Médiatiser l'annotation discursive support à l'herméneutique

Lortal Gaëlle

Laboratoire ISTIT/Tech-CICO– Université de technologie de Troyes
12, rue Marie Curie BP 2060 10010 Troyes Cedex
gaelle.lortal@utt.fr

Résumé – Abstract

Nous présentons une étude pour le développement d'un outil adapté à une activité coopérative médiatisée centrée sur le document. En TCAO, l'annotation permet d'échanger et d'argumenter en mode asynchrone dans un but de conception. Nous proposons une définition de l'annotation suivant ses fonctions d'étiquetage (annotation computationnelle du Web Sémantique) et de communication (commentaire du Web Social). Si les annotations computationnelles servent à optimiser l'interopérabilité logicielle ou la recherche d'informations (indexation), les commentaires ont un but coopératif cognitif. Il faut donc développer un outil permettant de prendre en compte les deux facettes de l'annotation, (1) l'interprétation du document, et (2) la récupération d'informations. Dans le cadre de l'optimisation de récupération d'information, il est possible d'utiliser des outils de T.A.L. pour aider l'utilisateur à mieux organiser ses annotations, via une indexation fine. Nous souhaitons aider l'utilisateur en lui proposant des candidats termes à l'indexation grâce à des Topic Maps (cartes de thèmes) respectant différents points de vue (de domaine, organisationnel, argumentatif), et créées semi-automatiquement. Nous avons opté pour un développement itératif de notre outil (en cours).

This paper aims at describing a study about the development of an annotation tool adapted to collaborative activity around a document. Annotating is central in Computer Supported Collaborative Work, as mediated work needs means to argue for design. In this framework, we propose a definition of annotation merging Semantic Web view of computational annotation (tagging) and Social Web view of annotation (comments). Computational annotations are used by specific software search, while comments are used to interpret the document. From this study of annotations, we identify some functionalities to include on an annotation tool merging the social and the computational annotation to enable discourse about a document in a distributed environment. Annotations are indexed (computational annotation), but also organised in a discourse. To develop this kind of tool, we can use N.L.P. tools to support user annotating. We propose indexing functionalities via a semi-automatically created Topic Maps (TM) displaying terms candidates. These TMs are multi-dimensional and focused on domain, organisation, and arguing. Our tool is still under iterative development.

Keywords – Mots Clés

Annotation, Coopération, Interprétation, Traitement Automatique des langues, Classification
Annotation, Collaboration, Sensemaking, Natural Language Processing, Classification

Introduction

L'herméneutique vise l'interprétation de textes sacrés ou juridiques. Elle s'appuie sur un dispositif d'annotation complexe. Aujourd'hui, nous passons des échanges discursifs autour de documents « papier » et des techniques herméneutiques, à des échanges discursifs médiatisés se modelant sur les supports informatiques disponibles. Nous postulons qu'il est possible d'améliorer la coopération autour de documents numériques en adaptant les techniques d'interprétation herméneutiques à un environnement médiatisé, principalement par la médiatisation de l'activité d'annotation pour une interprétation collective (Weick, 1979). Nous nous focaliserons sur les principes de développement d'un outil d'annotation collaboratif et principalement sur le soutien possible à apporter à un utilisateur par l'intégration d'outils du Traitement Automatique du Langage (T.A.L.). Pour permettre une herméneutique numérique, nous proposons de médiatiser l'annotation pour l'interprétation de documents. Nous présentons tout d'abord deux aspects de l'annotation, identifiés par l'étude de ses propriétés : l'aide à l'interprétation d'une part et l'aide à l'indexation automatique d'autre part. Dans une seconde section, nous exposerons les différents problèmes de médiatisation que pose l'annotation visible sous forme d'index comme de commentaire. L'annotation possède une valeur argumentative essentielle puisqu'elle est la représentation que ce fait un lecteur d'un document ou d'un autre commentaire déposé sur un document. En cela, l'annotation porte intrinsèquement la trace du locuteur. Elle est dépositaire à la fois du rôle du lecteur dans sa prise de position par rapport au document, du rôle de l'acteur dans le contexte collectif où il annote et enfin d'une thématique. Il est difficile de gérer cette structuration multi-points de vue des annotations, comme les fils multiples créés par les échanges annotatifs. Dans une troisième partie, nous postulons que seule la participation de l'annotateur peut permettre une bonne gestion de cette structuration complexe. Nous développons alors notre proposition de soutien de l'annotateur dans ces activités d'indexation grâce à l'utilisation d'outils de T.A.L. intégrés à un outil d'annotation. L'outil d'annotation permet aux utilisateurs d'échanger et de construire une interprétation partagée de documents, quant à la récupération des fragments d'interprétation (les annotations), leur soutien est effectué par des outils d'extraction de termes et de classification de termes/concepts. Nous présentons ensuite notre première version d'un outil d'annotation collaboratif et enfin nos perspectives de développement par la confrontation à divers terrains.

