

L'annotation comme support à la collaboration autour de documents : l'outil AnT&CoW

Lortal G.⁽¹⁾, Lewkowicz M.⁽¹⁾ et Todirascu-Courtier A.⁽²⁾

(1) Laboratoire ISTIT - Tech-CICO, Université de technologie de Troyes, 12 rue M.Curie, BP 2060, 10010 Troyes Cedex France

(2) EA 1339- Linguistique, Langues et Parole, Université Marc Bloch, 22 rue René Descartes 67084 Strasbourg Cedex France

Nous présentons un outil Web qui implémente le standard Annotea du W3C, destiné à la collaboration autour de documents au travers d'annotations dans des projets de conception. Il permet une interaction médiatisée entre les participants, les membres échangeant des commentaires sur la base de documents. Les projets se constituent aujourd'hui autour de documents représentatifs du domaine à propos desquels les membres du projet échangent. Les discussions prennent la forme d'annotations ancrées à différents fragments documentaires. Pour faciliter l'exploitation de ces annotations par les membres du projet, nous proposons de les indexer selon trois dimensions (domaine, organisation, argumentation), représentées par des ontologies semi-formelles créées semi-automatiquement grâce à des outils de T.A.L. traitant les documents du projet. Ces outils constituent également une aide à l'utilisateur pour indexer l'annotation par des termes-clés représentatifs de l'annotation dans chacune des dimensions.

Keywords: Annotation, Collaboration, Web Sémantique, Web Social, Traitement Automatique des Langues.

1 Introduction

Le document représente un support classique de partage de la connaissance. Les travaux en Gestion des Connaissances (GC) se sont souvent concentrés sur les informations contenues dans ce document et ont donné naissance à de nombreux outils et méthodes pour indexer, stocker, partager ou retrouver ces informations. Notre objectif est de nous concentrer sur la collaboration entre des utilisateurs autour d'un document. Dans une perspective de TCAO (Travail Collaboratif Assisté par Ordinateur), nous souhaitons proposer un outil de travail coopératif support aux interactions permettant la conception d'un document par un groupe, en conservant les traces de ce processus de conception.

Un des moyens de collaborer de manière asynchrone sur un document, c'est-à-dire de partager ses opinions ou de débattre d'une question, est d'annoter le document. Nous adoptons ce point de vue focalisé sur le document en tant que construction, sur la connaissance élaborée grâce aux argumentations et aux échanges produits par des utilisateurs autour d'un document. Nous souhaitons donc fournir un outil qui permettrait aux utilisateurs d'annoter des documents, de stocker ces annotations ainsi que de les retrouver afin de comprendre la logique de conception du document. Cet outil gèrera ainsi la connaissance en action, ou « knowing », par opposition au « knowledge », comme différencié par [COO99] et [PFE99], et s'insèrera dans le courant initié par [BAN96] de mémoire organisationnelle constructive (en complément à une mémoire organisationnelle passive ou active) où les outils ont pour but d'aider à une construction d'espaces communs d'information en prenant en compte l'aspect sociologique et psychologique des organisations.

Dans cet article, nous présentons tout d'abord notre positionnement théorique et le déroulement de nos recherches dans un contexte de travail de groupe en projet dans le domaine de la conception logicielle et de produit. Ensuite, nous définissons les spécifications de l'outil d'annotation nécessaire pour soutenir l'activité de ces groupes, puis décrivons les moyens mis en œuvre pour atteindre nos objectifs, comme l'utilisation d'outils de Traitement Automatique du Langage (TAL) ou de standards du Web sémantique (WS). Enfin, nous présentons l'architecture et l'interface de notre outil d'annotation.

2 Positionnement

2.1. Web Social et Web Sémantique :

Dans la perspective actuelle du Web Sémantique (WS), nombre d'outils et de modèles sont mis en place en vue de permettre l'annotation. Ces outils prennent en compte des annotations que nous qualifions de « computationnelles » au sens où elles sont insérées dans la représentation de la page qu'elles annotent afin d'améliorer sa description, dans un but d'optimisation des logiciels de recherche d'information (par exemple un moteur de recherche).

Nous nous intéressons ici d'avantage à une utilisation humaine d'annotations comme moyen de coopération pour un utilisateur. Cette orientation se retrouve dans un ensemble grandissant d'applications du Web visant à fournir des espaces de rencontre accroissant la conscience mutuelle entre les partenaires (mutual awareness) dans les interactions distantes (forum, chat, messagerie instantanées, etc.). L'annotation est alors un élément de construction d'échanges non structurés.

Nos travaux se situent dans une approche socio-sémantique du Web, telle que définie dans [ZAC04], et qui vient compléter la définition du Web cognitivement sémantique proposée par [CAU02]. Le Web Cognitivement Sémantique visait à intégrer dans les recherches et les pratiques de développement des applications du WS l'ensemble des activités de conception initiale des représentations, de maintenance au fil de l'eau et d'évaluation de la pertinence des résultats des requêtes.

Le Web socio-sémantique (W2S) vise à soutenir des activités de coopération dans lesquelles les interactions s'appuient également sur des informations ou des documents partagés par un collectif poursuivant, au moins pour un temps, des objectifs communs. Vis-à-vis de ces objectifs, le W2S doit contribuer à la construction d'une représentation structurée tant du domaine que du collectif.

