

Vers l'automatisation de la construction de systèmes de médiation pour le commerce électronique

C. Reynaud, G. Giraldo

Université Paris-Sud, CNRS UMR 8623, INRIA-Futurs
L.R.I., Bâtiment 490, 91405 Orsay cedex
{cr,giraldo}@lri.fr

I. Introduction

Le commerce électronique est un domaine d'application en plein essor. Beaucoup de services sont d'ores et déjà accessibles et les évolutions du Web ne feront qu'amplifier ce développement.

Aujourd'hui, un utilisateur qui désire bénéficier d'un service de commerce électronique en demande l'exécution après avoir fourni toutes les informations utiles selon les modalités propres au service. Si l'utilisateur ne connaît pas l'adresse du service qui l'intéresse, la recherche de celui-ci s'effectue via un moteur de recherche qui, à partir de l'expression du besoin retourne un ensemble d'adresses de sites pouvant répondre à ce besoin, ces sites étant sélectionnés très souvent par application d'une technique basée sur des mots-clefs. Ce n'est qu'après consultation de chacun des sites que l'utilisateur s'aperçoit que les services associés aux sites qui lui ont été proposés sont ou ne sont pas pertinents par rapport à la demande effectuée.

Le Web sémantique a pour objectif de permettre à des machines d'accéder aux sources d'information et aux services du Web. Cette évolution pourra permettre d'automatiser la découverte des services pertinents par rapport à une requête, et également de combiner automatiquement plusieurs services lorsque la demande de l'utilisateur ne peut être satisfaite par un seul service prédéfini mais correspond à un service complexe. En effet, une description sémantique des ressources du Web permettra de les identifier de façon plus précise. Par ailleurs, il sera possible d'exécuter automatiquement les services du Web.

Les travaux que nous effectuons se situent dans le cadre du Web Sémantique. Nous nous plaçons du point de vue de l'utilisateur final afin de l'aider à utiliser les services de commerce électronique accessibles via le Web, en lui proposant un système de médiation. Un système de médiation est une interface entre l'utilisateur et les services du Web d'un domaine d'application donné. Il doit donner l'impression à l'utilisateur d'utiliser un unique système alors que la satisfaction de sa demande peut exiger de composer plusieurs services délivrés par des fournisseurs différents.

Nous présentons le contexte d'étude de notre travail et l'architecture des systèmes de médiation pour le commerce électronique auxquels nous nous intéressons. Nous décrivons ensuite la méthode que nous proposons pour automatiser la construction de l'ontologie au cœur de ces systèmes.

II. Contexte d'étude

2.1. L'intégration de services dans le cadre du commerce électronique

Le commerce électronique est un domaine en plein essor avec des enjeux économiques importants.

L'évolution vers le Web Sémantique passe par l'intégration des services accessibles de façon à ce qu'un utilisateur ait l'impression qu'il s'adresse à un système centralisé et homogène. Résoudre ce problème d'intégration signifie apporter des solutions aux problèmes d'hétérogénéité tant syntaxiques que sémantiques.

Pour favoriser l'interopérabilité syntaxique, il est nécessaire d'utiliser un ou des langage(s) standard(s) pour décrire le contenu des ressources du Web. Beaucoup de travaux portent sur ces problèmes de standardisation. Deux groupes de travail du W3C étudient ces problèmes : (1) RDFCore [1] dont les travaux s'appuient sur RDF, un langage déclaratif permettant de décrire en XML les ressources du Web indépendamment de tout domaine d'application et RDF-S [2], le langage de typage associé, (2) WebOnt [3] qui travaille à l'élaboration d'un langage standard pour les ontologies du Web en partant de la proposition DAML+OIL.

Il ressort par ailleurs des travaux effectués, qu'étant donnée la diversité des connaissances à représenter et de la façon dont ces connaissances doivent être exploitées, plusieurs langages standards sont nécessaires, ces langages devant pouvoir cohabiter en pouvant être traduit dans un même format d'échange.

