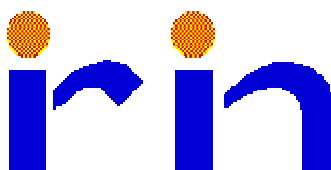


BONOM : un système Multi–Agents pour la recherche d’informations sur Internet dirigée par la connaissance.

S. Cazalens, E. Desmontils, C. Jacquin, P. Lamarre

IRIN, Rue de la Houssinière BP92208, F44322 Nantes cedex 3, France

--IC--



Research Report

N°00.1

March 2000

S Cazalens, E. Desmontils, C. Jacquin, P. Lamarre

**BONOM : un système Multi-Agents pour la recherche d'informations sur Internet
dirigée par la connaissance.**

Les rapports de recherche de l'institut de Recherche en informatique de Nantes sont disponibles aux formats [Postscript@](#) et [PDF@](#) à l'URL:

<http://www.sciences.univ-nantes.fr/irin/Vie/RR/>

Research reports from the institut de Recherche en informatique de Nantes are available in [Postscript@](#) and [PDF@](#) formats throughthe l'URL:

<http://www.sciences.univ-nantes.fr/irin/Vie/RR/>

March 2000 by S Cazalens, E. Desmontils, C. Jacquin, P. Lamarre

BONOM : un système Multi-Agents pour la recherche d'informations sur Internet dirigée par la connaissance.

Sylvie Cazalens, Emmanuel Desmontils, Christine Jacquin et Philippe Lamarre

Résumé

La croissance très importante du réseau Internet nécessite actuellement des outils de recherche de plus en plus performants permettant de discerner efficacement les informations intéressantes sans pour autant surcharger de manière abusive l'encombrement des canaux de communication. Le projet BONOM entre dans ce cadre en proposant de déployer une (ou plusieurs) société(s) d'agents intelligents d'informations à l'échelle de la planète. Ces sociétés sont spécialisées dans certains domaines de la connaissance leur permettant ainsi de répondre intelligemment aux requêtes des utilisateurs. Cette connaissance spécifique, maintenue sous forme d'ontologies, intervient à tout moment et à tous les niveaux de la société pour le déploiement des agents (leur organisation, leur configuration...) ainsi que pour la création, la circulation et le traitement des requêtes de recherche d'informations.

1 Introduction

L'objectif général du projet BONOM [1] est d'améliorer la gestion, la recherche et la consultation d'information sur le réseau Internet, sans toutefois le surcharger. Nous proposons une nouvelle architecture, basée sur des agents d'information distribués « intelligents » dans laquelle les gros moteurs de recherche actuels (tels Yahoo ou Altavista), bien qu'intégrables, ne sont plus des points de passage obligés lorsqu'on ne sait à quelle adresse chercher.

Les outils de recherche d'information sur Internet sont de quatre types [2] :

- Géographique ;
- Thématiques ;
- Par Index ;
- Par Méta-Index.

Une société BONOM est un *outil semi-automatique de recherche d'informations par index* sur Internet. Seulement, contrairement aux outils semi-automatiques classiques (outils d'aide à la consultation de moteurs de recherche, aspirateurs...), il ne demande pas de charger les pages possibles en local pour les analyser. L'utilisateur n'a pas besoin des grosses ressources (espace disque, processeurs) puisque les informations sont distribuées sur le réseau et que la recherche est effectuée de façon distribuée par les différents agents de la société

Le choix de cette nouvelle architecture se base sur un certain nombre de constatations aussi bien au niveau des sites de diffusion d'information qu'au niveau des utilisateurs.

Du point de vue des sites, le projet a pour but de fournir aux rédacteurs de pages et aux gestionnaires des outils pour :

- faciliter la structuration et la maintenance du site ;
- mieux contrôler la vue externe de ce site ;
- étudier les caractéristiques des visiteurs et leur comportement.

Du point de vue des utilisateurs, le projet vise à les aider :

- dans l'élaboration et la reformulation des requêtes ;
- dans la prise de connaissance et la compréhension des résultats obtenus.

Nous envisageons de mettre en place des outils permettant à l'utilisateur d'effectuer une recherche en rédigeant des requêtes non redondantes et plus pertinentes. Les agents interviennent au niveau de la formulation des requêtes en proposant une aide à la construction, à l'extension et, éventuellement, à la reformulation. Pour cela, il doit être capable d'exploiter le profil de l'utilisateur. Après la phase de recherche, les procédures de filtrage et de classification permettront d'exploiter au mieux les résultats de la requête.

Les choix faits dans ce projet permettent de mettre en place un mécanisme de requêtes persistantes : lorsque, après interrogation, un utilisateur aura trouvé une page ou un serveur Web ou même un thème qui l'intéresse, il pourra demander à être informé en temps réel de toutes les modifications (mais aussi ajouts de pages) correspondant à sa requête. Les mécanismes de persistance actuels, bien moins élaborés, permettent seulement la surveillance d'une page qui doit déjà être identifiée.

L'objectif de ce papier est de présenter l'organisation de la société BONOM et, plus particulièrement, la formalisation de la connaissance et son exploitation pour répondre le plus justement à une requête. La prochaine section présentera le projet BONOM plus en détail. Ensuite, nous présenterons l'organisation et la gestion des connaissances dans la société BONOM (section 3). Nous présenterons ensuite l'exploitation de cette connaissance pour la création et le fonctionnement des agents de la société : en particulier pour les agents sites (section 4). Nous donnerons aussi les principes pour l'exploitation des connaissances pour les agents reformulateurs (section 5) et les agents personnels (section 6). Finalement, avant de conclure nous donnerons l'état d'avancement du projet en ce qui concerne l'implémentation.

2 Le projet BONOM

2.1 Communautés de la société BONOM

BONOM est une *société d'agents collaboratifs* organisée en *système fédéré* [3,6] : c'est une structure hiérarchique où des agents dits « facilitateurs » ou « médiateurs » dirigent les requêtes selon leur nature depuis les agents personnels (l'utilisateur) vers les agents sites (les sites web).

« communauté simple ») possède plusieurs facilitateurs dont les rôles sont identiques (figure 1).