Les espaces de coopération offerts par le W2S doivent donc fournir un ensemble de fonctionnalités à caractère communicationnel et documentaire proche de celles offertes par les applications de groupware associant communication, partage d'information et mise en relation plus ou moins automatisée des acteurs par le biais de workflow. Mais, à la différence de ces applications qui reposaient sur des environnements propriétaires, le W2S s'inscrit dans la philosophie d'ouverture du WS. Par ailleurs, conformément à notre vision de la coopération structurellement ouverte, les applications du W2S doivent permettre aux acteurs de remodeler, voire de construire, la structure des espaces de coopération dans lesquels se déroulent leurs interactions. Cette malléabilité nécessaire des espaces du W2S, de même que leur caractère potentiellement ouvert et distribué, les rend conforme à notre vision des systèmes d'information situés et distribués et contribue à expliquer pourquoi l'approche du WS, tout à la fois élargie dans la direction du Web Cognitivement Sémantique et du W2S, nous semble offrir de bons principes de conception pour notre application.

2.2 Soutenir l'activité d'annotation dans son ensemble : discours et indexation

Pour identifier les fonctionnalités d'un outil de travail coopératif soutenant l'activité d'annotation, il est nécessaire non seulement d'observer des terrains où l'annotation est médiatisée, mais aussi de comprendre la place de cette activité d'annotation. Par son histoire même, l'annotation (issue des techniques d'herméneutique) [LOR05] est un énoncé explicatif lié à un texte qui peut aussi être reconnue dans son évolution comme un exercice discursif public sous forme de « sentences », de « commentaires » voire de « sommes » (résumés de connaissances sur un sujet [LIB01]). Elle possède ici en plus de sa caractéristique relationnelle et textuelle, une caractéristique discursive, d'outil énonciatif permettant l'argumentation. L'annotation étant fortement liée au document, elle possède une forte caractéristique contextuelle, y compris dans son énonciation et sa position argumentative. Nous parlerons donc désormais d'annotation discursive, suivant la dichotomie discours/texte courante dans le champ de la linguistique où le discours est l'« inclusion d'un texte dans son contexte » [CHA02]. L'annotation discursive permet l'élaboration d'une interprétation partagée du document. Son contexte est formé notamment par le rôle de son auteur, son contenu sémantique, sa place dans le fil de discussion. Cette contextualisation est essentielle pour tracer la logique de conception d'un texte, d'une interprétation car elle est à l'origine de création de sens.

Dans notre problématique de conception d'outil collaboratif, il serait intéressant d'étudier plusieurs terrains. Cependant, cette activité est peu soutenue dans un cadre collaboratif. Il est donc difficile de trouver un terrain observable de l'utilisation d'annotation dans un cadre collaboratif. Adaptant les résultats d'une étude sur le billet de [LAB05], nous adoptons une définition de ces échanges médiatisés comme éléments de genre épistolaire de forme brève. Les échanges étudiés possèdent certaines caractéristiques du type : Brièveté, caractère informel, caractère informationnel, caractère séquentiel, caractère relationnel (accompagnement d'une pièce jointe). Cette

dernière caractéristique de mise en relation d'un texte avec un autre objet (textuel ou autre) est centrale dans notre étude.

En effet la TCAO est fortement centrée sur les documents. Le document est le médium permettant aux participants d'un projet collectif de se constituer une réalité temporaire partagée (temporarily shared social reality - TSSR) [ROM74] dans un cadre de travail médiatisé. L'annotation dans un cadre de travail coopératif, qu'elle ait un objectif de planification (un post-it informatif laissé sur un dossier et destiné à l'équipe suivante) ou argumentatif (annotation en marge exprimant une opinion) est toujours liée à un document partageant aussi cette caractéristique de dialogue épistolaire (adressé) de forme brève. Contrairement à un message de forme plus longue ou plus « complet » qui peut être compris seul, les annotations sont fondamentalement liées. Par cette caractéristique relationnelle, nous rapprochons donc ces fragments textuels que sont les annotations des éléments de ce type d'échanges épistolaires médiatisés que sont les mels ou les échanges d'un forum toujours liés à un/des fil(s) de discussion structurant(s). L'annotation en tant que fragment textuel lié à un document ou un objet, prend son sens par sa contextualisation, tout comme un « post » envoyé à un forum en contribution à un fil de discussion.

Le caractère relationnel et contextuel de l'annotation a amené le domaine informatique à utiliser l'appellation « annotation » pour l'insertion de balise dans du texte. L'annotation est utilisée comme méta-donnée pour favoriser le décodage d'information et l'interopérabilité des systèmes dans le WS. L'annotation est alors un type de balise référentielle, liant différents objets ou pointant vers un chemin dans un réseau. Conceptuellement, on peut considérer ces annotations comme des index au sens étymologique d'« indicateur ». La balise informatique, contrairement à un simple indicateur, appartient au document qu'elle vise et qu'elle modifie de l'intérieur ; elle est l'objet utilisé par les outils d'annotation automatique, d'étiquetage.

Fort de toutes ces caractéristiques dérivées de différents points de vue sur l'annotation (historique, échange épistolaire médiatisé de forme brève), nous basons notre recherche à la suite de [ZAC03], sur l'étude de deux terrains en conception coopérative médiatisée. Le travail en conception se base sur des échanges qui construisent non seulement une interprétation de documents centraux de l'activité (Cahier des charges, étude de besoin) mais qui mène à une conception de produit fini. Les équipes observées travaillent de deux manières principales, en synchrone (travail et échange en face à face au sein d'un groupe) et en asynchrone (travail distribué s'effectuant pour les différents acteurs du projet dans des lieux et temps différents). Pour soutenir les phases de travail asynchrones, les participants s'appuient sur l'utilisation de techniques de médiatisation des activités, avant tout médiatisation de la communication. Nous avons donc regroupé pour notre analyse les échanges par courrier électronique d'un projet de conception logicielle (projet MIAMM, corpus de 130 mels) et ceux d'un projet de conception produit, en mécanique (projet Air Campus, corpus de 150 mels). Sans détailler l'étude de ce corpus, il apparaît que les échanges se structurent en discussions thématiques (problème à résoudre, une interprétation à construire) autour de documents (texte, maquette, code). Ces terrains sont des lieux représentatifs de l'activité d'annotation que nous cherchons à soutenir.