L'interopérabilité passe également par la résolution du problème de l'hétérogénéité sémantique. Les ressources accessibles via le Web sont des ressources autonomes, indépendantes, conçues indépendamment les unes des autres en utilisant un vocabulaire propre, que les concepteurs ont choisi librement. Des solutions à ce problème d'hétérogénéité sémantique doivent être trouvées. Dans certains domaines du commerce électronique, des travaux de normalisation sont en cours. Leur objectif est de fournir un vocabulaire standard de description des transactions commerciales propres à un domaine.

Ces travaux sur les langages et sur l'élaboration d'un vocabulaire partagé par une communauté d'utilisateurs permettent de définir des façons normalisées de communiquer entre services d'un même domaine et offrent ainsi une solution au problème de l'interopérabilité. Néanmoins, les solutions proposées ne sont pas suffisantes. Le Web Sémantique est une infrastructure qui devra permettre aux utilisateurs humains d'accéder à un immense espace de ressources. Ainsi, les solutions du Web Sémantique doivent prendre en compte leurs besoins en essayant de les satisfaire au mieux et le plus « intelligemment » possible. Elles doivent faciliter l'expression des besoins. Enfin, il devra être possible de délivrer des résultats adaptés, voire personnalisés. Nous énonçons dans la section qui suit quelques raisons pour lesquelles les systèmes de médiation viennent compléter l'usage de formalisme ou de vocabulaire communs.

2.2. Vers des systèmes de médiation

Même si les ressources du Web sont décrites de façon homogène, il est néanmoins nécessaire de concevoir des systèmes de médiation qui aident l'utilisateur à exprimer ses besoins, à découvrir l'ensemble de tous les services pertinents étant donnée sa requête et à composer, si nécessaire, plusieurs services.

En effet, les organismes de standardisation définissent les transactions commerciales en décrivant le contenu des messages échangés de façon très détaillée, ce niveau de détail correspondant au niveau requis par les membres des organismes de standardisation, des professionnels essentiellement. Les informations présentes au sein des différents messages sont très nombreuses car elles reflètent les besoins de chaque organisation participant au

processus de standardisation. Les termes choisis sont consensuels, le consensus étant toutefois obtenu essentiellement entre professionnels. Ils correspondent parfois à des codes.

Voici, à titre d'illustration, une partie d'un message représenté sous forme de XML-Schema défini par l'OTA [5], un organisme de standardisation de messages dans l'industrie du voyage. Ce message concerne une demande de disponibilité d'un vol.

```
<OTA_AirAvailRQ xmlns:xsi=http://www.w3.org/2001/XMLSchema-instance
xsi:noNamespaceSchemaLocation="OTA_AirAvailRQ.xsd" DesiredOptionQuantity="10"
DirectFlightsOnly="false">
  <POS>
    <Source AgentSine="BSIA1234PM" PseudoCityCode="2U8" ISOCountry="US"
ISOCurrency="USD">
      <UniqueId URL=" http://provider1.org/OTAEngine/" Type=" OTA_AirAvailRQ"
Id="ID0507G4325" Instance="2001-06-03T13:09:21"/>
    </Source>
  </POS>
  <OriginDestinationInformation>
    <OriginLocation LocationCode="LON" CodeContext="IATA"/>
    <DestinationLocation LocationCode="LAX" CodeContext="IATA"/>
    <TravelDateTime TravelDate="2001-09-11T09:00:00" TravelTimeIndicator="DepartureTime"
TimeWindowBefore="P3D" TimeWindowAfter="P3D"/>
  </OriginDestinationInformation>
  ....
</OTA_AirAvailRQ>
```

Ainsi le niveau de détail, la quantité d'information et les termes retenus sont déterminés en fonction des besoins de professionnels et non pas des utilisateurs finaux, ce qui explique qu'il est très difficile pour un utilisateur final de fournir les données en entrée d'un service en étant simplement guidé par le contenu de ces messages.

Des systèmes de médiation facilitant la communication entre l'utilisateur final et les services disponibles sur le Web sont absolument nécessaires.

III. Architecture d'un système de médiation pour le commerce électronique

Nous proposons une architecture de système de médiation pour le commerce électronique dans un domaine pour lequel existent des organismes de standardisation des transactions commerciales (cf. figure n°1).

