L'utilisation combinée d'un outil d'analyse syntaxique (Syntex (8)) et d'un logiciel d'analyse distributionnelle (Upery (6)) nous a permis d'étudier les formes sous lesquelles apparaissaient les concepts. Mais l'organisation hiérarchique des classes conceptuelles, implicite dans les textes, a été fournie par un expert.

4.2.2 Constitution du corpus

Pour la définition du corpus, nous nous situons dans l'acception de (12), selon laquelle un corpus est " *une collection de textes [...] constituée à partir de critères linguistiques ou extra-linguistiques pour évaluer une hypothèse linguistique ou pour répondre à un besoin applicatif* ".

En vue de la construction d'une RTO, le critère primordial pour constituer un corpus est la prise en compte de l'application pour laquelle est bâtie la RTO. Dans notre cas, comme un des objectifs consiste à rechercher des informations dans une base d'expériences, nous avons sélectionné l'ensemble des fiches de réparation comme corpus.

L'influence de la nature de l'application se traduit par la prise en compte de l'utilisateur. En effet, les fiches retenues comme corpus de départ sont rédigées par des experts du diagnostic automobile mais devront être comparées à des requêtes posées par des garagistes. Il est donc indispensable de vérifier la bonne adéquation entre les concepts et la terminologie utilisés par chacun des deux groupes. Dans notre cas, une étude d'ergonomie menée auprès de concessionnaires a permis de montrer que les deux groupes manipulaient des termes et des concepts très proches. Il n'a donc pas été nécessaire de modifier le corpus pour pallier un manque de ce type.

L'indexation conceptuelle d'une base de fiches évolutive nous a confronté à un autre problème intéressant : il nous fallait savoir si la RTO resterait adaptée aux nouveaux textes à indexer. Dans un domaine technologique comme celui de l'automobile, de nouveaux documents viennent régulièrement enrichir la base de fiches initiale. Le domaine couvert par le corpus évolue donc avec l'apparition de nouveaux concepts et de nouveaux termes relatifs à de nouveaux symptômes, à l'intégration de technologies de pointe ou à la mise sur le marché de nouveaux types de véhicules. Nous avons donc dû concevoir des mécanismes de prise en compte de l'évolutivité du corpus pour le processus d'indexation de la base d'expériences⁴. Ceux-ci ne sont pas encore implantés et testés mais ils permettront d'effectuer des mises à jour (terminologiques et/ou structurelles) de l'ontologie qui devront garantir sa cohérence.

Dans le but d'adapter la RTO aux autres modes de raisonnement, nous avons pensé accroître le corpus avec des textes portant sur des connaissances différentes de celles disponibles dans les fiches de réparation. En ceci, les documents de conception semblaient une piste intéressante. Cependant,

du fait de leur structure non textuelle (sous forme de schémas) et de leur focalisation sur les composants (peu utiles pour nos besoins), ceux-ci se sont avérés inadéquats.

Enfin, si les fiches de réparation semblaient suffire pour construire la RTO en question, les principes de modélisation d'une ontologie exigent d'avoir accès à des connaissances définitoires. Comme la base d'expériences s'appuie sur des connaissances opératoires, elle permet la construction d'un modèle de la tâche et du domaine, mais pas l'élaboration d'une ontologie. Ne disposant pas de documents appropriés (e.g. documents de formation au métier de garagiste), nous avons décidé d'utiliser un second type de ressource cognitive, à savoir un expert du domaine. Celui-ci nous permettrait de plus de valider par ses connaissances la structure ontologique bâtie à partir du corpus.