Notre outil devant soutenir l'annotation selon ses différentes caractéristiques et ses différentes formes (discursive donc textuelle, indexante donc balise) dans un contexte coopératif, il doit se baser sur des techniques souples et adaptables aux traitements cognitif et informatique. Ainsi, nous proposons de nous fonder sur les standards d'annotation existants, et de les enrichir pour une indexation plus fine et une exploitation plus aisée de ces annotations. Le but de ce travail est de soumettre un ensemble de fonctionnalités pour le développement d'un outil d'aide à l'annotation et à l'indexation de ces annotations, selon un modèle multidimensionnel.

Afin d'aider l'utilisateur à indexer les annotations qu'il a créées, nous proposons une indexation semi-automatique, fondée sur l'utilisation de méthodes et d'outils de Traitement Automatique du Langage (TAL). Ces outils TAL seront d'une part utilisés pour créer et mettre à jour les ontologies, bases de l'indexation, et d'autre part, ils permettront d'aider l'utilisateur à indexer ses productions et à naviguer au sein de l'ensemble des annotations. Puisque cet outil doit être pratique et mis à jour, nous avons choisi d'utiliser des ontologies semi-formelles et de les structurer en Topic Maps, un formalisme de représentation de connaissances adapté à la navigation [BRI04].

3 Quels standards et outils d'annotation ?

3.1 Les standards d'annotation existants

Notre but étant d'ajouter des commentaires à un ensemble de documents, il faut se concentrer sur la question de l'ancrage de ces annotations et la forme de ses méta-informations dans le document original. Cette

problématique est abordée dans le domaine du WS dont le but est d'enrichir les ressources Web en leur liant des informations descriptives structurées pour améliorer leur accessibilité, leur recherche et l'utilisation de l'information. Nous allons maintenant décrire les outils proposés dans ce domaine, que nous allons réutiliser et enrichir dans l'optique de notre projet.

Le WS identifie trois types d'annotations : (1) de simples méta-données (du type date de modification, auteur, etc.), (2) des annotations que nous qualifions de « computationnelles » dans la mesure où elles s'adressent à des programmes en leur permettant de tirer meilleur parti des ressources annotées [BRE01], [VOL03], [ROU01], et (3) des annotations que nous qualifions de « sociales » puisqu'elles s'adressent au lecteur, à l'utilisateur humain, lui permettant d'être un participant actif du Web.

Les outils développés depuis le début des années 90 permettent de réviser des textes à l'aide de commentaires ou d'explications, de justifier des décisions. En général, ils sont composés de divers modules permettant de visualiser, créer, stocker et rechercher des annotations. Celles-ci sont définies par une ancre, des attributs et un corps. Elles sont stockées sur un serveur dédié (serveur d'annotations), et peuvent être classées selon leurs attributs, avoir un statut public, privé, ou être partagées par un groupe défini. Le serveur contient les informations sur la localisation de l'annotation (le document sur lequel l'annotation a été créée ou sa place dans le document), son style (police, couleur,...), son contenu (texte et attributs), et la fonction de l'annotation (ses liens par exemple). Les annotations sont le plus souvent reliées en arbre, ce qui facilite la navigation et leur gestion.

Toutes ces recherches ont mené à la définition du standard Annotea [ANN03] par [KAH01] du W3C, basé sur une description des annotations en RDF [BRI04] et améliorant la collaboration au travers de méta-données partagées basées sur des annotations Web, des signets, et leurs combinaisons. Plusieurs serveurs (ZAnnot [ZAN03], Annotea) et clients (Annozilla [ANZ04], Amaya [AMA05]) d'annotations implémentent le standard Annotea. Le serveur d'annotations ZAnnot conserve les annotations dans une base de données RDF, et les utilisateurs peuvent interagir avec le serveur par le client Annozilla, plug-in du navigateur Mozilla, afin de rechercher une annotation, en créer une nouvelle ou en supprimer une autre. Une annotation est décrite par un ensemble de méta-données (ses attributs définies par un schéma RDF) et un corps. L'avantage de la notation en RDF est qu'il est possible de la personnaliser, en ajoutant par exemple au schéma de l'annotation, des attributs ou un ensemble de valeurs de ces attributs. Cette solution technique est donc intéressante puisqu'il est donc possible d'ajuster le modèle à notre besoin d'indexation multidimensionnelle (voir section 5). En effet, les méta-données proposées par les standards du Web pour indexer les annotations (nom de l'auteur, date, thème, type d'annotation, etc.) ne sont donc pas suffisantes dans notre contexte. En fait, avec ce type d'index, nous ne pouvons pas conserver le contexte organisationnel (rôles, profil des participants, etc.), le domaine contextuel (lexique spécifique, mots-clés du domaine, concepts, etc.), ni le type d'argumentation (proposition, opposition). Nous proposons donc de compléter les attributs Annotea préexistants par d'autres attributs liés à un usage « socio-sémantique » des annotations dans notre problématique de soutien cognitif de la coopération dans des projets de conception.

Nous allons maintenant décrire et classer ces outils d'annotations existants, et clarifier notre positionnement.

3.2 Les outils d'annotation existants

A l'heure actuelle, plusieurs clients d'annotation sont disponibles, issus d'initiatives du WS. La plupart adopte ce que nous avons appelé une approche « computationnellement sémantique » ayant pour objectif d'indexer des pages Web plus ou moins automatiquement. Ils sont utilisés pour la création de méta-données et certains se basent sur des ontologies pour soutenir l'annotation computationnelle : OntoMat-annotizer [HAN02] ; Melita [DIN03] ; MnM [DOM02]. Les annotations computationnelles sont liées géographiquement à une partie d'une page Web et rajoutent des informations supplémentaires, mais elles n'aident pas à la coopération ou aux interactions entre les lecteurs d'une même page. Ce sont des méta-données qui indexent une page, et permettent aux moteurs de recherche un meilleur rappel d'informations ou de pages.