Cette architecture est spécifique à un ensemble de services variés et distribués d'un domaine donné. C'est une solution à l'intégration de services proposés par des fournisseurs offrant des services divers mais spécifiés à l'aide d'un même vocabulaire. A chacun de ces services sont associées des descriptions fonctionnelles correspondant à la définition standardisée des transactions traitées. Ces transactions sont données sous la forme de descriptions des messages d'entrée et de sortie. Les descriptions sont homogènes syntaxiquement et sémantiquement car elles sont exprimées dans un même langage et dans les termes définis par l'organisme de standardisation.

Le système de médiation comporte une ontologie des services du domaine d'étude. Cette ontologie est un support à l'expression des demandes des utilisateurs. Elle offre un vocabulaire structuré et exprimé à un meilleur niveau d'abstraction que le niveau de description des messages. C'est un ensemble de prédicats modélisant le domaine des services auxquels on s'intéresse. Elle permet, par ailleurs, de combiner les différents services accessibles. Selon l'approche médiateur, l'intégration d'information est en effet fondée sur l'exploitation de

descriptions homogènes et uniformes des fonctionnalités des services intégrés dans les termes de l'ontologie.

Les réponses à une requête posée à un médiateur sont calculées par un moteur de requêtes générique utilisable quel que soit le domaine d'application du médiateur. Notre objectif est de réutiliser le moteur de requêtes développé dans le cadre du projet PICSEL [4].

Un des problèmes à résoudre pour automatiser la construction des systèmes médiateurs concerne alors la construction de l'ontologie. Notre objectif est d'automatiser sa construction. Nous proposons une méthode décrite en section 4 basée sur l'exploitation des transactions proposées par l'organisme de standardisation et utilisées pour décrire les services dont nous facilitons l'accès.