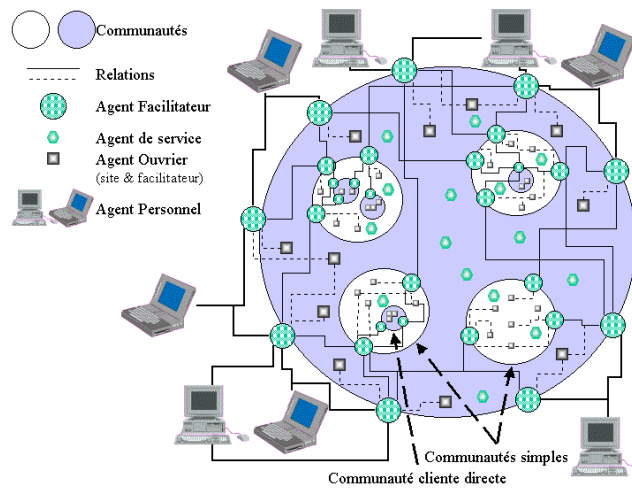


Figure 1. Une société BONOM

De façon à diriger les requêtes de recherche d'informations le plus rapidement et le plus pertinemment vers les sites concernés, la société est donc organisée en différentes *communautés simples*, éventuellement imbriquées, qui sont spécialistes d'un domaine informationnel particulier appelé *thème*. Les agents d'une communauté simple ne traiteront que les requêtes qui concernent leur domaine, les autres seront proposées aux communautés imbriquées. La notion qui relie ici les agents d'une même communauté simple est donc celle de domaine informationnel. Dans un premier temps, nous avons choisi de considérer une structure de domaines particulière qui ressemble à celles des « newsgroups » actuels. Ce choix a été fait, car ce système est connu et fonctionne plutôt bien. Un domaine informationnel sera donc une succession de thèmes, chacun indiquant une spécialisation du thème qui précède, comme par exemple « /recherche/informatique/système/linux ».

Notons que nous ne nous intéressons pas pour l'instant aux sociétés auto-organisatrices qui décideraient elles-mêmes s'il convient de rajouter un domaine informationnel. Nous considérons que comme dans les *newsgroup*, ce sont les utilisateurs et les responsables de la société qui décident. Cependant, nous nous proposons de fournir des outils de haut niveau pour aider les intervenants humains dans la gestion de la société (organisation des agents, gestion de la connaissance...).

Une *communauté* de la société BONOM (au sens large) est une communauté simple et l'ensemble des communautés qui lui sont rattachées appelées *communautés clientes*, car elles vont y rechercher des requêtes. Les communautés simples qui sont directement attachées à une autre communauté simple sont appelées *communautés clientes directes*. Une communauté simple est attachée à un thème. Les communautés simples (et leur thème) sont organisées sous forme d'une arborescence.

Dans la société BONOM, la connaissance est présente à tous les niveaux de la structure. Contrairement aux systèmes existant où elle ne se trouve qu'au niveau de l'utilisateur et au niveau des sites. Les requêtes peuvent ainsi être dirigées avec beaucoup plus de justesse.

2.2 Les agents de la société BONOM

Les agents de la société BONOM sont des *agents d'information intelligents* [3,4,5,7], car en plus de posséder les caractéristiques des agents intelligents (autonomes, capables de communiquer, réactifs et ayant des capacités d'apprentissage, de raisonnement, de planification...), ils sont aptes à gérer des requêtes d'information. Nous distinguons quatre classes d'agents dans la société :

- Les *agents personnels* qui assistent l'utilisateur dans la formulation des requêtes et dans la prise de connaissance des réponses ;
- Les *agents ouvriers*, qui vont effectuer un certain travail sur les requêtes de recherche d'information ;
- Les *agents facilitateurs* qui effectuent un travail de routage des requêtes et de concentrateurs de réponses (d'où leur nom de « BaGate », concentré de « Bag » et « Gates ») ;
- Les *agents de service* qui proposent des traitements spécifiques mais qui n'interviennent pas directement dans le traitement d'une requête.

d'un utilisateur. Les créations et les gestions de ces agents ne sont pas centralisées. Le choix des outils mis en oeuvre reste local et leur programmation est donc peu contrainte. Seuls le langage de communication et quelques fonctionnalités élémentaires sont imposées pour qu'il puisse y avoir coopération.

2.2.1 Les agents personnels

L'agent personnel, que l'on peut aussi appeler *agent utilisateur* [3], est responsable de la gestion des requêtes d'un utilisateur (reformulation, extension...) et des réponses qu'il obtient (filtrage, classification...). En prenant en compte le profil de l'utilisateur, il est apte à mettre en place des requêtes ciblées et pertinentes ainsi que d'effectuer des présentations de résultats adaptés. Il propose les requêtes à la société BONOM soit par l'intermédiaire des BaGates d'entrée dans la société soit aux BaGates d'une communauté particulière (thème donné) soit directement à un agent site.

2.2.2 Les agents ouvriers

Parmi les ouvriers se trouvent, entre autres, les *agents sites* et les *agents reformulateurs*. Un agent ouvrier est rattaché à une et une seule communauté simple, c'est-à-dire à un thème donné et donc à une catégorie de connaissance particulière. Ces agents sont en relation avec un ou plusieurs facilitateurs d'une communauté où ils vont chercher des requêtes à traiter.

L'agent site (ou *agent de données* [3]) est responsable des informations de son site et, en particulier, de l'exploitation et de la structuration des informations (indexation, utilisation d'ontologies et de méta-connaissance...). Il garantit que les informations qu'il gère sont régulièrement mises à jour et sont cohérentes par rapport à sa place dans la société. De plus, il a une connaissance avancée de leur organisation. Il peut ainsi répondre avec justesse à des requêtes très précises. Les agents personnels et les agents sites sont reliés par un réseau d'agents facilitateurs.

Les agents reformulateurs (ou *agents de requêtes* [3]) sont chargés de manipuler les requêtes par des reformulations, des généralisations, des spécialisations... Ils sont aussi capables de construire des requêtes portant sur des connaissances « proches » de celles auxquelles fait référence la requête (par l'utilisation des ontologies et des relations entre concepts...).

2.2.3 Les agents facilitateurs

La société d'agents facilitateurs assure les échanges entre les différents types d'agents, en les mettant en communication de manière pertinente. Un agent facilitateur est dédié à un domaine. Il reçoit une requête par un agent personnel ou par d'autres agents facilitateurs. Il diffuse la requête à d'autres agents facilitateurs encore plus spécialisés ou à des agents sites. Il gère également le retour des résultats et une partie de leur filtrage. Les agents facilitateurs sont donc attachés à une communauté dont ils sont les seuls moyens d'accès. Une communauté possède généralement plusieurs agents facilitateurs qui se connaissent mutuellement¹.