D'autres clients d'annotation adoptent une approche plus sociale, ayant pour objectif de faciliter la communication, sans prévoir de fonctionnalités pour l'indexation ou la récupération d'annotation, même si ces annotations peuvent être triées sur des méta-données rudimentaires telles que la date de création ou l'auteur : Yawas [DEN00] ; CritLink [KAP98] ; XLibris [PRI98] ; etc. Ces outils d'annotation considèrent l'annotation comme un commentaire, conception partagée par certains logiciels propriétaires ou certains plug-in de logiciels spécifiques, où les commentaires ne sont ni indexés ni différenciés du document [WIN03]. Les annotations sont parfois stockées à part sur des serveurs d'annotations [ACR04] et organisées de façon minimaliste. Cependant,

ces outils d'annotation ne permettent pas de relier les annotations entre elles qui ne peuvent donc pas représenter un ensemble d'échanges entre des utilisateurs à propos d'un document. Même si ces outils d'annotation soutiennent plus facilement l'interaction que les outils d'annotation computationnels, ils ne sont pas suffisants pour notre propos.

Nous pouvons finalement classer schématiquement les outils d'annotation en deux familles ; l'une se concentre sur l'indexation de pages Web favorisant leur rappel, tandis que l'autre se concentre sur la communication humaine au travers de commentaires. Dans une visée de conception d'environnement support au travail coopératif, on peut déplorer l'absence de possibilités de gestion des annotations ou d'interactions entre utilisateurs dans ces deux familles d'outils. Nous proposons donc de les enrichir grâce aux techniques d'indexation du WS et de soutenir l'utilisateur dans son activité d'annotation documentaire, en vue de l'aider à travailler de manière coopérative. De plus, nous proposons d'autres fonctions d'annotation telles que le multi-ancrage (permettant de relier plusieurs fragments de documents) ou la possibilité de répondre à une annotation. Nous exposons dans la partie suivante les fonctionnalités d'une application soutenant la coopération autour de documents, en nous appuyant sur une approche socio-sémantique du Web.

4 Spécifications de l'outil AnT&CoW

Nous avons défini précédemment l'annotation comme un type de méta-donnée située, reliée à un autre document. Cet ensemble est relié à divers paramètres tels que le temps, le lieu, les participants, son statut public ou privé, sa sémantique, ce qui signifie que l'annotation est une entité composée de plusieurs parties telles que son ancre (ou ses ancres) dans un document, ses attributs, et un corps (le texte de l'annotation). Nous considérons aussi que l'annotation est une trace du processus de coopération qui a deux fonctions principales : la planification (gestion du projet, micro-organisation) et la révision (argumentation, annotation constituante du corps d'un document).

Afin de permettre un classement plus fin de ces annotations, nous proposons, suivant [ZAC03], d'étendre l'indexation de l'annotation collaborative non seulement par des dimensions spécifiques aux domaines (thèmes), mais aussi par une dimension cognitive grâce à une dimension argumentative (conservant la trace des décisions et des négociations entre les participants humains) ainsi que par une dimension organisationnelle, se servant du rôle de l'acteur pour souligner l'importance d'une décision. Ces trois dimensions, de domaine, argumentative et organisationnelle sont nécessaire à la contextualisation de l'annotation.

4.1 Des ontologies sémiotiques pour l'indexation d'annotation selon trois dimensions

Nous proposons de décrire chacune des trois dimensions listées ci-dessus par une ontologie. Dans une perspective de WS, les ontologies sont censées représenter exhaustivement les connaissances d'un domaine spécifique, structurées en hiérarchie de concepts par des relations entre ceux-ci. Chaque concept est bien défini par toutes ses propriétés, et l'expert doit donc spécifier entièrement les relations entre les concepts. Cependant, les experts humains ont souvent des définitions conflictuelles de certains concepts pour lesquels plusieurs définitions sont en concurrence. Parallèlement, des mécanismes d'inférence spécifiques calculent la cohérence et la consistance de ces ontologies. Cette opération, tout comme la construction, est longue et coûteuse et il n'est pas toujours possible de réutiliser des ontologies existantes. En effet, d'un côté, il existe des ontologies génériques (EuroWordNet [VOS98], DOLCE [GAN03]) qui ne sont pas adaptées à une application spécifique à un domaine puisqu'elles ne contiennent pas de définitions de concepts spécifiques à un domaine. Et d'un autre côté, on rencontre des ontologies de domaine souvent onéreuses et peu disponibles même si leur portabilité est augmentée par le respect de standards du W3C (OWL, RDF). Ainsi, il est difficile d'élaborer une représentation du contenu sémantique de pages Web, même à l'aide d'ontologies.

Pour éviter ces inconvénients, le W2S propose d'utiliser des ontologies moins formelles dont l'objectif principal est d'aider l'utilisateur à naviguer dans le Web. Dans cette perspective, les concepts peuvent être moins spécifiés et il n'est donc pas nécessaire d'en lister toutes les propriétés. Les standards comme Topic Maps (TM) (Norme ISO, [BIE99]) sont définis pour ces ontologies semi-formelles. Le formalisme TM définit un réseau de concepts couvrant des connaissances de domaine. Les thèmes (topics) sont définis par de simples URL afin que tous les utilisateurs en aient la même définition. Les thèmes sont hiérarchiquement organisés (relations « est un ») et associés par des relations horizontales (« partie de », « utilisé par »). Aucun mécanisme de cohérence n'est utilisé par les TM.