4.3 Structuration de l'ontologie

4.3.1 Concepts centraux et rôles de l'ontologie

Une fois le corpus d'étude constitué, le choix des concepts essentiels à modéliser peut se faire sur la base de l'usage des termes en corpus, et à l'aide de connaissances a priori sur le domaine et la tâche. Dans le cas de notre étude, la particularité du corpus n'est pas liée aux termes utilisés, mais à la structuration des fiches de réparation, qui reflète la démarche du garagiste. Les champs de la fiche directement utilisés dans l'application, " constatation " et " diagnostic " correspondent l'un à la description des *symptômes*, l'autre à celle de *composants*, de *tests d'état*, d'*hypothèses de panne* (cause possible d'un symptôme) et de *normes* si l'on reprend le vocabulaire de la modélisation à la KADS de la tâche de diagnostic. La structure de la fiche comporte en plus la notion de *réparation* dans le champ " remède après-vente ". Selon le modèle de la tâche associée au module RXP (fig. 3), seuls nous intéressent les rôles de *symptôme*⁵ (en entrée) et de *prestation* (en sortie). En effet, nous adoptons le point de vue d'un garagiste en phase de diagnostic de panne. Nous avons choisi d'intégrer ces concepts dans l'ontologie selon une modélisation discutée dans la partie suivante.

Dans le cadre d'une stratégie descendante de construction, ces rôles nous ont permis de tracer les grandes lignes de l'ontologie à réaliser et de définir les types de concepts du domaine à identifier. De plus, les fiches étant très structurées, chacune des parties fait référence à des concepts de types particuliers : les champs " constatation " et " diagnostic " ont ainsi été exploités en priorité pour trouver des concepts caractérisant les *symptômes* et les *prestations*.

De plus, le choix des concepts a été guidé par la nécessité de disposer d'un ensemble d'index judicieux pour la recherche de fiches, capables d'isoler chaque type de panne. Dans cette optique, la notion de symptôme était un concept

⁴Pour plus de détails, voir en 4.5.3

⁵Nous le définissons ici plus formellement comme une observation non conforme à un modèle de fonctionnement nominal.

primordial, tant pour son pouvoir discriminant sur les différentes sortes de panne que pour sa fréquence d'utilisation dans le diagnostic automobile. Les concepts ont donc été identifiés en réponse aux contraintes imposées par la tâche et l'application ciblées.

4.3.2 Modélisation d'un concept

Au delà du choix des concepts et des rôles, la tâche oriente la manière de les décrire. En effet, la notion de symptôme est un des rôles principaux dans un raisonnement de diagnostic. Dans la plupart des méthodes de résolution applicable au diagnostic (22), Le symptôme doit permettre, en référence à un modèle de fonctionnement de l'objet à diagnostiquer, de repérer un écart par rapport au fonctionnement attendu et d'identifier un composant en panne. Ici, la représentation de la notion de symptôme doit donc aiguiller vers des dysfonctionnements visibles du véhicule ou de parties de véhicule, à savoir des prestations. Du fait de la forte structuration des fiches, on s'attend à trouver les symptômes dans le champ " constatation ". Les termes identifiés dans ce champ grâce à l'extracteur de terme (Syntex et Upery) renvoient à des concepts spécifiques de type problème, **prestation** et contexte :

- " la **climatisation** ne fonctionne pas *au ralenti* "
- " allumage du **témoin ABS** "
- " *en accélération*, le **moteur fume noir** "

La notion de symptôme est représentée en faisant appel à ces concepts. Il est difficile de trouver un critère ontologique de différenciation qui organiserait les symptômes : toute différence entre deux symptômes renvoie à la comparaison des prestations et/ou des problèmes associés. Nous avons donc modélisé le rôle de symptôme sous la forme de concept défini (au sens de la logique de description). Formellement, nous le définissons comme un problème affectant une prestation dans un certain contexte (fig. 4). Ainsi, les concepts définis nous semblent adaptés à la représentation d'un rôle comme symptôme.

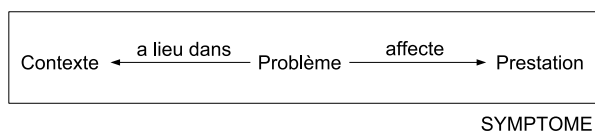


FIG. 4 – Représentation du symptôme

La représentation de symptômes particuliers suppose donc d'identifier les différentes prestations (voir 4.3.3) et les problèmes possibles. Le concept de problème est spécialisé selon des critères que nous a fourni l'expert. La différenciation reflète la manière de percevoir les problèmes par l'utilisateur (fig. 5), parmi lesquels :

- comportement attendu non réalisé / comportement non prévu,
- observation instantanée / observation moyennée.