2.2.4 Les agents de service

Enfin, la société BONOM comporte un certain nombre d'agents dits « de service » n'intervenant pas directement dans la gestion d'une requête de l'utilisateur mais permettant la construction, la maintenance et la cohérence de la société ainsi que de la connaissance qu'elle détient. Parmi ces agents, nous trouvons les agents fournisseurs d'ontologies, les agents fournisseur de thème, les agents indexeurs... Plusieurs de ces agents seront décrits dans les sections suivantes.

2.3 Les requêtes

Comme nous l'avons vu, une société BONOM est un système fédéré. Les messages circulant sur ce réseau d'agents sont « routés » par les agents facilitateurs. Ceux-ci s'envoient mutuellement des messages (requêtes d'un utilisateur ou réponses). C'est donc un *système par passage indirect de messages* [3] entre les agents personnels et les agents sites. Il y a trois classes de messages dans la société BONOM : les *requêtes d'architecture*, les *requêtes d'information* (parfois appelées *requêtes de données*) et les *réponses*.

L'ajout et la suppression d'un agent font l'objet de messages spécifiques dans la société BONOM : les requêtes d'architecture. Ces requêtes garantissent l'insertion de l'agent dans la communauté

¹ Ici, nous considérons une société bien formée où chaque communauté simple est associée à un thème différent et où tous les agents facilitateurs de cette communauté se connaissent. Seulement, au cours de la « vie » de la société et des événements techniques ou humains, cette cohérence n'est pas toujours maintenue. Aussi, il existe tout un protocole entre les agents facilitateurs permettant de réunir les communautés de même thème et de garantir qu'ils se connaissent tous [1]. De plus, toute l'organisation de la société BONOM est décrite par

tous les facilitateurs d'un thème soient bien dans la même communauté simple et qu'ils se connaissent entre eux.

Les requêtes d'information sont les requêtes des utilisateurs (éventuellement reformulées). Si une requête arrive dans une communauté simple (c'est-à-dire sur un agent facilitateur de cette communauté simple) et qu'elle ne peut être dirigée convenablement, elle « tourne » alors parmi les facilitateurs de la communauté simple jusqu'à trouver celui lui permettant d'avancer. Les requêtes des utilisateurs sont des combinaisons de termes (simples ou complexes, précis ou flous...) connecté par des opérateurs logiques. Ces requêtes acceptent aussi des quantificateurs logiques et des qualificatifs. Il est ainsi possible de proposer des requêtes généralistes ou encyclopédiques. En réalité, pour éviter un trop grand nombre de connexion réseau, les requêtes ayant un destinataire commun sont regroupées.

En ce qui concerne les réponses des agents sites, elles sont remontées par les facilitateurs qui, périodiquement, les regroupent en paquets en effectuant aussi un premier filtrage (doublons...).

Le langage supportant les requêtes de l'utilisateur, les réponses des agents est basé sur la norme XML [8] et peut s'apparenter aux langages de communication entre agents tel que KQML[9] ou le FIPA ACL [10].

3 Les connaissances dans le projet BONOM

3.1 Qu'est-ce qu'une ontologie ?

Les ontologies sont devenues ces dernières années un domaine à part entière de la recherche en ingénierie des connaissances (conféré le nombre de conférences qui y font références).

Ce terme est issu du domaine de la philosophie et a comme signification originelle : « explications systématiques de l'existence ». Les chercheurs en ingénierie des connaissances ont, sans fondamentalement en changer le sens, donné d'autres définitions de ce concept, mieux adaptées à leur souci de recherche. En parcourant la littérature, on peut s'apercevoir qu'il y a foison de définitions complémentaires qui sont fortement liées au point de vue de l'auteur et à l'utilisation qu'il fait des ontologies [11] [12]... Certains ont un point de vue plus formel et s'attachent à travailler sur un modèle abstrait des ontologies, d'autres au contraire, utilisent des ontologies dans des applications bien précises et en ont une vue plus pragmatique.

En ce qui nous concerne, nous prendrons comme définition, celle issue de [13] : « les ontologies fournissent le vocabulaire commun d'un domaine et définissent de façon plus ou moins formelle le sens des termes et les relations entre ces derniers ». Ceci nous autorisera à pouvoir appeler *ontologies*, des *hiérarchies de concepts* plus ou moins formalisées.

Du point de vue de la représentation des connaissances, les ontologies sont décrites en mettant en jeu la suite de composants suivants [11] :

- les concepts ;
- les relations entre concepts (générique / spécifique, partie / tout...) ;
- les fonctions qui sont des relations particulières ;
- les axiomes ;
- les instances (qui réifient le modèle).

Il est à noter que les ontologies ne comprennent pas toutes ces cinq composants comme, par exemple, Wordnet [14] qui est un thesaurus issu du monde de la recherche en langage naturel.

Du point de vue des langages de description d'ontologie, divers types de langages existent, certains formels comme KIF [15], FLOGIC [16], LOOM [17] et d'autres plus proches du langage naturel comme le langage de description des termes, de leur sens et de leur relation dans le thesaurus Wordnet.

3.2 Ontologies et recherche d'information sur Internet

A l'heure actuelle, les projets de recherche d'informations sur Internet utilisent la notion d'ontologies pour aider le processus de recherche de document, mais surtout au niveau de la reformulation de la requête. Par exemple, le robot de recherche IICA développé par un groupe de chercheurs japonais [18] est un système de navigation sur Internet basé sur les ontologies. Ces dernières étant utilisées pour déterminer des similarités entre les pages candidates issus du Web et les mots clefs donnés par l'utilisateur ainsi que pour catégoriser les pages des sites Web. Dans cette application, les ontologies sont de simples taxonomies construites à partir de thesaurus (type Wordnet) et de documents techniques.

Par ailleurs, une importante communauté de chercheurs s'intéressent à l'annotation manuelle de page Web à l'aide d'ontologies afin d'apporter de la sémantique à ces dernières. Nous

travaillent à la constitution d'une ontologie qui modélise la communauté acquisition des connaissances et qui est ensuite utilisée pour annoter manuellement les pages Web de cette communauté afin que le moteur de recherche Ontobroker [21] puisse faire une recherche intelligente de documents via ces pages annotées.