Puisque les TM ne nécessitent pas de définitions précises des concepts et sont conçues pour soutenir l'utilisateur dans sa navigation Web, nous avons adopté ce formalisme pour représenter nos ontologies dimensionnelles.

Dans notre système, les ontologies organisationnelle et argumentative sont construites manuellement. La première est fondée sur une analyse sociale des acteurs et la seconde sur une analyse cognitive et pragmatique des interactions par les échanges. L'ontologie de domaine, elle, nécessite une combinaison de techniques issues du Traitement Automatique des Langues (TAL) et d'opérations manuelles sur le choix des termes et des concepts. Ces trois ontologies sont stockées sur un serveur d'ontologies qui permet une récupération aisée des concepts. Portons notre attention sur ces techniques de TAL.

4.2 La construction d'ontologies de domaine contextuelles par les outils et méthodes issus du T.A.L.

En raison du peu de disponibilité d'ontologies de domaine et de l'utilisation impossible d'ontologies génériques, nombre de projets utilisent les techniques de TAL pour extraire semi-automatiquement des termes (des instances de concept) [JAC03], pour créer des agrégations de termes (concepts) [CIM04] ou encore pour extraire des relations entre termes [BUI04]. L'expert doit alors nommer les agrégats (clusters) en tant que concepts et doit éventuellement définir les relations entre concepts.

Dans le système d'annotation que nous proposons, les techniques de TAL sont utilisées pour deux objectifs : pour la construction et la mise à jour des ontologies de domaine à partir des corpora, mais aussi pour l'indexation des annotations. La première tâche est effectuée hors-ligne, en extrayant les termes d'un corpus sélectionné et en proposant une hiérarchie de thèmes simple (où un terme est équivalent à un thème). La seconde tâche est le soutien à l'utilisateur en lui proposant semi-automatiquement d'indexer ses annotations selon les trois dimensions (les index du type nom de l'auteur, date ou titre sont automatiques).

Dans le cas de la première tâche, la construction d'une ontologie initiale à partir de corpora, nous sommes parties du texte pour créer une ontologie semi-formelle (structurée en thèmes) en utilisant un extracteur de termes. Nous avons choisi LIKES [ROU96] qui est un extracteur de termes simple qui identifie des séquences de mots fréquents (segments répétés) apparaissant dans le corpus. Les segments répétés sont des candidats termes potentiels organisés en arbre, regroupés selon leur tête et affichés selon leur fréquence d'apparition. Les candidats termes sont utilisés afin de sélectionner les thèmes de notre ontologie. Les sorties sont filtrées en vue d'éliminer les candidats termes incorrects (les termes finissant par une préposition, une conjonction). La plupart des candidats termes correspondent à un patron Tête + Modifieur.

Nous avons développé un outil (GenTMInd) qui identifie les relations hiérarchiques entre les termes via des règles heuristiques et qui structure l'ensemble en TM. Ainsi, un terme qui correspond à un patron Tête + Modifieur est un sous-concept du concept Tête. Pour le moment, les candidats thèmes sont identifiés parmi des syntagmes nominaux simples (un syntagme nominal suivi par un syntagme prépositionnel au plus). Ces hypothèses et ces règles heuristiques ne sont pas suffisantes pour identifier toutes les relations hiérarchiques ou tous les candidats thèmes pertinents. Une fois un corpus pertinent rassemblé, nous étendrons la recherche de candidats sur un ensemble de verbes spécifiques au domaine. Nous explorerons le contexte de chaque candidat thème afin d'identifier plus de relations entre les thèmes. S'il est possible de trouver d'autres candidats thèmes apparaissant fréquemment dans le contexte, cela signifierait que des relations horizontales doivent être ajoutées entre deux candidats thèmes. Par exemple, dans notre domaine de conception mécanique, le terme « flasque » apparaît à plusieurs reprises dans le contexte du terme « moteur ». On pourrait donc ajouter une relation « concern » entre « flasque » et « moteur », à faire valider par l'expert.

La seconde tâche, la proposition de termes (mots-clés) à l'utilisateur afin qu'il indexe ses annotations, implique que les outils TAL soient capables d'analyser un texte court (l'annotation soumise par l'utilisateur) et de faire correspondre les éléments de ce texte aux thèmes de l'ontologie pour chacune des dimensions. Pour cette opération, nous souhaitons utiliser un algorithme de mise en correspondance entre le corps de l'annotation et la partie du document annoté et l'arbre TM construit par le système. Le système d'annotation proposera alors à l'utilisateur des mots-clés ou des syntagmes-clés spécifiques au domaine ainsi que des types argumentatifs. L'utilisateur décidera ensuite si l'indexation proposée est pertinente et s'il souhaite la conserver comme méta-donnée de son annotation. En créant son annotation, l'utilisateur décide si son annotation est ancrée à une ou plusieurs parties du document ou des documents. Ainsi, nous prévoyons une indexation complexe avec un multi-ancrage. Une fois la validation effectuée par l'utilisateur, l'annotation est stockée avec ses méta-données sur le serveur d'annotations.

Le prochain pas dans l'implémentation de cet outil est d'adapter un extracteur de termes plus efficace du type FASTR [JAC99], afin d'identifier les candidats termes dans le corps des annotations et d'extraire une hiérarchie de concept par les techniques de clustering [CIM04].

Nous allons maintenant présenter l'architecture distribuée et l'interface de notre outil d'annotation, intégrant les standards du W3C et des outils de TAL.