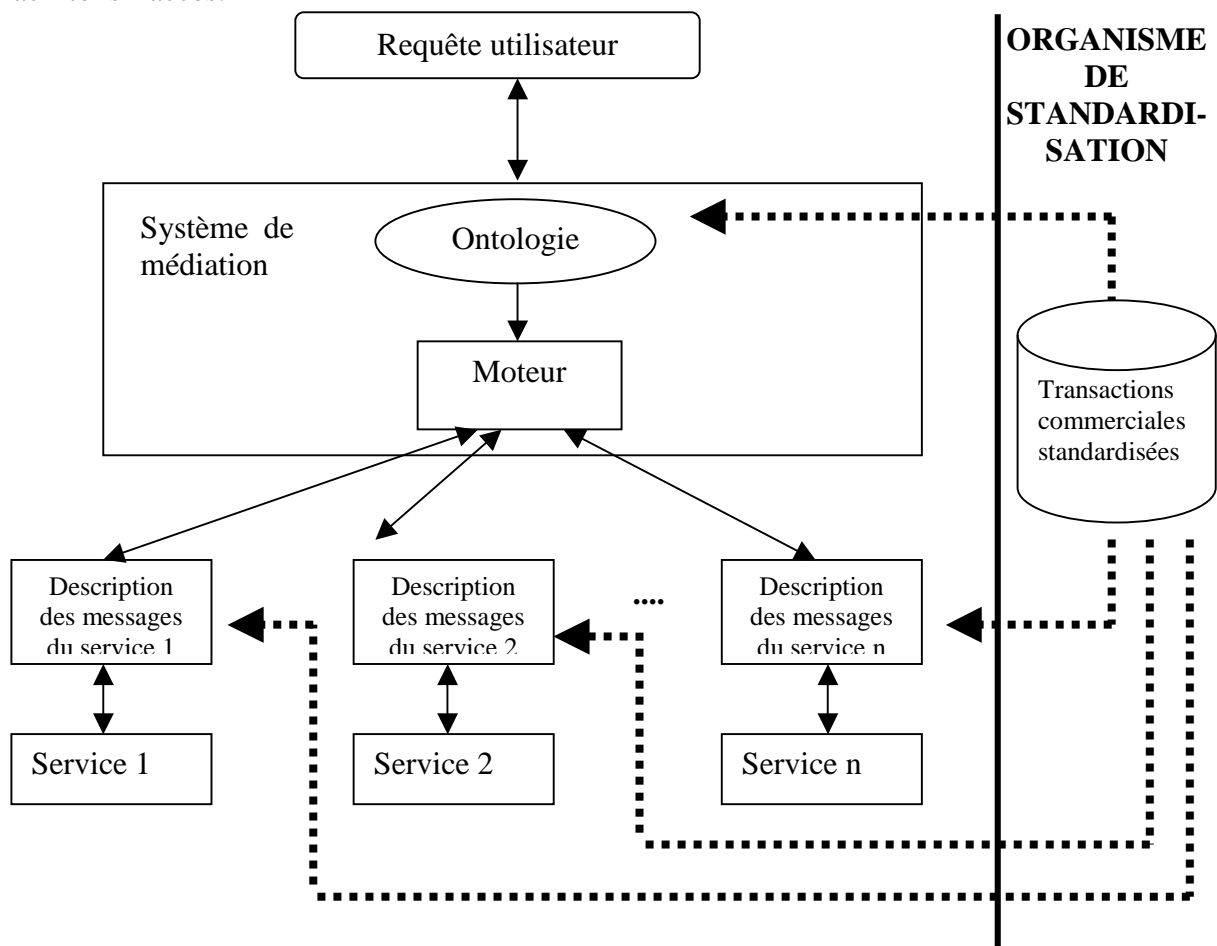


Figure n°1 : Architecture d'un système de médiation pour le commerce électronique

IV. Méthode de construction automatisée de l'ontologie

Nous nous sommes intéressées à la construction d'une ontologie représentant une catégorisation des services pour lesquels un standard de transactions est proposé sous forme de DTDs.

L'ontologie que nous souhaitons construire doit comprendre un ensemble de **classes** d'objets du domaine, chaque classe étant caractérisée par des **propriétés** et pouvant être reliée par des **relations** spécifiques au domaine à d'autres classes. Pour une classe donnée, le modèle précise la classe qui la généralise, éventuellement ses propriétés spécifiques ou bien l'ensemble des propriétés nécessaires et suffisantes d'un objet pour appartenir à cette classe. Le formalisme de représentation adopté devra permettre d'utiliser l'ensemble des outils logiciels développés dans le cadre du projet PICSEL I [4], en particulier le moteur de requêtes.

Nous proposons de construire cette ontologie en 2 étapes. Une version initiale très simple de l'ontologie est tout d'abord construite à la main de façon à fixer le point de vue à adopter pour classer les services du domaine étudié. Cette ontologie initiale est ensuite enrichie semi-automatiquement à partir des descriptions de transactions standardisées.

4.1. La construction d'une ontologie initiale simple

L'ontologie initiale construite manuellement est simple dans la mesure où elle ne correspond qu'à deux niveaux d'une hiérarchie de concepts associés aux services auxquels on veut faciliter l'accès (cf. figure n°2).