3.3 Agents BONOM et ontologies

Au sein de la société BONOM, les ontologies permettent de formaliser la connaissance de cette société non seulement au niveau de l'utilisateur et des serveurs Web mais aussi au niveau des agents reformulateurs. Ainsi, le routage des requêtes et des réponses sera optimisé. Les ontologies sont donc utilisées à divers niveaux de l'architecture :

- Au niveau des agents sites (un site physique pouvant supporter plusieurs agents sites spécialistes d'un domaine donné). A ce niveau, les ontologies, sélectionnées par le gestionnaire du site (humain), participent au processus d'indexation du site afin de dégager au mieux des représentants du contenu du site. De plus, ces connaissances permettent de répondre de manière plus judicieuse aux requêtes.
- Au niveau des agents reformulateurs qui sont chargés de transformer la requête d'information qu'ils reçoivent (de l'étendre, de la restreindre...) afin qu'elle colle au mieux à la demande de l'utilisateur de la société.
- Au niveau des agents personnels qui sont l'interface entre l'utilisateur et la société BONOM. Les ontologies permettent à un utilisateur de formuler sa requête en fonction d'une ontologie qu'il aura éventuellement sélectionnée ainsi que d'effectuer un filtrage et une classification performante des réponses.

Dans la société BONOM, la connaissance est donc formalisée à l'aide d'ontologies permettant ainsi des réponses exhaustives aux requêtes du point de vue des pages proposées mais aussi des connaissances (des concepts) concernées. Cette connaissance est mise à disposition des agents de la société via un *agent fournisseur d'ontologie*. Cet agent de service s'appuie sur une base de données d'ontologies sélectionnées par des intervenants humains qui en assurent la cohérence et la couverture. Notons aussi, qu'il pourrait être fait appel à des serveurs proposant des ontologies standardisées et attestées (comme le serveur Ontolingua [22])...

3.4 Structure en thèmes et ontologie

Au niveau de l'organisation thématique, une société BONOM est organisée de manière hiérarchique selon une arborescence de thèmes (similaire à celui des *newsgroup*). A un thème donné correspond une communauté simple de la société BONOM. La connaissance associée à cette structuration est de deux types :

- une connaissance relative à la hiérarchie de thèmes qui est une ontologie simplifiée dont les concepts sont les thèmes ;
- une connaissance relative aux thèmes.

La connaissance relative à un thème est un ensemble d'ontologies regroupant (figure 2) :

- les ontologies des agents site de la communauté simple associées au thème. Ce sont les ontologies qui sont utilisées pour l'indexation de ces sites ;
- les ontologies représentatives des connaissances des communautés liées au thème.

Une ontologie représentative est construite en prenant en compte toute la connaissance propre à une communauté (elle représente une synthèse de la connaissance de la communauté), c'est à dire l'ensemble des ontologies des agents site de la communauté simple et des communautés simples clientes directes. A l'heure actuelle, dans la société BONOM, c'est un intervenant humain qui est chargé pour chaque thème de la construction de l'ontologie représentative. Il est donc le garant de sa validité et de sa cohérence par rapport aux connaissances du thème. Ces ontologies seront utilisées, par exemple, par les agents reformulateurs afin de modifier la requête et la cibler vers une communauté plus spécialisée.

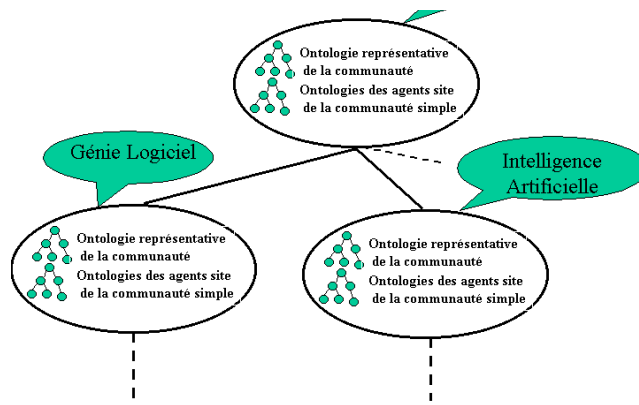


Figure 2. La hiérarchie de thèmes

Les connaissances que nous venons de présenter, celles relatives à l'arborescence de thèmes et celles plus spécifiques aux thèmes, sont disponibles dans la société BONOM via un agent de service appelé *fournisseur de thèmes*. Cette connaissance est maintenue, comme pour l'agent fournisseur d'ontologies, par un intervenant humain pour en assurer la cohérence. Dans les sections suivantes, nous présentons comment les connaissances sont exploitées principalement au sein des agents site mais aussi des agents reformulateur et des agents personnel.

4 Exploitation des connaissances par l'agent site

4.1 Principe

Pour insérer son site dans la société BONOM, le gestionnaire doit déterminer le (ou les) communauté(s) susceptible(s) d'accueillir son site. Dans ce but, un *agent indexeur* l'aide à déterminer l'ensemble des ontologies significatives par rapport au contenu de son site. A partir de cet ensemble d'ontologies, le gestionnaire de site génère des agents sites qui seront chacun spécialisé dans un (des) thème(s) donné(s) et donc pour une partie des ontologies significatives déterminées précédemment. Ensuite, chaque agent site est inséré dans la société sur le(s) thème(s) faisant référence à une ontologie proche des ontologies dont il dispose.

4.2 Détermination des ontologies significatives

La création des agents site et la sélection des ontologies est supervisée par le gestionnaire du site. En effet, connaissant au mieux la sémantique des informations proposées par son site, nous considérons qu'il est le mieux placé pour déterminer les caractéristiques qu'il désire mettre en évidence. Par conséquent, la détermination des ontologies significatives du contenu du site s'effectue par un échange entre le gestionnaire du site et les agents fournisseur d'ontologie et fournisseur de thèmes.

4.2.1 Initialisation (figure 3)

A partir de sa connaissance du site, le gestionnaire propose à l'agent fournisseur de thème, via l'agent fournisseur d'ontologie, un ensemble de concepts C_i avec $i \in [1, n]$ qui lui semblent significatifs du contenu de son site. L'agent fournisseur d'ontologie détermine alors un ensemble de couples $\langle \{O_j / j \in [1, m_i]\}, C_i \rangle$ avec $i \in [1, n]$ qui à un concept donné C_i associe l'ensemble des ontologies qui y font référence. Cet ensemble est déterminé en s'appuyant sur la structure actuelle de la société (ontologies utilisées) en sélectionnant celles référençant un ou plusieurs concepts.