5 Architecture et Interface d'AnT&CoW

Respectant le standard Annotea du W3C, l'architecture de notre système d'annotation est une architecture distribuée client/serveur (FIG.1) :

Partie client : comme mentionné précédemment, le but est d'annoter des documents qui sont accessibles par un navigateur Web. Pour cela, nous avons choisi Annozilla, plug-in pour le navigateur Mozilla qui est un client Annotea répondant à notre objectif. Utilisant les XPointer, les standards DOM et de nombreuses fonctions de l'infrastructure Mozilla (XPConnect, composants XPCOM), Annozilla offre des possibilités de création, de mise à jour et de suppression d'annotations sur un document ou une partie de document et permet de les stocker sur un serveur local (usage individuel) ou distant (usage partagé).

Partie serveur : Nous avons choisi le client Annotea ZAnnot développé sur la plateforme Zope [LAT03] qui possède un Object DB, un serveur Web et plusieurs autres composants (ZClass, Zope Products,...). ZAnnot tire profit de la plateforme Zope et gère à la fois les requêtes envoyées par le client Annozilla et la fonction de réponse à une annotation.

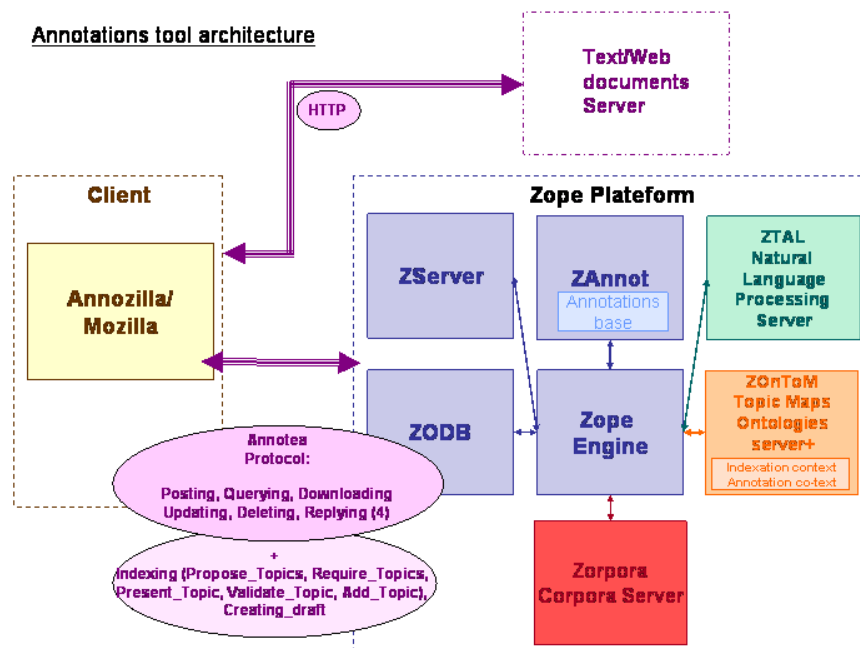


FIG. 1 – Architecture d'AnT&CoW

Adaptant Annozilla pour l'annotation de documents selon nos besoins, nous avons implémenté les fonctions de réponses entre annotations, la création de sommes et le mécanisme d'indexation. Pour classifier les annotations, nous avons étendu le schéma d'annotation d'Annotea en lui ajoutant les méta-données correspondant aux trois dimensions qui seront sauvegardées au format RDF de la même manière que les autres méta-données et le corps de l'annotation. Pour des raisons de cohérence, les Topic Maps des différentes dimensions sont actuellement stockées au format RDF et ne sont pas modifiables par l'utilisateur.

Nous fournissons donc à l'utilisateur une interface (FIG.2, encadré 1) lui permettant de gérer les thèmes des différentes dimensions et de naviguer dans les annotations. Lorsqu'un membre du projet ouvre un document dans son navigateur Web, il a la possibilité d'ouvrir dans la partie gauche de la fenêtre principale, le plug-in Annozilla, qui permet aussi bien d'annoter que de récupérer les annotations organisées selon leurs attributs

définis plus haut. Les annotations sont localisables par une ancre dans le document (un crayon). Si l'annotateur décide de créer une nouvelle annotation, cette annotation apparaît dans une nouvelle fenêtre contenant son corps et les champs d'indexation. L'utilisateur peut décider d'effectuer un multi-ancrage grâce à cet outil. C'est-à-dire que s'il souhaite répondre à une annotation précédemment déposée ou s'il souhaite justifier une remarque qu'il fait par la création d'un lien vers une autre partie du document, notre outil peut gérer différentes ancres pour une seule annotation. C'est cette fonctionnalité de multi-ancrage qui permet de créer des fils thématiques. L'utilisateur peut aussi décider d'éditer les annotations dans un nouveau document (FIG.2 encadré 2), laissant apparaître le contexte argumentatif (fil de discussion) dans lequel les annotations trouvent leur place.

