

Annotations sur le Web : notes de lecture

E. DESMONTILS, C. JACQUIN

IRIN, Université de Nantes
2, rue de la Houssinière, BP92208
F-44322 NANTES Cedex 3, France
{desmontils, jacquin}@irin.univ-nantes.fr

Résumé

De nombreux systèmes de partage d'information existent de nos jours mais les spécificités du Web en font des outils extrêmement difficiles à exploiter. Les outils d'annotation visent à améliorer échange, communication et interopérabilité sur le Web. Notre objectif est de faire une synthèse des caractéristiques des annotations. Cette étude générale nous permet de caractériser les annotations sémantiques.

Mots-clés : Annotation, Système d'annotation, Annotation libre, Annotation sémantique, Méta-donnée

1 INTRODUCTION

L'arrivée du Web, et des technologies associées, confirme l'inspiration de V. Bush qui prônait dès 1945 la nécessité de proposer des outils interactifs pour partager l'information (Bush, 1945). De nombreux systèmes de partage d'information existent : cela va du Web aux outils très évolués de travail collaboratif (Lotus Notes...). Cependant, ce qui est possible avec ces derniers ne l'est plus avec le Web. Ces outils visent un groupe très réduit de personnes travaillant souvent ensemble, avec un vocabulaire commun, sur des thèmes proches et donc avec des habitudes spécifiques. Par contre, le Web concerne potentiellement des millions de personnes non seulement ne se connaissant pas mais ayant en plus des centres d'intérêt différents, des habitudes différentes, des cultures différentes... De plus, sur le Web, l'information est fortement distribuée, extrêmement volumineuse, évolutive, volatile, très "bruitée", très hétérogène et souvent très peu structurée. Dans ce contexte, il est nécessaire de proposer des méthodes et des outils pour comprendre, manipuler et partager des documents, pour mettre en place des services pertinents et performants.

L'annotation sémantique à partir d'ontologie semble actuellement l'approche la plus prometteuse pour partager et exploiter l'information sur le Web. Plus généralement, les outils d'annotation visent à améliorer communication et interopérabilité sur le Web. Ces annotations permettent d'associer des notes de lectures aux documents, de partager de l'information, d'effectuer des tâches rédactionnelles en groupe... Grâce aux systèmes d'annotation, le lecteur devient aussi rédacteur. Plus généralement, dans un contexte comme le Web, le système passe du "one-to-many" (un rédacteur et des millions de lecteurs) au "many-to-many" (tout utilisateur du Web est Lecteur/Rédacteur) (Zohar, 1999).

Notre objectif est de faire une synthèse des caractéristiques des annotations¹ (section 2) et de proposer une caractérisation des annotations sémantiques (section 3).

2 CARACTÉRISATION DES ANNOTATIONS

2.1 Description des annotations

Une annotation est une information graphique ou textuelle attachée à un document et le plus souvent placée dans ce document. Cette place est donnée par une ancre.

1. Voir aussi (Bremer, 2002; Denoue, 2000; Desmontils & Jacquin, 2002a; Garfunkel, 1999; Heck *et al.*, 1999; Perry, 2001; Vasudevan & Palme, 1999; Zohar, 1999)

Les annotations font référence à des entités diverses : un ensemble de documents, un document, un passage, une phrase, un terme, un mot, une image... Les annotations peuvent prendre plusieurs formes comme :

- des icônes (par exemple pour décrire des avis en utilisant des étoiles, des points d'interrogation...),
- des symboles de liens (pour décrire des associations, des relations entre mots, paragraphes ou chapitres),
- des notes textuelles en marge, en bas de page ou en fin de document repérées dans le texte² par des icônes (numéros, étoiles...),
- des mises en formes typographiques (surlignage, soulignage, italique...),
- des redécoupages de texte (à l'aide d'accolades, de numérotation de passages...),
- des images,
- des sons...

2.2 Dimensions des annotations

C. C. Marshall (Marshall, 1998) propose de caractériser les annotations selon différentes dimensions. Ces dimensions sont des espaces mono-dimensionnels continus. Elles décrivent les propriétés de l'annotation au niveau de sa structuration, de sa fonction et de son rôle dans la communication rédacteur/lecteur. Dans cette section, nous allons décrire ces dimensions en spécifiant les cas typiques, extrêmes, mais il est souvent possible de trouver des exemples d'annotations "intermédiaires".