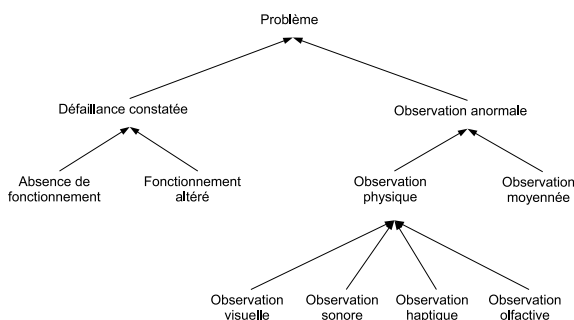


FIG. 5 – Partie supérieure de l'ontologie des problèmes

4.3.3 Organisation hiérarchique des concepts

L'application pour laquelle est développée la RTO influence évidemment la structuration des concepts entre eux. Nous soulignons ici un impact de l'application moins évident, à savoir celui sur le degré de décomposition de l'ontologie. En effet, il faudra plus ou moins différencier certains concepts d'un point de vue hypéronymique, de manière à servir au mieux les besoins applicatifs.

Pour que la RTO assure la collaboration de plusieurs méthodes de raisonnement au sein du projet MODE, nous avons dû détailler avec précision les différents niveaux de prestations véhicule. L'objectif poursuivi était double : fournir une vision du domaine proche de celle d'un réparateur en garage (et rendre l'utilisation de l'ontologie plus " intuitive "), mais aussi avoir la possibilité d'isoler certains sous-systèmes suspects au cours d'un processus de diagnostic à base de modèles (MBR). Pour cela, nous nous sommes inspirés des approches (11; 18), pour lesquelles le véhicule est considéré comme un ensemble de systèmes réalisant des fonctions pour le conducteur ou les passagers. Dans ce cadre, la prestation telle qu'elle a été définie en 4.1 correspond à une macro-fonction.

Le module RXP essaie d'établir un lien entre les connaissances comportementales et structurelles sur un système d'une part (*comment fonctionne le système ?*) et les connaissances téléologiques d'autre part (*quel est le but du système ?*). Plus concrètement, le module doit faire l'association entre des équations de fonctionnement d'un sous-système du véhicule et la prestation que celui-ci remplit pour le client. Pour cela, le module RXP doit proposer au garagiste une série de tests dans le but de parvenir à un diagnostic de la panne. Il faut donc disposer d'une structure hypéronymique des prestations suffisamment détaillée pour qu'on puisse les distinguer selon leur mode d'activation (essuyage avant par commande manuelle / essuyage avant automatique). Nous avons mis en place cette structure en nous appuyant sur des hiérarchies disponibles auprès de constructeurs automobiles et sur les prestations mention-

nées dans les fiches. Nous obtenons un squelette schématisé en partie sur la figure 6.