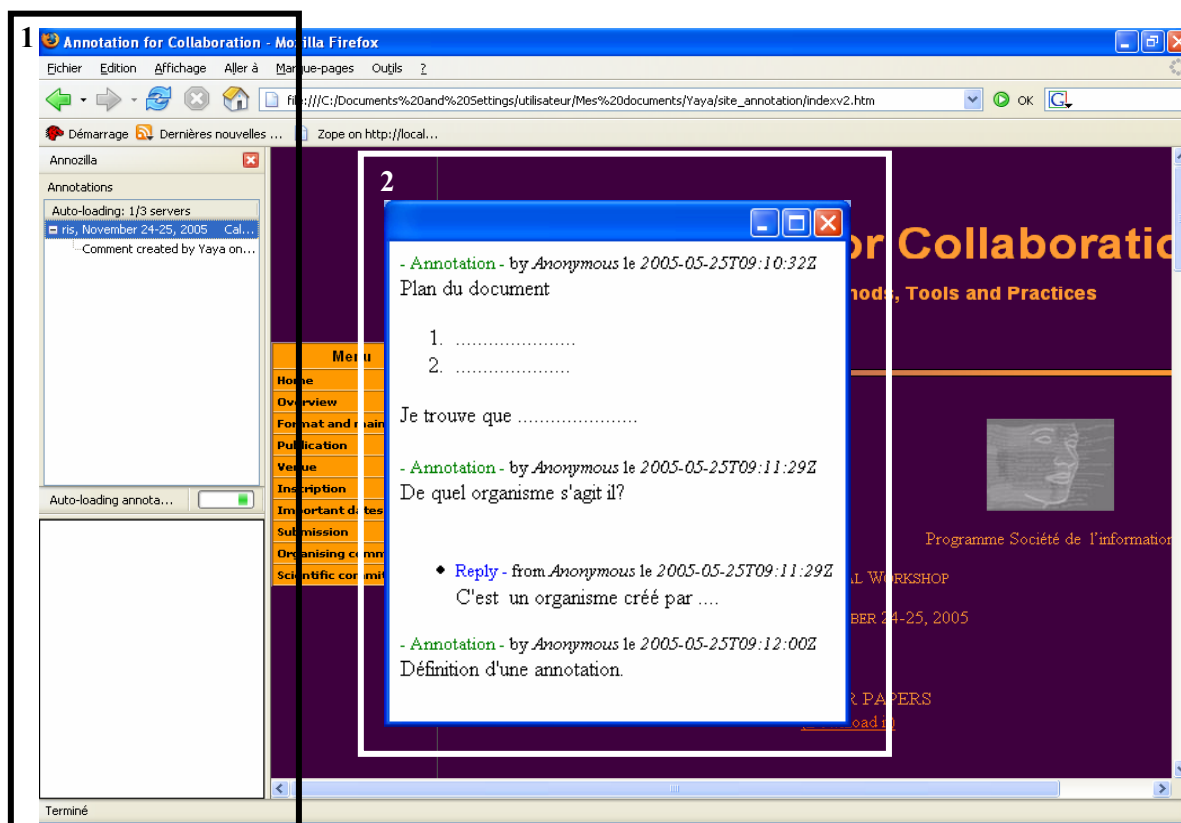


FIG. 2 –Interface d'AnT&CoW

La prochaine étape de notre développement consiste à intégrer dans notre architecture les éléments d'indexation (c'est-à-dire les ontologies et les outils de TAL permettant l'organisation, la récupération et l'affichage d'annotations sélectionnées) et relier le serveur d'ontologie au serveur d'annotation afin que les TM représentant les dimensions puissent servir à une indexation semi-automatique. La mise en ligne de ce serveur d'ontologie permettra également leurs mises à jour, par les utilisateurs ou par les outils de TAL.

6 Conclusion

Dans cet article, nous avons présenté un outil Web destiné à soutenir les interactions au travers des annotations. Cet outil a pour objectif d'être un complément à des phases de travail synchrone et permet de rendre explicite la compréhension d'un document et la logique de conception d'une idée entre différents acteurs impliqués dans une activité commune. Nous considérons ainsi gérer la connaissance dans l'action, en soutenant les interactions humaines. Notre outil d'annotation respecte le standard Annotea du W3C et utilise des techniques de TAL à la fois pour créer des ontologies semi-formelles permettant d'indexer les annotations et pour aider l'utilisateur à indexer ses annotations par le biais d'une indexation semi-automatique. Ce travail est toujours en cours et certains points restent encore à améliorer. Sur le plan de l'utilisation des outils de TAL, la prochaine étape consiste à adapter un extracteur de terme plus performant, comme FASTR [JAC99], à identifier des candidats termes dans le corps de l'annotation et à extraire une hiérarchie de concepts en adaptant des

techniques de clustering [CIM04]. Sur le plan de l'outil lui-même, la prochaine étape consiste à fournir à l'utilisateur une interface lui permettant de gérer les thèmes des trois dimensions (domaine, organisationnel et argumentatif) et de naviguer au mieux à l'intérieur de l'ensemble des annotations stockées. Cela nous permettra de débiter une phase d'expérimentation avec les utilisateurs visant à tester la pertinence de nos résultats.