1. L'annotation, un outil herméneutique collaboratif

Dans un contexte d'échanges autour de documents entre différents acteurs pour la conception d'une représentation partagée d'une problématique, nous nous sommes penchés sur une discipline coopérative structurée autour des textes, l'herméneutique. L'herméneutique se fonde sur des annotations à des fins de communication sur l'interprétation de textes. Par exemple, la glose est un énoncé explicatif lié à un texte qui devient un exercice rhétorique scolaire au Moyen-Âge, puis un exercice discursif public sous forme de commentaires, de sentences, voire de sommes, résumant les connaissances sur un sujet (De Libera, 2001). Les

débats sont publics et se fondent sur les argumentations de plusieurs orateurs en formation universitaire. Dans ce contexte l'annotation est à la fois un objet lié à un document et un outil énonciatif permettant l'argumentation. Sa forme brève et son caractère relationnel peuvent être expliqué, comme dans la génétique du courrier électronique selon (Labbe & Marcoccia, 2005) par son appartenance aux formes de dialogues épistolaires brefs tels que le billet selon (Haroche-Bouzinac, 2000). Les caractéristiques de ces genres d'énoncés sont la brièveté, leur caractère informel, informationnel, séquentiel et relationnel. Dans l'usage du billet ou du courrier électronique, on observe des mises relations (type message d'accompagnement d'une pièce jointe) où le message met en lien un auteur, un destinataire, et un document. Dans un cadre de travail collaboratif, l'annotation, qu'elle soit dans un but de planification (un post-it informatif laissé sur un dossier et destiné à l'équipe suivante) ou dans un but argumentatif (annotation en marge exprimant une opinion) possède ces caractéristiques de dialogue épistolaire (adressé) de forme brève. Une autre facette de l'annotation se définit par cette caractéristique relationnelle, l'index. Index est alors pris avec son sens étymologique d'« indicateur ». Il est alors un type de balise référentielle, liant divers objets ou pointant vers un chemin dans un réseau. La balise informatique, elle, appartient au document qu'elle vise et qu'elle modifie de l'intérieur ; elle est l'objet recherché par les outils d'annotation automatique, d'étiquetage).

L'annotation étant fortement liée à un 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[1]/texte courante dans le champ de la linguistique. 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. Les annotations, fragments textuels ancrés à un document et au point de vue de leur auteur, deviennent alors des fragments discursifs liés au texte, ou reliant différents textes. Motivant par ailleurs la conception d'autres objets (un document, un produit, une autre annotation), par les échanges qu'elle génère, on considère alors l'annotation comme un DOcument Pour l'Action (DoPA) suivant (Zacklad, 2004). Selon toutes ces caractéristiques, l'annotation peut alors être décrite comme un continuum allant du computationnel au cognitif selon des degrés de formalité différents (voir fig.1).