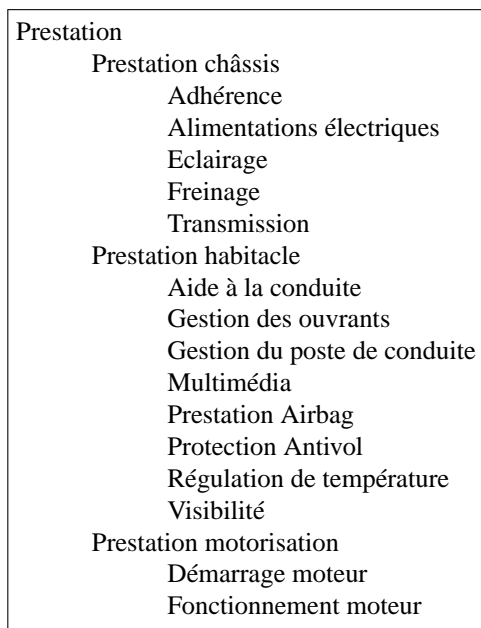


FIG. 6 – Partie supérieure de l’ontologie des prestations

A l’inverse, l’indexation conceptuelle des fiches ne nécessite pas une modélisation poussée des composants de fonctionnement qui entrent dans la réalisation de ces prestations. En effet, la recherche au sein des fiches ne se fait pas sur la base des pannes identifiées sur des composants mais sur celle des problèmes constatés sur des prestations. La structure de cette partie de l’ontologie s’en est trouvée simplifiée. Toutefois, si une fonctionnalité supplémentaire concernant les composants devait être ajoutée (e.g. associer certains composants avec leur schéma électrique), il faudrait alors utiliser des critères de différenciation pour détailler les prestations en types de composants selon les besoins.

À la suite de cette étape de structuration, l’ontologie résultante comporte quatre concepts généraux bien distincts (fig. 7), environ 300 concepts primitifs différents (27% sous prestation, 27% sous composant, 26% sous contexte et 20% sous problème) et des concepts définis de type symptôme. Parmi les concepts primitifs, seuls ceux de type prestation peuvent être considérés comme des rôles joués par des composants. Les symptômes sont aussi des rôles, définis à partir d’un problème, d’une prestation et d’un contexte (éventuellement vide).

4.4 Choix du langage de formalisation

Selon les contraintes, une ontologie peut être représentée de façon plus ou moins formelle (14). Il faut donc bien analyser les besoins pour lesquels est construite cette ressource afin de choisir un langage de description adéquat.

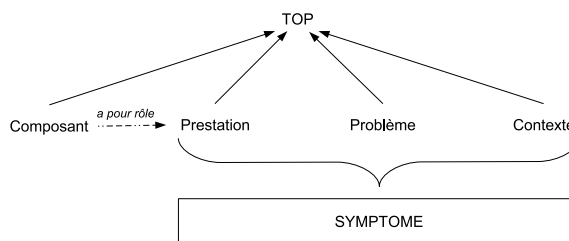


FIG. 7 – Partie supérieure de l’ontologie résultante

Actuellement, trois langages principaux de représentation d’ontologie sont répertoriés par (4) dans le cadre du Web sémantique :

1. Les cartes topiques, standard ISO à vocation annotative,
2. RDF Schema, extension de RDF permettant la spécification de classes et de relations élémentaires,
3. OWL, inspiré des logiques de description, avec une plus grande expressivité grâce aux notions de concept, propriété et axiome.

Selon (4), seuls les deux derniers peuvent être qualifiés de langages de définition d’ontologies. Généralement, les cartes topiques sont utilisées dans le cadre du Web Sémantique pour marquer des relations entre ressources ou leur associer des méta-données.

Dans notre projet, nous souhaitons que l’ontologie permette d’optimiser la recherche d’information en raisonnant sur les relations entre concepts ou leurs propriétés. En effet, en faisant des inférences sur certains objets du domaine, en l’occurrence ici sur les symptômes, une requête peut être reformulée et renvoyer à la recherche d’un ensemble plus large ou plus pertinent de concepts. Par exemple, on sait qu’une " fumée noire à l’échappement " est symptomatique d’une " surconsommation de carburant ". Lorsque le garagiste spécifie un symptôme de surconsommation, il serait intéressant de lui demander s’il constate des fumées noires, ou de rechercher a priori des fiches traitant des deux symptômes. Ceci élargira le champ de la prospection effectuée par le module RXP dans les fiches.

De plus, les choix de modélisation faits préalablement peuvent imposer certaines contraintes supplémentaires sur le langage à utiliser. Dans le cadre de notre expérience, il est nécessaire de disposer d’un langage permettant de représenter un symptôme sous la forme d’un concept défini.