2.2.1 Dimensions liées à la structuration de l'annotation

Une première classe de dimensions concerne la structuration de l'annotation. La première dimension proposée décrit le niveau de formalisation de l'annotation : c'est la dimension Formelle - Informelle. Les annotations sont représentées de manière plus ou moins structurée, i.e. plus ou moins basées sur une formalisation rigoureuse. Les notes peuvent aller du renseignement de champs spécifiques (par exemple par réponse à un questionnaire) au texte en langage naturel (mais aussi utilisation de sigles intuitifs...).

Une autre dimension concerne la signification propre de la note, c'est-à-dire entre l'explicite et le tacite. Une annotation explicite se suffit à elle-même (destinée à une autre personne que le rédacteur ne connaît pas forcément) alors qu'une annotation tacite demande une connaissance complémentaire (table de lecture...) et est destinée à un lecteur instruit des conventions adoptées (souvent le rédacteur lui-même). Ces dernières sont difficiles à interpréter, parfois même, après un certain temps, par son auteur lui-même ! Par exemple, les notes de bas de page sont le plus souvent des annotations explicites alors que les annotations de lecture sont assez souvent une codification propre au lecteur (même si C. C. Marshall a mis en évidence la possibilité de trouver des traits communs dans les comportements d'annotation).

Il faut noter que certaines annotations peuvent "se déplacer" sur les axes des dimensions au cours du temps. Par exemple, les smiley ("frimousse" en français !) étaient au départ des annotations informelles et sont devenues (à travers tout les livres sur Internet) des annotations formelles (des règles de construction "normalisées" sont proposées...).

2.2.2 Dimensions concernant les fonctions de l'annotation

Une classe de dimensions concerne les dimensions décrivant les fonctions des annotations. La première dimension concerne l'utilisation des annotations dans les processus de lecture et de rédaction. Pour aller plus loin que C. Marshall, nous pensons que les annotations ont cinq utilisations majeures :

1. *information, illustration, extension du document* ; lorsque le lecteur rédige ces annotations, il devient alors rédacteur³,
2. *forum*, en permettant à un ensemble de lecteurs de débattre sur le document,
3. *opérationnalisation de l'information*, en permettant de déterminer les concepts présents dans le document (nous verrons cela avec les annotations sémantiques en section 3),

2. Ce sont les ancres (comme pour cette note !).

3. Ce type d'annotation est présente dans beaucoup de systèmes d'annotation sur le Web. C'est une plus-value importante sur le document.

4. *aide au processus rédactionnel*, en permettant d'indiquer des consignes de rédaction (corrections, mouvements d'informations...),
5. *support de lecture* (la mise en évidence de passages importants...) permettant l'appropriation du texte par lecteur, l'annotation est alors le reflet de l'engagement par rapport à un texte d'un lecteur qui le personnalise (trace de lecture) afin de faciliter un futur retour.

C. C. Marshall propose aussi une dimension concernant le rôle de l'annotation par rapport au niveau de lecture du document. Certaines notes (hyper-liens par exemple) permettent une lecture extrêmement superficielle du document (surf). À l'opposé, des annotations sont utilisées pour une lecture approfondie d'un texte⁴. Entre les deux, il y a les notes de lecture rapide comme l'utilisation de la typographie (mise en gras ou en italique par exemple) ou le surlignage.

Remarque : D'une certaine manière, les attributs typographiques peuvent être considérés comme des annotations permettant une aide à la lecture, particulièrement le souligné, le surligné, le gras et l'italique.

C. C. Marshall propose aussi une dimension qui concerne la "durée de vie" de l'annotation, c'est la dimension temporelle. En effet, une annotation peut être permanente ou éphémère. Les annotations permanentes sont des annotations faites pour durer. Dans ce cadre, nous trouverons par exemple toutes les annotations de bas de page dans les documents... Ce sont généralement les annotations produites par le rédacteur et laissées dans le document final. À l'opposé, les annotations éphémères sont très souvent des annotations d'aide à la rédaction du document. Une spécificité des annotations de documents numériques est que l'annotation n'est pas statique. En effet, sa forme et son contenu peuvent évoluer au cours du temps. L'annotation elle-même peut être placée pour une durée de vie limitée.