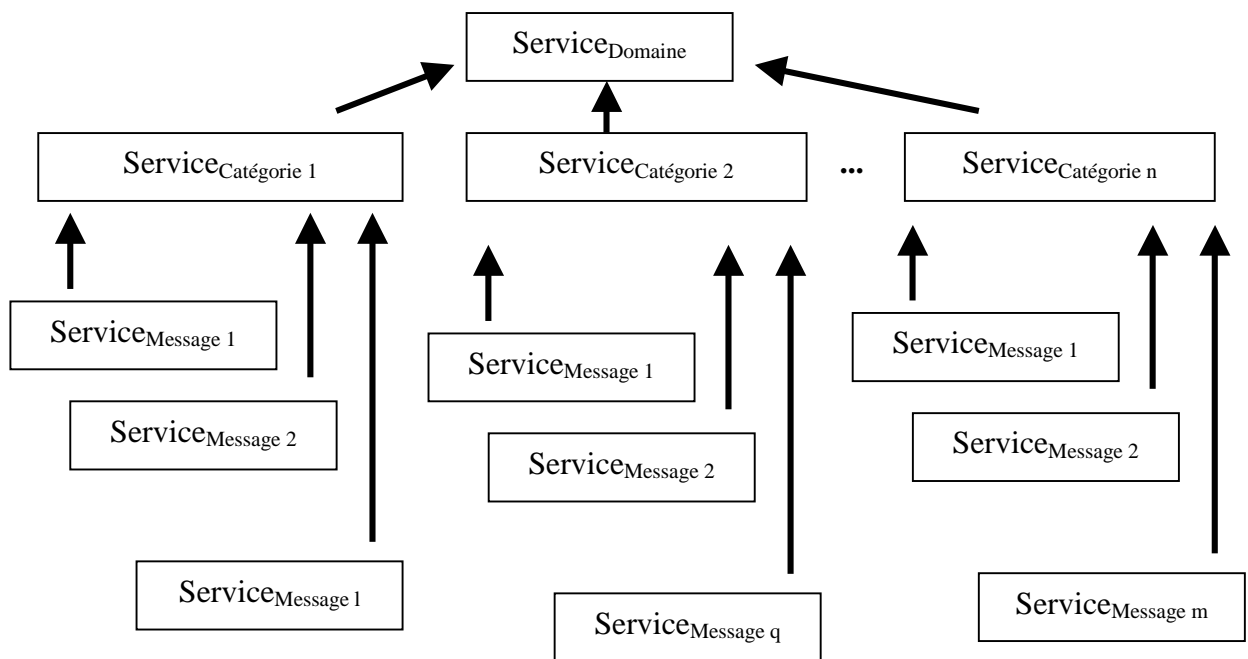


Figure n°2 : Hiérarchie de concepts initiale

Les termes associés aux concepts du niveau 2 ($Service_{Message\ i}$) correspondent aux noms des messages définis par l'organisme de standardisation. Le niveau racine et le niveau 1 sont construits manuellement. L'objectif est de présenter à l'utilisateur un ensemble de services d'un certain domaine organisés en catégories. Une telle hiérarchie servira de support à la formulation de la demande de l'utilisateur qui n'aura pas à sélectionner les services qui l'intéressent parmi un vaste ensemble de services diversifiés mais pourra faire cette sélection parmi les services appartenant à des catégories données (Exemples de catégories appliquées au domaine du tourisme: vehicle service, air service, hotel service, ... Exemples de messages pour ce même domaine: hotel search requirement, hotel availability requirement, hotel reservation requirement, ...).

4.2. L'enrichissement de l'ontologie initiale

La hiérarchie initiale de services est automatiquement enrichie à partir des transactions standards définies. L'approche retenue comporte 3 phases : (1) une phase d'extraction d'éléments utiles pour compléter l'ontologie initiale (classes, propriétés, relations), exploitant la structure des DTDs qui décrivent les messages standards du domaine, (2) une phase d'organisation des éléments acquis, (3) une phase de représentation des connaissances préalablement collectées et organisées dans un formalisme à base de classes interprétable par le moteur de requête PICSEL.