Pour ces différentes raisons, nous avons choisi de manipuler l’ontologie résultante sous le format OWL DL.

4.5 Impacts sur la terminologie

4.5.1 Spécificité terminologique

Comme nous nous plaçons dans un paradigme de construction de RTO à partir de textes, nous partons de l’hy-

Classe sémantique	Proportion	Exemples de termes	Profondeur moyenne dans RTO
composant	31%	sonde de température, calculateur d'injection	4,08
contexte	20%	moteur coupé, régime moteur	non disponible ⁶
problème	17%	bruit de claquement, mauvais fonctionnement	3,91
prestation	13%	feux de détresse, démarrage moteur	3,92
symptôme	6%	ralenti instable, démarrage impossible	non pertinent ⁷
divers	13%	...	-

FIG. 8 – Répartition par classe sémantique pour les 200 termes les plus fréquents du corpus

pothèse selon laquelle une partie importante des termes présents dans le corpus sont spécifiques du domaine concerné. La figure 8 nous conforte dans cette idée : on constate que parmi les 200 termes les plus fréquents, au moins 70% peuvent être reliés à des classes sémantiques spécifiques au domaine du diagnostic automobile (à savoir composant, contexte, prestation et symptôme).

Une autre observation sur le corpus nous permet de mesurer l'influence du domaine sur la spécificité de la terminologie. En effet, si l'on s'intéresse à la profondeur moyenne à laquelle on trouvera le concept désigné par un terme donné (i.e. nombre de noeuds concepts traversés pour parvenir à lui depuis le haut de l'arborescence hypéronymique), on mesure une profondeur moyenne de 4, avec peu de variations selon la grande catégorie ontologique à laquelle appartient le concept (fig. 8). Plus intéressant encore, 87% des termes pris en compte correspondent à une profondeur dans l'ontologie strictement supérieure à 3. Ces résultats pourraient s'interpréter en termes de technicité du domaine : dans les niveaux supérieurs de la RTO, les critères de séparation entre différents concepts proviendraient plus d'un besoin de classement que de réelles occurrences des termes correspondants dans le corpus.

4.5.2 Précision de la terminologie

Le domaine détermine également les termes à intégrer dans la RTO, et ce en fonction de la spécificité terminologique du corpus. En effet, indexer un corpus de langue générale supposerait de disposer d'une ressource très large et générique dont l'utilisation peut nécessiter une phase de désambiguïsation (due à la polysémie des termes). Au contraire, lorsque l'on se trouve sur un domaine particulier, il est indispensable de faire référence à une RTO spécialisée d'un domaine et plus précise : de nombreux termes polysémiques en l'absence de tout contexte prennent une sémantique plus précise sur un domaine particulier. Par exemple, le verbe intransitif "patiner" possède deux sens principaux dans le Trésor de la Langue Française (TLF)⁸. Le premier correspond à l'action d'évoluer sur la glace ou le bitume à l'aide de patins tandis que le second se rapporte à un

dérapiage par manque d'adhérence au sol. Ces définitions se distinguent notamment par la nature du sujet (personne ou objet). En examinant le corpus, nous constatons que le verbe "patiner" est utilisé uniquement avec des objets du domaine automobile : "une courroie patine", "la poulie patine", "le véhicule patine", "l'embrayage patine"⁹. Pour notre RTO, ce verbe a donc un sens univoque (celui de la seconde définition) et ne renvoie qu'à un seul concept.

Dans un même ordre d'idée, afin de représenter fidèlement les usages linguistiques d'une communauté, le modèle d'une RTO peut aller à l'encontre de conceptions de sens commun. Ainsi, le terme "témoin" défini entre autres par le TLF comme une "chose qui permet de constater, de vérifier" est par essence un hypéronymie de "voyant" ("lampe s'allumant pour attirer l'attention sur un danger, un dysfonctionnement ou pour indiquer qu'un appareil fonctionne"). Toutefois, l'observation des expansions de ces termes en corpus nous a révélé qu'ils étaient utilisés de manière équivalente : environ 55% de termes spécifiant "témoin" (e.g. "témoin de charge batterie", "témoin de climatisation") ont un équivalent avec "voyant" ("voyant de charge batterie", "voyant de climatisation"). Ceci nous a conduit à regrouper les deux termes comme des synonymes désignant un même concept (dont la définition est celle d'un voyant).