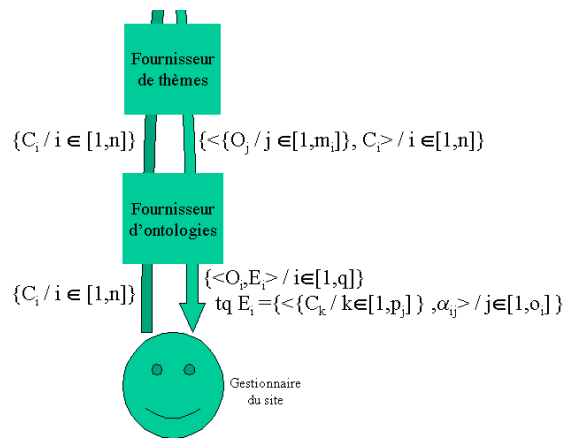


Figure 3. Initialisation

Ensuite, à partir de l'ensemble précédent, l'agent fournisseur d'ontologie (qui éventuellement rajoute des ontologies proches ou en relation avec celles déterminées précédemment), quantifie la force de la relation qui unit les concepts à chaque ontologie. Dans ce but, pour une ontologie donnée O_i , il détermine des ensembles de concepts qui lui sont le plus fortement liés (en utilisant un principe similaire à l'appariement des index avec les concepts de la section 4.2.3) ainsi qu'un coefficient α_{ij} qui mesure pour chaque ensemble précédant, le degré de représentativité de l'ensemble des concepts par l'ontologie concernée, c'est-à-dire l'ensemble suivant :

$\{ \langle O_i, E_i \rangle \} / i \in [1, q]$ tels que : $E_i = \{ \langle \{ C_k / k \in [1, p_j] \}, \alpha_{ij} \rangle$ et $O_j / j \in [1, o_i]$

Actuellement, nous établissons la valeur de ce coefficient α_{ij} , en effectuant une moyenne pondérée relative aux concepts directement présents dans l'ontologie, à ceux fortement liés à des concepts de l'ontologie et à ceux qui le sont faiblement.

4.2.2 Processus de choix des ontologies

Le gestionnaire du site sélectionne les ontologies qui lui semblent les plus pertinentes, en ajoutant éventuellement à cet ensemble de nouvelles ontologies qui ne sont pas encore utilisées dans la société BONOM.

Toute nouvelle ontologie proposée doit être validée via l'agent fournisseur d'ontologie. En effet, dans un souci de cohérence, toute nouvelle ontologie doit être ajoutée selon un format standard, mise en relation avec les autres ontologies (et éventuellement avec les thèmes par l'intermédiaire du fournisseur de thèmes) et validée par intervention humaine.

Une fois les ontologies validées, l'agent indexeur effectue une indexation du site (en fonction de ces ontologies) et une instanciation des ontologies. Ces résultats sont ensuite présentés au gestionnaire de site. S'ils ne lui conviennent pas, il peut réitérer le processus précédent (figure 3), en redéfinissant de nouveaux concepts (par exemple mis en évidence par l'indexation...) ou en proposant de nouvelles ontologies qui lui semblent significatives.

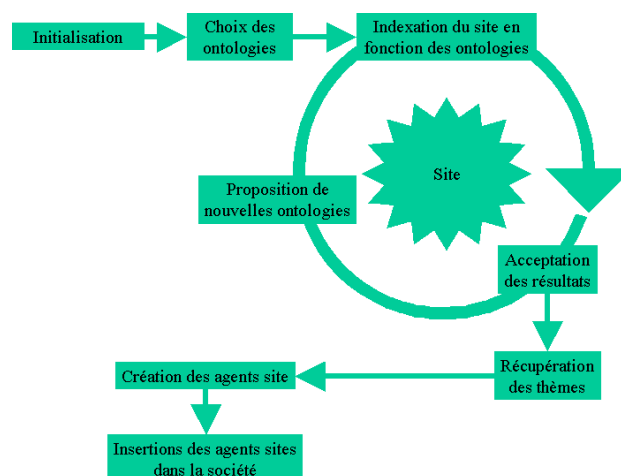


Figure 4. Processus de création des agents site

4.2.3 Indexation du site

A ce niveau, le gestionnaire de site a sélectionné un ensemble d'ontologies qui lui paraissent refléter le contenu de son site. L'agent indexeur a la charge de l'indexation des pages du site en prenant en compte cet ensemble d'ontologie. Sa tâche est donc d'instancier les concepts des

textes ont une place prépondérante dans leur compréhension. De plus les pages web, comportent des informations de type typographique (les marqueurs HTML) qui peuvent appuyer très fortement la recherche de termes précédents. En prenant en compte ces deux principes, le processus d'indexation commence par l'acquisition des termes des pages, en s'appuyant sur des méthodes d'extraction terminologique qui prennent en compte les marqueurs HTML. Ensuite nous déterminons les concepts relatifs aux termes sélectionnés. Afin d'effectuer le passage du niveau linguistique (termes pondérés) au niveau conceptuel, nous utilisons des ontologies de type Thésaurus. En effet, dans les thesaurus comme Wordnet, qui sont des thesaurus issus du monde de la recherche en langage naturel, chacune de leurs entrées est un terme qui est décrit par ses différents *synsets* (les différents ensemble de synonymes qui lui sont relatifs, donc ses différents sens), et pour chaque synset, ses hyperonymes, ses hyponymes, ses méronymes, etc. Enfin, le processus d'indexation se termine par l'appariement des concepts retenus pour les pages avec des concepts issus des ontologies sélectionnées précédemment par le gestionnaire de site.

Le processus d'indexation est donc découpé en cinq phases (figure 5) :

1. **collecte des pages du site et pour chaque page** [23] :

- a) Détermination de la catégorie grammaticale des mots à l'aide d'un étiqueteur grammatical [24],
- b) Lemmatisation, utilisation de patrons syntaxiques [25] et exploitation des marqueurs HTML pour récupérer les termes.