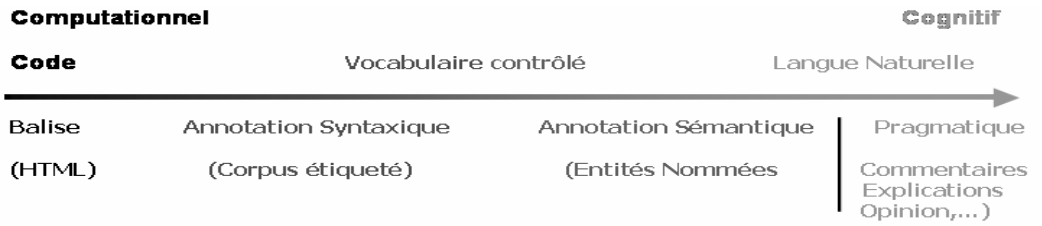

Figure 1 : Continuum d'annotation

Considérant l'annotation comme une balise permettant l'indexation d'un fragment textuel aussi bien que comme un énoncé porteur du rôle de l'auteur, il s'agit pour nous de médiatiser

[1] « inclusion d'un texte dans son contexte » (Charaudeau et Maingueneau, 2002).

les interactions humaines autour d'un document numérique par des annotations discursives, tout en permettant leur structuration par des annotations computationnelles.

2. Un fragment de discours à médiatiser

Dans le domaine de la conception ou de la rédaction (par exemple en Rédaction Collaborative Assistée par Ordinateur), le développement de collecticiels pertinents doit permettre de soutenir les échanges et la construction d'une interprétation collective autour de documents numériques textuels. L'herméneutique nous propose un contexte de travail adapté et nous avons proposé un modèle pour notre outil inspiré de l'herméneutique (Lortal *et al.*, 2005). Dans le développement d'un outil, il faut prévoir de pallier plusieurs faiblesses du discours médiatisé et d'être représentatif des caractéristiques multidimensionnelles des annotations discursives.

2.1 Faiblesses de la médiatisation du discours

Suivant le Web Social (WSo) (Jordan *et al.*, 2003 ; Reed *et al.*, 2004), nous considérons que l'annotation est un élément de collaboration dans un projet de conception collaborative. Elle n'est pas simplement une métadonnée sur un document, mais plutôt un élément discursif autour d'un document permettant une communication et l'émergence d'idées autour de ce document partagé par un groupe concepteur. Cependant, on constate dans les outils du WSo s'essayant à ce type d'annotation une perte de la structure de l'information et du document. Des phénomènes de digression thématique et de décomposition thématique (Topic decay) sont notés respectivement par (Marcoccia, 2004) et (Herring, 1999). Ils sont dus au multi-threading (plusieurs thèmes dans un fragment) courant dans les discussions et problématiques dans les discussions médiatisées.

Pour pallier ces déstructurations, nous proposons de suivre les recommandations du Web Sémantique (WS) (Berners Lee *et al.*, 2001). Le WS recommande nombre de formalisations standardisées afin d'homogénéiser les ressources Web et permettre l'interopérabilité des applications. Ainsi, l'accent est mis sur l'amélioration des moyens à mettre en œuvre pour la récupération des documents principalement au niveau de l'indexation des documents (Volz *et al.*, 2003). Cette indexation est bien souvent soutenue automatiquement, par des outils d'annotation sémantique améliorant la communication machine (par ex. OntoMat-annotizer (Handschuh *et al.*, 2002)). Cependant, l'indexation automatique ne tient pas compte des liens sémantiques ou pragmatiques liés à la coréférentialité de la langue que l'on peut considérer comme un type d'indexation en langue naturelle.

2.2 Structuration multi-points de vue

Nous basons notre recherche à la suite de (Zacklad *et al.*, 2003), 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 aussi 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 mèls) et ceux d'un projet de conception produit, en mécanique (projet Air Campus, corpus de 150 mèls). 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. Par l'observation de ces terrains où l'annotation discursive est fortement contextualisée nous pouvons confirmer, la proposition de (Zacklad *et al.*, 2003) sur les multi-points de vue de l'annotation dans le domaine de la conception. Ainsi, nous proposons une indexation de l'annotation discursive pour la collaboration selon trois dimensions: (1) spécifique au domaine (thèmes), (2) argumentative (conservant la trace des décisions et des négociations entre les participants humains), et (3) organisationnelle (se servant du rôle de l'acteur pour souligner l'importance d'une décision).

Cette indexation sera soutenue par des ontologies semi-formelles et permettra un rappel important et une récupération optimale des annotations. En effet, elles permettent une indexation fine qui pour l'utilisateur se traduit par une ré-organisation possible des annotations selon sa tâche et son point de vue.