4.5.3 Association des termes aux concepts

Le nombre et la variété de termes associés aux concepts dépendent étroitement des besoins applicatifs pour lesquels la RTO est bâtie. Si par exemple l'ontologie que nous avons construite avait pour but de servir de structure de base d'un portail dans le cadre du Web sémantique, elle aurait une composante terminologique beaucoup moins importante. Dans le cas de l'indexation conceptuelle, l'association (automatique) de concepts à des fragments de textes requiert de caractériser les formes linguistiques exprimant les concepts, c'est-à-dire les termes les désignant ou bien les contextes dans lesquels ils sont présents. Etant donné que les fiches sont peu rédigées, nous avons choisi de nous appuyer sur les termes pour identifier la présence de concepts. De ce

⁶La structuration de cette partie de la RTO n'est pas terminée.

⁷Pour explication, voir en 4.3.2.

⁸<http://atilf.atilf.fr/>

⁹Nous savons que ce sont des objets du domaine soit par les résultats préalables de l'étude du corpus, soit par le recours à un expert.

fait, la RTO doit comporter le plus de termes possibles associés aux concepts.

Etant donnée cette nécessaire richesse terminologique, nous avons essayé de faciliter le travail de recherche des termes associés à chaque concept. Ainsi, une fois déterminée la structure de l'ontologie, l'étape suivante consistera à "peupler" l'ontologie en utilisant les spécificités de l'algorithme d'indexation. En effet, l'application d'indexation conceptuelle impose de pouvoir comparer deux fiches de la base d'expériences selon les symptômes de panne décrits. Cela présuppose d'avoir systématiquement reconnu au moins un symptôme dans chaque fiche. Par la structure même des fiches, nous sommes assurés qu'elles satisfont toutes cette exigence. L'algorithme d'indexation des fiches a pour tâche de retrouver dans le champ symptôme les termes associés à un problème et à une prestation qui sont reliés dans l'ontologie (et éventuellement un contexte d'apparition de la panne). Il peut échouer pour deux raisons :

- il trouve un problème et une prestation mais ceux-ci ne sont pas associés dans l'ontologie,
- il ne trouve pas d'occurrence d'au moins un des deux concepts.

Le premier cas nous permettra de corriger l'ontologie en y ajoutant sous forme de concepts définis les symptômes que nous aurions oubliés pendant la phase de construction¹⁰. Le second cas est celui qui nous intéresse ici : l'algorithme ne trouve pas de terme approprié car celui-ci n'est sans doute pas encore rattaché au concept qu'il désigne. En présentant à l'analyste le champ symptôme problématique, l'algorithme lui permettra ainsi de repérer plus rapidement des termes à rattacher à un concept (problème ou prestation). On s'attend à ce que le processus de peuplement, fastidieux au départ, identifie de moins en moins de nouveaux termes avec le temps. Une fois que la totalité de la terminologie du domaine aura été associée aux concepts appropriés, l'algorithme sera - en théorie - entièrement silencieux. Si l'enrichissement de la base de fiches s'accompagne d'une évolution de la terminologie du domaine, il sera possible de procéder à une nouvelle phase de peuplement. Si de nouveaux concepts apparaissent, leur insertion cohérente dans l'ontologie selon les critères de différenciation en vigueur relèvera de la responsabilité de l'expert chargé de cette maintenance.

5 Conclusion et perspectives

Entre l'application cible, la tâche et le domaine, il est difficile de faire clairement la part des influences sur le processus de modélisation et le contenu du modèle construit. L'application cible (dans notre cas la recherche d'information dans des textes) englobe en partie le fait que le modèle reflète les connaissances d'un domaine mises en oeuvre pour réaliser une certaine tâche (le diagnostic de panne).