2. **détermination des index.** Cette phase consiste à déterminer les concepts caractéristiques à partir des termes précédents, sachant qu'à un terme donné peuvent correspondre plusieurs concepts différents (les différents « synset » de Wordnet). L'appariement entre concept et terme d'une page s'effectue selon le contexte en recherchant, pour un terme donné et pour un sens donné, le nombre de mots du texte qui sont des synonymes, des hyperonymes de niveau 1 de ce terme... Le sens sélectionné, donc le concept retenu est celui qui obtient le plus fort poids lors de cette recherche. Chaque concept c_j significatif engendre un index ind_j composé de l'ensemble des synonymes de ce concept ($synset_j$) et les pages (p_{ij}) dans lesquelles il se trouve. Donc, un index ind_j est de la forme : $\langle synset_j, \{p_{ij} / i \in [1, n_j]\} \rangle$. Pour indiquer l'importance relative d'un concept dans les pages dans lesquelles il est présent par rapport à tous les concepts référencés présents dans ces mêmes pages, nous calculons aussi le *coefficient de pertinence* d'un index (entre 0 et 1). Ce coefficient est calculé de la manière suivante :

$$C_p(ind_k) = \frac{\sum_{p=p_k}^{pnk} \sum_{s \in synset_k} \sum_{t \in p/t=s} cm_t}{\sum_{c \in \mathcal{C}} \sum_{p=p_k}^{pnk} \sum_{s \in synset_c} \sum_{t \in p/t=s} cm_t}$$

où cm_t est le poids du terme t en fonction du marquage HTML le concernant, \mathcal{C} l'ensemble des concepts indexés.

3. **appariement des index avec les concepts des ontologies.** Pour pouvoir faire l'appariement, il faut qu'obligatoirement tout concept Co_{ij} d'une ontologie O_i soit lié à au moins une forme lexicale qui est son représentant linguistique. Cette forme lexicale permet de sélectionner dans le thésaurus (le même qui a été utilisé lors de l'extraction de termes) l'ensemble des synsets candidats $S(Co_{ij})$. Parmi ceux-ci sont retenus uniquement un sous-ensemble $S'(Co_{ij})$, qui contient les synsets apparaissant comme synset d'un index. La phase suivante consiste à déterminer quel est le(s) synset(s) (élément de $S'(Co_{ij})$) qui sera(ont) retenu(s) pour l'appariement. Cette sélection s'effectue en exploitant les relations du concept Co_{ij} avec les autres concepts de son ontologie ainsi que les relations des termes du synset avec les autres termes du thésaurus en prenant en compte les relations équivalentes. La phase de sélection calcule un *coefficient d'appariement* Ca_k (entre 0 et 1) permettant de sélectionner les plus probables. Deux cas de figure se présentent alors : aucun synset n'est retenu, le concept ne peut pas s'apparier ; un ou plusieurs sont retenus, alors tous les index correspondant sont appariés sauf si un index apparaît valide (coefficient proche de 1). Notons que ceux dont le coefficient est proche de 0 sont supprimés.

4. **amélioration des appariements.** Afin d'éviter au maximum les ambiguïtés sur les concepts des ontologies (un même index apparié avec deux concepts différents), tout index ayant un fort coefficient d'appariement (proche de 1) avec un certain concept, est retiré des concepts avec qui il a été aussi apparié. Un concept donné est donc caractérisé par un ensemble de couple qui pour chaque synset possible donne son coefficient de pertinence par rapport aux pages et son coefficient d'appariement avec le concept.

5. **caractérisation des ontologies.** Pour chaque ontologie est ensuite calculé un ensemble de

- de mettre en évidence le degré de couverture de cette ontologie,
- d) la mesure relative au nombre de concepts de l'ontologie référencés dans les pages afin de caractériser le pouvoir représentatif de cette ontologie.

Remarque : l'indexation peut mettre en évidence des index dont les concepts s'apparient mal ou pas du tout aux concepts des ontologies proposées. Ces concepts sont alors proposés au gestionnaire du site qui, s'il le désire, peut demander une recherche d'ontologies les concernant.

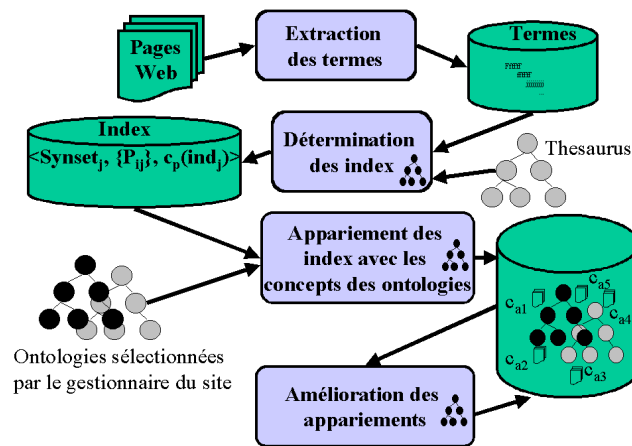


Figure 5 : Processus d'indexation d'un site

4.3 Insertion dans une communauté

A ce niveau, le gestionnaire de site dispose d'un ensemble d'ontologies (valide dans le cadre de la société BONOM) qui lui semblent pertinentes pour son site. De plus, le site est indexé selon ces ontologies.

Le gestionnaire de site envoie au fournisseur de thème l'ensemble des ontologies O_i (et les résultats des coefficients calculés après indexation) qu'il a sélectionné. Ce dernier propose alors un ensemble de couples $\langle Thème_j, \{O_{kj}\} \rangle$ possibles.

Le gestionnaire du site demande alors la création d'autant d'agents sites que de couples qui lui conviennent. Ces agents sont insérés dans les communautés simples associées aux thèmes visés.

Par ailleurs, le gestionnaire du site peut demander la création de nouveaux thèmes. Le bien fondé de cette demande est validé par un intervenant humain, tel que le gestionnaire de la société BONOM (par exemple par l'organisation d'un vote comme dans les newsgroup). En cas de création d'un nouveau thème, les ontologies le concernant, ainsi que le thème lui-même, sont mis à jour dans les connaissances des agents fournisseurs d'ontologie et fournisseurs de thème.

4.4 Fonctionnement d'un agent site

Un agent site est spécialisé pour un site Web donné en fonction des ontologies qu'il possède. De plus, le site est indexé selon ces ontologies. Aussi, lorsque cet agent reçoit une requête (construite, elle aussi, à l'aide de ces mêmes ontologies) concernant son domaine de connaissance, il est capable de trouver des informations susceptibles d'y répondre. Ces réponses sont les adresses des pages concernées et/ou des informations mises en forme issues de ces pages.