La phase d'extraction est semi-automatique. Dans un premier temps, on acquiert de façon entièrement automatique un ensemble de termes-classes, de termes-propriétés et de relations, par application d'heuristiques. L'idée de base est la suivante. Une classe est vue comme une représentation abstraite d'un ensemble d'objets « complexes » repérés, dans les DTDs, par le fait qu'il s'agit d'éléments décomposables. Etant donnée que l'approche repose sur l'exploitation de nombreux DTDs représentatives du domaine, leurs composants devraient apparaître dans au moins une des DTDs de l'échantillon. En suivant ce raisonnement, nous obtenons une méthode de repérage des termes associés aux propriétés : les termes associés à des éléments qui ne sont jamais décomposés dans aucune des DTDs. Ce processus d'extraction s'accompagne par ailleurs de traitements éliminant les doublons, extrayant les différents mots d'expressions composées, remplaçant les abréviations par leur signification en clair et éliminant certains termes jugés non pertinents pour l'ontologie. Ces traitements exploitent, d'une part des fichiers contenant les explications des abréviations et la liste des termes jugés non pertinents. D'autre part, ils exploitent un thésaurus (WordNet) pour identifier les termes pluriels et les termes composés retenus en tant que classes pour le domaine d'étude. Le processus complet d'extraction est semi-automatique dans la mesure où les termes extraits automatiquement sont présentés à l'utilisateur qui juge, après une lecture rapide, de l'opportunité d'enrichir les fichiers de termes « non pertinents » et des abréviations, et de la nécessité de re-exécuter le processus d'extraction automatique pour prendre en compte les nouvelles versions de ces fichiers. Une fois le contenu de ces fichiers arrêté, le concepteur analyse de façon plus fine les résultats produits par le logiciel, ce qui peut conduire à éliminer des termes, modifier certains noms, transformer certains termes-classes en termes-propriétés. Nous présentons ci-dessous des exemples de termes extraits de l'exploitation de DTDs fournis par l'OTA dans le domaine du tourisme.

Exemples de termes-classes : airBook, air itinerary, airline, flight segment, customer, hotel chain, hotel, travel club, etc.

Exemples de termes-propriétés : phone, phone number, payment type, personal service, price, program code, program description, program name, etc.

La phase de structuration consiste à raccrocher les éléments extraits des concepts composant la hiérarchie initiale construite manuellement. Dans le contexte dans lequel nous nous plaçons ici, la phase de structuration est facilitée dans la mesure où les concepts du niveau 2 de la hiérarchie initiale construite manuellement sont des noms de messages standards. Ils font partie des composants extraits automatiquement, pour lesquels des propriétés et des relations avec d'autres concepts ont été trouvées. La structuration consiste alors à construire, pour chacun de ces concepts, un réseau de relations les liant à d'autres classes du domaine (autres classes extraites liées et jugées pertinentes par le concepteur), chacune des classes étant caractérisées par des propriétés.

5. Conclusion

Cet article décrit la nécessité des systèmes de médiation en illustrant ce besoin dans le domaine du commerce électronique. Par ailleurs, il présente les principes de base d'une approche de construction automatisée de l'ontologie, au centre des systèmes médiateurs et dont l'automatisation est indispensable pour que ces systèmes passent à l'échelle du Web. La partie construction automatisée de l'ontologie a été implémentée et une expérimentation concernant le processus d'extraction a été menée à partir de DTDs fournis par l'OTA (Open Travel Alliance) [5]. L'OTA est un consortium qui regroupe plus de 150 organisations relatives à l'industrie du voyage : des agences de voyage, des hôtels, des agences de location de voitures, des compagnies aériennes, etc. En association avec la DISA (Data Interchange Standards Association), cet organisme a développé des standards de communication basés sur XML pour faciliter l'emploi du commerce électronique. Deux séries de données ont été publiées, l'une sous forme de 15 DTDs (317 lignes), l'autre sous forme de 77 XML-Schema (5717 lignes) que nous avons traduites sous forme de DTDs avec le logiciel XML Spy [6]. L'expérimentation, qui a porté sur les deux séries de données, a donné des résultats satisfaisants décrits dans [7], nous encourageant à poursuivre l'approche.

Références

[1] <http://www.w3.org/TR/rdf-primer>

[2] <http://www.w3.org/TR/rdf-schema>

[3] <http://www.w3.org/TR/web-ont-req>

[4] Marie-Christine Rousset, Christine Froidevaux, Hélène Gagliardi, François Goasdoué, Chantal Reynaud, Brigitte Safar, "Construction de médiateurs pour intégrer des sources d'information multiples et hétérogènes : le projet PICSEL", Revue I3, à paraître.

[5] <http://www.opentravel.org>

[6] <http://www.xmlspy.com>

[7] Gloria Giraldo, Chantal Reynaud, "Construction semi-automatique d'ontologies à partir de DTDs relatives à un même domaine", 13èmes journées francophones d'Ingénierie des Connaissances, Rouen, 28-30 Mai 2002.