¹⁰Pour cela, il suffira de rajouter un lien *affecte* entre le problème et la prestation trouvés dans la fiche.

La particularité de notre application est justement qu'elle n'automatise pas la tâche en effectuant une résolution de problème, mais en réalisant une recherche d'information. De notre expérience, il ressort que l'application et le besoin de raisonner sur les connaissances modélisées déterminent le processus de modélisation (choix des modes d'analyse et des logiciels de TAL) et le degré de formalisation des connaissances. Pour leur part, la tâche et le domaine caractérisent le contenu de l'ontologie (choix des concepts, manière de les définir, niveau de détail).

La notion de rôle (au sens de classe conceptuelle utilisée dans le raisonnement) tient ici une place charnière : à un rôle, sont associés des types de connaissances à définir en tant que concepts du domaine. Nous avons choisi de placer les rôles utilisés pour le diagnostic au sein de l'ontologie, qui couvre ainsi le domaine des connaissances du spécialiste en diagnostic automobile, suffisantes pour retrouver des cas analogues. Dans l'ontologie, les rôles sont représentés comme des concepts de haut niveau (i.e. les prestations), lorsqu'ils mettent en jeu un ou plusieurs concepts qui ne jouent pas d'autre rôle (i.e. les composants), ou comme des concepts définis (i.e. les symptômes) quand ils impliquent la mise en relation de plusieurs concepts. Il s'agit là d'un choix pragmatique dont nous devons évaluer la généralisation dans d'autres contextes analogues, où le système réalise une recherche d'information à propos d'une tâche.

Ainsi, la définition des concepts rend compte d'un point de vue particulier sur le domaine. Même si l'ontologie à construire a pour vocation d'indexer des documents, ce projet confirme que le vocabulaire et les connaissances présents dans ces documents ne suffisent pas toujours pour obtenir des définitions. Celles-ci se trouvent davantage dans des documents de type pédagogique ou de formation, ou bien elles sont à recueillir auprès d'experts. De même, comme l'application utilisant l'ontologie vise la recherche d'information, la composante terminologique associée à l'ontologie doit être d'autant plus riche. Ceci confirme le besoin d'adapter la terminologie à l'application et au domaine à couvrir.

Au delà de ces énoncés, il serait intéressant de localiser explicitement ces influences dans une méthode comme TERMINAE. Il s'agit d'une première perspective à notre travail. De plus, la notion de RTO telle que nous l'avons utilisée renvoie plus à un modèle conceptuel d'un domaine qu'à une ontologie au sens où le seul principe ontologique appliqué est la différenciation. Une approche complémentaire serait de déterminer la manière de prendre en compte l'application ciblée pour construire une ontologie selon des méthodes plus contraintes comme OntoSpec (17) ou OntoClean (15).

Références

- [1] N. Aussenac-Gilles, B. Biébow et S. Szulman. D'une méthode à un guide pratique de modélisation de