3. Utilisation des outils de T.A.L.

Nous offrons de soutenir cette collaboration discursive autour de documents par un système permettant d'annoter ces documents avec une finalité d'interprétation et d'appropriation, et de gérer les annotations une fois créées. Cette finalité n'est pas assistée par les outils informatiques d'annotations actuels du WSo qui ne permettent que des annotations isolées pauvrement indexées (date, nom d'auteur) et difficilement utilisables comme support aux interactions au sein d'un collectif (liées à une application client ou intégrées à un document). Ainsi, partant de théories rhétoriques, nous définissons les fonctionnalités d'un outil support à l'herméneutique en adaptant un modèle de l'activité d'annotation discursive instrumentée au travers de primitives de conception (Lortal *et al.*, 2005). Suite aux études de corpus, nous proposons, pour gérer ces annotations discursives, une indexation multi-dimensionnelle (domaine, argumentation, organisation) basée sur des techniques de Traitement Automatique de la Langue (T.A.L.). Les termes clés de chacune des trois dimensions seront représentés par des ontologies semi-formelles en Topic-Maps (TM) (Biezunski *et al.*, 1999). Puisque nous souhaitons soutenir l'utilisateur dans sa navigation thématique, nous adoptons ce formalisme TM permettant d'élaborer une représentation du contenu sémantique de document moins formelle, où les concepts de l'ontologie créée peuvent être moins spécifiés et il n'est donc pas nécessaire d'en lister toutes les propriétés. 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. L'utilisateur navigant dans les ontologies fera son choix parmi les concepts pour indexer son annotation. Cette phase d'indexation manuelle peut être fastidieuse et nous souhaitons donc l'assister grâce à des

outils de T.A.L. Les techniques de T.A.L. seraient donc utilisées dans deux objectifs : la construction et la mise à jour des ontologies de domaine à partir des corpora, et l'indexation des annotations.

3.1 Construction de l'ontologie de domaine

Comme les ontologies spécifiques à un domaine sont coûteuses et rarement disponibles, et que l'utilisation d'ontologies génériques est difficilement envisageable pour un domaine particulier, nombre de projets utilisent les techniques de T.A.L. pour extraire semi-automatiquement des termes (des instances de concept) (Jacquemin, 2003), pour créer des agrégations de termes (concepts) (Cimiano *et al.*, 2004) ou encore pour extraire des relations entre termes (Buitelaar *et al.*, 2004). La construction d'une ontologie initiale du domaine se base pour le moment sur l'utilisation d'un extracteur de termes, LIKES (Rousselot *et al.*, 1996), qui identifie des séquences de mots (segments répétés) apparaissant fréquemment 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 résultats sont facilement utilisables pour extraire une ontologie et nous l'avons choisi pour construire un premier noyau d'ontologie. 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. Le principal désavantage de LIKES reste le nombre important de termes candidats, ce qui nécessite une étape de nettoyage manuel de la hiérarchie obtenue. L'étape suivante consistera donc à adapter un extracteur de termes plus performant, tel que SYNTEX (Bourigault & Fabre, 2000) pour identifier des candidats termes dans le corps de l'annotation, construire et maintenir la TM.

3.2 Indexation de l'annotation

Soutenir l'indexation d'annotation selon trois dimensions implique que les outils T.A.L. puissent analyser un texte court (l'annotation) et faire correspondre les éléments de ce texte et ceux de son co-texte d'ancrage aux concepts, occurrences des concepts et contexte des occurrences formant l'ontologie du domaine. La structuration en TM permet le stockage de différents éléments co/contextuels selon un point de vue (scope) et de conserver le lien entre les concepts et leurs réalisations linguistiques (occurrences et contextes). Pour ces traitements, nous souhaitons utiliser un algorithme de mise en correspondance entre le corps de l'annotation et la TM construite par le système selon des principes d'analyse vectorielle (LSA, (Landauer et Dumais, 1997), algorithme de (Berry et *al.*, 1995). Le système d'annotation

proposera alors à l'utilisateur des mots-clés ou des syntagmes-clés spécifiques au domaine et à son argumentation. L'utilisateur décidera si l'indexation proposée est pertinente et la validera pour la conserver comme méta-donnée de son annotation. S'il ne trouve pas de concept adéquat, il pourra ajouter lui-même dans la TM l'élément manquant.