Un site Web donné comporte donc plusieurs agents sites. Chaque agent site caractérise une connaissance particulière du site. Cette connaissance est définie par une ou plusieurs ontologies. Par conséquent, les agents sites représentent un ensemble de points de vue sur le site. Ces agents exploitent cette connaissance et l'indexation pour répondre aux différentes requêtes. Elles peuvent être de différentes natures. En particulier, nous avons mis en évidence deux types de requêtes :

- les requêtes « généralistes » où l'utilisateur demande des informations générales sur un concept en omettant les détails trop précis et toutes les notions s'y rattachant plus ou moins directement ;
- les requêtes « encyclopédiques » où l'utilisateur demande des informations relatives à tout ce qui s'approche de près ou de loin au concept.

Il se trouve que notre choix d'indexation permet de différencier ces deux catégories de requêtes. Dans le cadre des requêtes généralistes concernant un concept, l'agent site sélectionne les pages référencées par les index associés à ce concept ainsi que ceux qui lui sont proches.

tous ses sous-concepts.

De manière générale, les réponses sont constituées : d'un ensemble de pages et de concepts qui leurs sont associés ainsi que, pour chacun de ces concepts, les coefficients de pertinence C_p et d'appariement C_a relatifs aux sous-ensembles de pages qui les concernent. De plus, ces réponses font référence à l'ontologie utilisée pour sélectionner ces pages. Ces coefficients sont éventuellement réajustés en fonction de l'avis de l'utilisateur par rapport aux pages fournies.

4.5 Exploitation des connaissances par l'agent reformulateur

Un agent reformulateur permet de créer d'autres requêtes d'information plus ciblées en fonction des éléments d'une requête initiale. Cette opération s'effectue aussi bien en modifiant ou en spécifiant la communauté cible de la requête initiale qu'en reformulant la requête elle-même. Au retour, les réponses aux requêtes d'information, qui comportent les coefficients de pertinence et d'appariement, permettent au reformulateur d'améliorer son traitement.

La création d'un agent reformulateur est généralement proposée par un responsable de sites ou tout autre acteur humain de la société. Cependant, il est possible de générer automatiquement un agent reformulateur au fur et à mesure de la création des thèmes. Le principe en est le suivant : chaque thème place un reformulateur dans la communauté pour laquelle elle est cliente directe. Ce reformulateur se base sur la connaissance de la communauté pour laquelle il est créé, c'est-à-dire sur l'ontologie représentative de la communauté (figure 6).

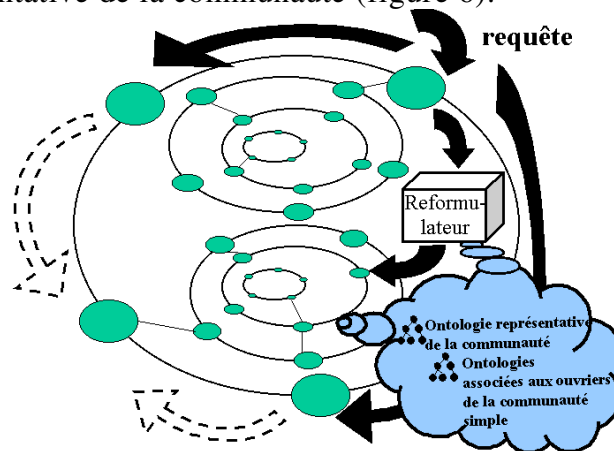


Figure 6. Rôle de l'agent reformulateur

Lorsque le reformulateur traite une requête à l'aide des ontologies représentatives et des relations avec d'autres ontologies, il la reformule afin de la diriger vers la communauté qu'il représente.

Remarque : les ontologies représentatives sont caractérisées par des coefficients sur les concepts permettant d'indiquer leurs fréquences d'utilisation dans la communauté. Ces coefficients sont la moyenne de tous les coefficients de pertinence des index appariés avec ces concepts. Ainsi, les coefficients des concepts de l'ontologie représentative de la société BONOM (le niveau zéro de la hiérarchie) fournissent une photographie de l'état réel des connaissances à un instant donné dans la société.

5 Exploitation des connaissances par l'agent personnel

L'agent personnel assiste l'utilisateur dans sa recherche d'information dans la société. Il l'aide à composer sa requête, s'occupe de la cibler (et éventuellement de la reformuler) en fonction de son profil. De plus, il effectue un dernier filtrage des résultats et s'attache à les présenter de manière adaptée (graphiques, arbres...) en effectuant une classification plus fine. Pour effectuer ces tâches, il dispose de la connaissance de la société BONOM formalisée à l'aide d'ontologies.

Lors de la formulation de la requête, l'agent personnel dispose d'abord des connaissances relatives au profil de l'utilisateur (domaines de recherches habituelles, profession, préférences déclarées...). Ces connaissances font éventuellement référence à des ontologies utilisées dans la société BONOM. Ainsi, l'agent personnel peut aider l'utilisateur à formuler la requête à partir de ces ontologies, lui proposer d'autres requêtes apparentées grâce aux relations entre les concepts des ontologies et, éventuellement, produire des requêtes apparentées sans que l'utilisateur n'intervienne. Les requêtes sont produites en leur associant les ontologies qui les caractérisent (éventuellement après échanges avec l'agent fournisseur d'ontologies et l'utilisateur) permettant

répondre. Ces derniers construisent alors leurs réponses (pages HTML, données mises en forme...) en leur associant éventuellement les ontologies utilisées pour construire la réponse, ainsi qu'une évaluation de la réponse (basées sur les résultats de l'indexation) précisant sa pertinence par rapport à la requête. Ces renseignements seront utiles à l'agent personnel pour mettre en forme les réponses, et les classer de manière cohérente, non redondante et judicieuse, afin que l'utilisateur en prenne connaissance.

6 Implémentation

Actuellement, un premier prototype de la société BONOM a été implémenté. La société se déploie correctement sur différentes plate-formes (Linux et Windows) et des requêtes d'information simples sont gérées. Ces requêtes sont construites à l'aide de la norme XML [8]. Les agents sont implémentés en Java 1.2 et en Python. Les communications entre agents de la société sont basées sur le bus CORBA [26] « ORBacus » (OOC [27]), ce qui permet une interopérabilité entre des agents qui peuvent être programmés dans des langages différents et supportés par des plate-formes différentes.