- connaissances à partir de textes. In F. Rousselot, éditeur, *Actes des 5e rencontres Terminologie et IA (TIA 2003)*, pages 41–53, Avril 2003.
- [2] N. Aussenac-Gilles, A. Condamines et S. Szulman. Prise en compte de l'application dans la constitution de produits terminologiques. In *Actes des 2e Assises Nationales du GDR I3*, pages 289–302. Cépaduès Editions, Dec 2002.
- [3] B. Bachimont. Art et sciences du numérique : ingénierie des connaissances et critique de la raison computationnelle, 2004. Mémoire d'habilitation à diriger des recherches, Université Technologique de Compiègne.
- [4] J.F. Baget, E. Canaud, J. Euzenat et M. Saïd-Hacid. Les langages du web sémantique. *Revue I3*, Hors Série 2004.
- [5] M. Baziz. *Indexation conceptuelle guidée par ontologie pour la recherche d'information*. PhD thesis, Université Paul Sabatier, 2005.
- [6] D. Bourigault. Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Actes de la 9e conférence annuelle sur le traitement automatique des langues (TALN 2002)*, 2002.
- [7] D. Bourigault, N. Aussenac-Gilles et J. Charlet. Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. In J.M. Pierrel et M. Slodzian, éditeurs, *Techniques informatiques et structuration de terminologies*, volume 18/1 of *Revue d'Intelligence Artificielle*. Hermes Sciences, 2004.
- [8] D. Bourigault et C. Fabre. Approche linguistique pour l'analyse syntaxique de corpus. In *Sémantique et Corpus*, volume 25 of *Cahiers de grammaire*. Université de Toulouse-Le Mirail, 2000.
- [9] J. Breuker et R. Hoekstra. Core concepts of law : taking common-sense seriously. In *Formal Ontologies in Information Systems FOIS-2004*, pages 210–221. IOS-Press, 2004.
- [10] J. Charlet. L'ingénierie des connaissances : développements, résultats et perspectives pour la gestion des connaissances médicales, 2002. Mémoire d'habilitation à diriger des recherches, Université de Pierre et Marie Curie.
- [11] L. Chittaro, G. Guida, C. Tasso et E. Toppano. Functional and teleological knowledge in the multimodeling approach for reasoning about physical systems : a case study in diagnosis. In *Proc. of the IEEE Transactions on Systems, Man and Cybernetics*, volume 23, 1993.
- [12] A. Condamines. Sémantique et corpus spécialisés : constitution de bases de connaissances terminologiques, 2003. Mémoire d'habilitation à diriger des recherches, CNRS & Université de Toulouse-Le Mirail.
- [13] N. Fridman-Noy et C. Hafner. The state of the art in ontology design : a survey and comparative review. *Artificial Intelligence Magazine*, pages 53–74. 1997.
- [14] A. Gomez-Pérez, M. Fernando Lopez et O. Corcho. *Ontological engineering : with examples from the area of knowledge management, e-commerce and the semantic web*. Springer, 2004.
- [15] N. Guarino et C. Welty. A formal ontology of properties. In R. Dieng et O. Corby, éditeurs, *12th International Conference in Knowledge Engineering and Knowledge Management (EKAW'00)*, pages 97–112. Springer Verlag, 2000.
- [16] G. Van Heijst, A. Schreiber et B. Wielinga. Roles are not classes : a reply to nicola guarino. volume 46 of *Journal of Human-Computer Studies*, pages 311–318. 1997.
- [17] G. Kassel. Integration of the dolce top-level ontology into the ontospec methodology, 2005. LARIA Research Report 2005-08, Université de Picardie Jules Verne. Available at <<http://hal.ccsd.cnrs.fr/ccsd-00012203>>.
- [18] Y. Kitamura et R. Mizoguchi. Ontology-based functional-knowledge modeling methodology and its deployment. In *Proc. of EKAW 2004*, pages 99–115, 2004.
- [19] A. Maedche. *Ontology learning for the Semantic Web*. Kluwer Academic, 2002.
- [20] T. Ait El Mekki et A. Nazarenko. Comment aider un auteur à construire l'index d'un ouvrage ? In Y. Toussein et C. Nédellec, éditeurs, *Actes du Colloque International sur la Fouille de Texte CIFT'2002*, pages 141–158, Octobre 2002.
- [21] C. Reynaud, N. Aussenac-Gilles, P. Tchounikine et F. Trichet. The notion of role in conceptual modeling. In R. Benjamins et E. Plaza, éditeurs, *Proceedings of EKAW97 - European Knowledge Acquisition Workshop*, pages 221–236. Springer Verlag, 1997.
- [22] G. Schreiber, H. Akkermans, A. Anjewierden, K. de Hoog, N. Shadbolt, W. Van De Velde et B. Wielinga, éditeurs. *Knowledge Engineering and Management : the CommonKADS Methodology*. MIT Press, 2000.