Dans le cadre d'une conception itérative du collecticiel d'annotation, un premier outil d'architecture distribuée client-serveur a été développé répondant au standard du W3C (Annotea, 2003). Il est basé sur le plug-in d'annotation du navigateur Mozilla-Firefox, (Annozilla, 2004). Ce plug-in est géré sur une plateforme Zope qui bénéficie d'un serveur d'annotation, (Zannot, 2004). Un serveur d'ontologie Topic Maps (ZOnToM) y a été ajouté afin de stocker une TM construite semi-automatiquement (Likes et GenTMInd) sur un corpus d'articles scientifiques (200 000 mots). Elle comprend environ 200 concepts avec 9 associations Tête+Modifieur par Tête. En plus de cette ontologie, ZOnToM conserve les occurrences des concepts et les co-textes des annotations déposées. La mise en correspondance n'est effectuée que par un algorithme temporaire à optimiser.

Conclusion et perspectives

Dans une problématique de médiatisation d'activité collaborative, l'annotation trouve sa définition dans un continuum allant de la balise au commentaire. L'annotation permet alors d'échanger autour d'un document et de construire une interprétation partagée, mais aussi d'indexer ce document en le replaçant dans son contexte. Dans cette présentation, nous nous sommes focalisée sur cette problématique d'indexation et l'utilisation d'outil de T.A.L.. En utilisant la première version de notre outil d'annotation sur nos terrains, nous pourrons évaluer nos hypothèses sur le statut de l'annotation dans différentes activités de conception collaborative (Rédaction assistée par ordinateur, conception produit et conception système).

Références

Annotea (2003), http://www.w3.org/2001/Annotea/

Annozilla (2004), http://annozilla.mozdev.org/

Berners-Lee T., Hendler J., et Lassila O. The Semantic Web, *Scientific American*, 2001

Berry M.W., Dumais S.T., et Shippy A.T. (1995), *A Case Study of Latent Semantic Indexing,* http://www.cs.utk.edu/~lsi/papers/index.html

Biezunski, M., Bryan, M., et Newcomb, S. R., (1999) *« Topic Maps »,* spécification ISO/IEC 13250, 3 Décembre 1999.

Bourigault D., Fabre C., Approche linguistique pour l'analyse syntaxique de corpus*, Cahiers de Grammaires*, n° 25, 2000, Université Toulouse - Le Mirail, p. 131-151

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

Charaudeau P. et Maingueneau D., (2002) article Discours in *Dictionnaire d'analyse du discours*, Seuil.

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.

De Libera A., 2000, *La philosophie médiévale*, Paris, PUF (« Que sais-je ? » 1044), 4e éd., 2000.

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.

Haroche-Bouzinac G., (2000) Une esthétique de la brièveté. *Revue de l'AIRE – Recherches sur l'épistolaire,* n° 25-26, p.49-51

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.

Jordan K.,Hauser J., Foster S., (2003) The augmented Social Network : Building idntity and trust into the next-generation Internet, in *firstmonday* vol.8, N.8, august 4th 2003 http://www.firstmonday.org/issues/issue8_8/jordan/

Labbe H. & Marcoccia M., (2005), Communication numérique et continuité des genres : l'exemple du courrier électronique, *Texto !*

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240

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.

Reed D., Le Maitre M., Barnhill B., Davis O., and Labalme F., (2004) The Social Web: Creating An Open Social Network with XDI in *Planet Network Journal* http://journal.planetwork.net/article.php?lab=reed0704

Rousselot, 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.

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.

Weick K.E., (1979) *The Social Psychology of organizing,* New York, Random House

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.

Zacklad, M. (2004). Processus de documentation dans les Documents pour l'Action (DopA) : statut des annotations et technologies de la coopération associées, in *actes du colloque Le numérique : Impact sur le cycle de vie du document pour une analyse interdisciplinaire*, 13-15 Octobre 2004, Montréal (Québec).

ZAnnot (2003), http://www.zope.org/Members/Crouton/ZAnnot/