En ce qui concerne les agents indexeurs, des expérimentations ont été faites à l'aide du thesaurus Wordnet et des ontologies que nous avons construites concernant les connaissances relatives aux sites internet des universités. Wordnet est un thesaurus à caractère généraliste, mais nous aurions pu tout aussi bien utiliser des thesaurus à caractère plus spécialisé mais ceux-ci ne sont pas si facilement disponibles. Tout agent indexeur a à sa disposition tout un arsenal d'outil à caractères linguistique : lemmatiseur, étiqueteur, analyseur de marqueurs HTML qui sont utilisés pour extraire les termes. A l'aide du thesaurus Wordnet les index sont ensuite déterminés et l'appariement des index avec les concepts des ontologies est réalisé.

Actuellement, les ontologies sont déterminées par le gestionnaire du site. Les agents reformulateurs et les agents personnels sont en cours de développement.

7 Conclusion

Le projet BONOM est une société d'agent d'information dont les connaissances sont formalisées à l'aide d'ontologies. Une particularité essentielle est que cette connaissance n'est pas centralisée mais est totalement répartie dans la société. En particulier, les agents site responsables des réponses aux requêtes sont des spécialistes d'un domaine de connaissance. Ces agents sites sont construits après détermination de la connaissance spécifique du site. Chacun d'eux ayant la responsabilité d'un point de vue sur le site. L'élément essentiel pour la détermination de ces connaissances est une indexation basée non pas sur de simples critères statistiques mais sur l'exploitation de connaissances linguistiques et de connaissances a priori sur le contenu du site (ontologie).

Dans l'avenir, nous allons mettre en place les agents reformulateurs et les agents personnels qui vont eux aussi exploiter des ontologies. Nous allons aussi affiner la détermination des coefficients de pertinence et des coefficients d'appariement en réalisant de plus amples expérimentations.

Pour l'instant, du fait de l'utilisation du thesaurus anglophone Wordnet, nous ne pouvons travailler que sur des sites en anglais. Cependant, nous prévoyons prochainement de travailler sur des sites en langue française. Notons qu'un des objectifs du projet BONOM est aussi de supporter des requêtes multilingues.

8 Références

- [1] S. Cazalens, E. Desmontils, C. Jacquin et P. Lamarre. BONOM : une société d'agents répartis pour la recherche d'informations sur Internet. Rapport de Recherche IRIN-???, janvier 2000.
- [2] H. Samier et V. Sandoval. *La recherche intelligente sur l'Internet et l'Intranet*. Seconde éditions. Hermès, 1999
- [3] D. S. HaverKamp and S. Gauch. Intelligent Information Agents: Review and Challenges for Distributed Information Sources. *Journal of the American Society for Information Science* 49(4), April 1998, pp. 304-311
- [4] M. Wooldridge and N. Jennings. Intelligent Agents: Theory and Practice. *Knowledge Engineering*, vol. 10, no. 2, June 1995.
- [5] J. Bradshaw. An Introduction to Software Agents. *Software Agents*. J. Bradshaw (Ed.) AAAI Press / MIT Press. 1997
- [6] M. R. Genesereth and S. P. Ketchpel. Software Agents. *Software Agents*. J. Bradshaw (Ed.) AAAI Press / MIT Press. 1997
- [7] H. S. Nwana. Software Agents: An Overview. *Knowledge Engineering Review*. 11(3). pp. 205-244. 1996
- [8] W3C. Extensible Markup Language (XML) 1.0. W3C Recommendation. Reference: REC-xml-19980210. 10 February 1998. <http://www.w3.org/TR/REC-XML>
- [9] T Finin, R. Fritzson, D. McKay, R McEntire . KQML as an agent communication langage , the proceedings of the third international conference on information and knowledge managment (CIKM'94), ACM Press, novembre 1994.
- [10] FIPA Agent Communication Language. <http://www.fipa.org/spec/fipa97/FIPA97.html>

- engineering. Stanford pp 57–63, 1997.
- [13] A Gomez–Perez. Développements récents en matière de conception, de maintenance et d'utilisation des ontologies. Actes du colloque Terminologie et intelligence artificielle de Nantes le 10–11 mai 1999, revue terminologies nouvelles pp 9–20
 - [14] G.A Miller. Wordnet: an Online Lexical Database. International journal of lexicography 3,4: pp 235–312, 1990.
 - [15] M.R. Genesereth and R.E. Fikes. Knowledge Interchange Format, version 3.0, reference manual. Technical Report, Logic–92–1, computer science department, Stanford University, 1992. <http://www.cs.umbc.edu/kse/>
 - [16] M.Kifer, G Lausen and J Wu. Logical Foundations of Object–Oriented and Frame–Based Languages. Journal of the ACM, 1995
 - [17] R MacGregor. Inside the LOOM Classifier. SIGART bulletin, 2 (3): pp 70–76, 1991.
 - [18] M. Iwazume, K Shirakami, K Hatadani, H Takeda, T Nishida. IICA: an Ontology–Based Internet Navigation System. Workshop on Internet based information systems, august 5, 1996, Portland, USA.
 - [19] V.R Benjamins, D. Fensel, A Gomez–Perez, S Decker, M. Erdmann, E. Motta, M Musen. Knowledge Annotation Initiative of the Knowledge Acquisition Community KA². Proceedings of the 11th Banff knowledge acquisition for knowledge–based system workshop, KAW'98, Banff, Canada, avril 1998, pp 18–23
 - [20] S Luke, L Spector, D Rager. Ontology–Based Knowledge Discovery on the World–Wide–Web. Proceedings of the workshop on internet–based information system, AAAI'96, Portland, Oregon, 1996.
 - [21] D. Fensel, S. Decker, M. Erdman, R. Studer. Ontobroker: the Very High Idea. Proceedings of the 11th international flairs conference, Sanibal, Island, 1998.
 - [22] A. Farquhar, R. Fikes and J. Rice. The Ontolingua Server: a Tool for Collaborative Ontology Construction. International Journal of Human–Computer Studies 46(May) pp. 707–728, 1997.
 - [23] J Gamet. Indexation de pages web, DEA de l'université de Nantes, juin 1998.
 - [24] E Brill., Some Advances in Transformation–Based Part of Speech Tagging. Proceedings of the Twelfth National Conference on Artificial Intelligence AAAI–94, 1994.
 - [25] B. Daille, Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques, Thèse, Paris 7, 1994
 - [26] Object Management Group. The Common Object Request Broker: Architecture and Specification. OMG TC Document 95–02–25, Revision 2.0, 1997.
 - [27] Object Oriented Concepts. ORBacus Home page. <http://www.ooc.com/ob>