Références

- [ACR04] Acrobat PDF, (2004), <http://www.adobe.com/support/techdocs/ac76.htm>
- [AMA05] <http://www.w3.org/Amaya/>
- [ANN03] Annotea (2003), <http://www.w3.org/2001/Annotea/>
- [ANZ04] Annozilla (2004), <http://annozilla.mozdev.org/>
- [BAN96] Bannon, L, et Kuutti, K. (1996), Shifting Perspectives on Organizational Memory: From Storage to Active Remembering. *Proceedings of the 29th IEEE HICSS, vol. III, Information Systems - Collaboration Systems and Technology*. IEEE Computer Society Press, Washington 1996, pages 156-167.
- [BIE99] Biezunski, M., Bryan, M., et Newcomb, S. R., (1999) *Topic Maps*, spécification ISO/IEC 13250, 3 Décembre 1999.
- [BRE01] Bremer J.M., et Gertz M., (2001) Web Data Indexing through External Semantic-carrying Annotations. In *11th IEEE Int'l Workshop on Research Issues on Data Engineering: Document management for data intensive business and scientific applications (RIDE-DM'2001)*, IEEE Computer Society, pp. 69-76.
- [BRI04] Brickley, D., et Guha, R.V., (2004) *Resource Description Language* - <http://www.w3.org/TR/rdf-schema/>, February 2004
- [BUI04] Buitelaar P., Olejnik D., Hutanu M., Schutz A., Declerck T., et Sintek, M. (2004), Towards Ontology Engineering Based on Linguistic Analysis, in *Proceedings of LREC'2004*, Lisbon, may 2004, ISBN 2-9517408-1-6, pp.7-11
- [CAU02] Caussanel, J., Cahier, J.P., Zacklad, M., et Charlet, J., (2002) Les Topic Maps sont-ils un bon candidat pour l'ingénierie du Web sémantique. In Bruno Bachimont, editor, *Actes des 6 emes Journées Ingénierie des Connaissances*, pages 233-52, Rouen, France, 28-30 mai 2002
- [CHA02] Charaudeau P. et Maingueneau D., (2002) article Discours in *Dictionnaire d'analyse du discours*, Seuil.
- [CIM04] Cimiano, P. Hotho, A., et Staab S. (2004), Clustering Concept Hierarchies from Text, in *Proceedings of LREC'2004*, Lisbon, may 2004, ISBN 2-9517408-1-6, pp. 1721-1724.
- [COO99] Cook, S. Brown, J. (1999), *Bridging Epistemologies: The Generative Dance Between Organizational Knowledge and Organization Knowing*, *Organization Science*, Vol.10, n°4, 1999, pp. 381-400.
- [DEN00] Denoue, L., et Vignollet, L., (2000) An annotation tool for Web browsers and its applications to information retrieval, in *proceedings of RIAO 2000*
- [DIN03] Dingli A., (2003) Next Generation Annotation Interfaces for Adaptive Information Extraction. In *6th Annual Computer Linguists UK Colloquium (CLUK 03)*, January, 2003, Edinburgh, UK
- [DOM02] Domingue J.B., Lanzoni M., Motta E., Vargas-Vera M., et Ciravegna F., (2002), Mnm: Ontology driven semi-automatic or automatic support for semantic markup. In *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02)*, October.
- [GAN03] Gangemi, A., Guarino, N., Masolo, C., et Oltramari, (2003) A. Sweetening WordNet with DOLCE, *AI Magazine* 24(3): Fall 2003, 13-24
- [HAN02] Handschuh, S., Staab S., et Ciravegna, F., (2002), S-cream - semi-automatic creation of metadata. In *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02)*, October.
- [JAC03] Jacquemin C. et Bourigault D. (2003), Term Extraction and Automatic Indexing, in Mitkov R. (ed), *The Oxford Handbook of Computational Linguistics*, Oxford University Press, 2003, pp. 599-615
- [JAC99] Jacquemin, C., et Tzoukermann, E. (1999), NLP for Term Variant Extraction: A Synergy of Morphology, Lexicon and Syntax. In T. Strzalkowski, editor, *Natural Language Information Retrieval*, pages 25-74, Kluwer, Boston, MA, 1999
- [KAH01] Kahan J., Koivunen M.-R., Prud'Hommeaux E., et Swick R.R. (2001) Annotea : an open RDF Infrastructure for Shared Web Annotations, *Proceedings of WWW10*, May 1-5 2001, Hong-Kong, pp. 623-632.
- [KAP98] Ka-Ping Y., (1998) *CritLink : Better hyperlinks for the WWW*. <http://crit.org/ping/ht98.html>, 1998

- [LAB05] Labbe H. & Marcoccia M., (2005), Communication numérique et continuité des genres : l'exemple du courrier électronique, *Texte !*
- [LIB00] De Libera A., (2000), *La philosophie médiévale*, Paris, PUF (« Que sais-je ? » 1044), 4e éd.
- [LAT03] Latteier A., Pelletier M., McDonough C., et Sabaini P. (2003), *The Zope Book*, Edition 2.6. http://zope.org/Documentation/Books/ZopeBook/2_6Edition/ZopeBook-2_6.pdf
- [LOR05] Lortal G., Lewkowicz M., Todirascu-Courtier A. (2005). Modélisation de l'activité d'annotation discursive pour la conception d'un collecticiel support à l'herméneutique, in *Actes de la conférence IC2005* p. 169-180.
- [PRI98] Price, M., Schilit, B., et Golovchinsky, G., (1998) XLibris: The active reading machine. In *proceedings of CHI'98 Human factors in computing systems*, Los Angeles, California, USA, vol.2 of Demonstrations: Dynamic Documents, pages 22-23, 1998
- [ROU01] Roussey C., Calabretto S., et Pinon J.-M., (2001) SyDoM: A Multilingual Information Retrieval System for Digital in *proc. International Conference ICC/IFIP On Electronic Publishing (ELPUB'2001)*, Canterbury (UK), 5-7 July 2001, p. 150-164,
- [ROU96] Rousset, F., Frath, P., et Oueslati, R. (1996), Extracting concepts and relations from Corpora. In *Proceedings of the Workshop on Corpus-oriented Semantic Analysis, European Conference on Artificial Intelligence, ECAI 96*, Budapest, 12 August 1996.
- [ROM74] Rommetveit, R. (1974). *On message structure: A framework for the study of language and communication*. London: John Wiley.
- [VOL03] Volz R., Oberle D., Motik B., et Staab S. (2003), KAON SERVER - A Semantic Web Management System? In: *Proceedings of the 12th World Wide Web, Alternate Tracks - Practice and Experience*, Hungary, Budapest, 2003.
- [ZAC03] Zacklad, M., Lewkowicz M., Boujut, J-F., Darses, F., et Détienne, F (2003), Formes et gestion des annotations numériques collectives en ingénierie collaborative, *actes des journées Ingénierie des Connaissances 2003*, Laval.
- [ZAC04] Zacklad, M., et Barbaud, X., (2004) Vers une application du Web Socio Sémantique pour la réalisation d'un système d'information destiné aux réseaux de santé, In *Second séminaire francophone du Web Sémantique Médical* - 9 mars 2004, Rouen
- [ZAN03] ZAnnot (2003), *Zannot* <http://www.zope.org/Members/Crouton/ZAnnot/>
- [ZWE94] Zweigenbaum, P; et Consortium MENELAS, MENELAS : an access system for medical records using natural language. In *Computer methods and programs in Biomedicine*, 45:117-120, 1994