2.2.3 Dimensions concernant le rôle de l'annotation dans la communication rédacteur/lecteur

Une dimension particulièrement importante concerne la relation lecteur/rédacteur, autrement dit sur le choix des destinataires potentiels de l'annotation. Une annotation peut être soit privée, c'est-à-dire que le rédacteur la destine à lui-même, soit publique. Dans ce dernier cas, plusieurs "degrés" sont possibles : le groupe de travail, l'institution... jusqu'au niveau le plus global (c'est-à-dire le monde entier !).

Le comportement du rédacteur et des lecteurs vis à vis du document et des annotations mène à des systèmes du type "one-to-many" (un rédacteur destine son document à un ou plusieurs lecteurs) ou "many-to-many" (un groupe de lecteurs/rédacteurs travaille en collaboration). Les systèmes d'annotation sur le Web visent à atteindre ce dernier cas.

Remarque : Il existe aussi des annotations, en particulier pour le support de lecture, ajoutées automatiquement au document permettant un accès plus facile que ce soit par la machine ou par l'humain. Le système XLibris (XLibris,) par exemple propose de mettre en évidence les mots représentatifs du document en utilisant la méthode statistique TFIDF. Ce même système ajoute une note aux termes déjà annotés. Celle-ci contient le résultat de la recherche des mots annotés par des moteurs de recherche classiques (Altavista...). Google surligne automatiquement les termes du document ayant fait l'objet de la recherche. Certains outils gardent aussi une trace des liens empruntés (Zeiliger, 2001), du temps passé sur les pages ou des événements souris. Ces informations seront ensuite proposées à l'utilisateur et éventuellement elle-même annotées (Zeiliger, 2001).

2.3 Exploitation des annotations

Les annotations sont multiples dans leurs formes et dans leurs fonctions. Par rapport au Web, les annotations les plus courantes sont des annotations en langage naturel (donc informelles) ou symboliques (surlignage en particulier), publiques (par définition du Web !), plutôt éphémères, explicites et le plus souvent associées à des termes. Dans la suite, nous appellerons ces annotations, des annotations libres. Elles sont le plus souvent utilisées comme support de lecture (commentaires, recherche de passages, surlignage, cartes de navigation) ou comme médium ayant rôle de forum. Dans le cadre du Web, se trouve un cas très particulier d'annotation : les annotations sémantiques. Elles sont utilisées dans le cadre de la recherche d'information, la classification, le résumé automatique et plus généralement, pour favoriser l'interopérabilité. Elles seront décrites en section 3.

Des travaux récents, comme (Denoue, 2000), proposent d'exploiter les annotations libres afin de capitaliser l'apport d'information issue de ces notes. Les annotations d'un document sont en effet exploitables

4. Cf. rôle des notes dans les œuvres de Tolkien par exemple.

pour améliorer la gestion du document. Une première utilisation des annotations peut être la constitution d'un résumé du document. En effet, les parties annotées peuvent être considérées comme importantes (positivement ou négativement) pour le lecteur. Aussi, construire un texte à partir de ces "morceaux" s'avère une méthode assez performante pour résumer le texte. Au-delà de l'indexation, dans le même ordre d'idée que le résumé automatique, (Denoue, 2000) propose une méthode de classification basée sur les termes annotés.

Les annotations sont aussi potentiellement intéressantes dans le cadre de la recherche d'information. En effet, les termes annotés peuvent être représentatifs du document. A cela, on peut ajouter l'exploitation du contenu des annotations dans l'indexation ou même au niveau de la recherche d'informations. Dans ce dernier cas, les annotations sont susceptibles de servir au niveau du retour utilisateur ("relevance feedback") pour étendre la requête ("query expansion"). En effet, l'analyse des passages annotés et des annotations permet de juger la pertinence des documents proposés et de construire d'autres requêtes plus adéquates (Golovchinsky, 1998).

Un certain nombre d'outils d'annotation sur le Web ont été proposés (Röscheisen *et al.*, 1994; Koivunen *et al.*, 2001; Yee, 1997; Imarkup Solutions, 2002; Denoue & Vignollet, 1999)... Pour des comparatifs techniques plus précis (architectures adoptées, techniques mises en œuvre, présentations des annotations choisies, types d'annotations...), il est possible de consulter un certain nombre de sites ou d'articles tels que (Garfunkel, 1999; Denoue, 2000; Bremer, 2002; Heck *et al.*, 1999; Perry, 2001; Vasudevan & Palme, 1999; Zohar, 1999; Desmontils & Jacquin, 2002a).

3 CAS DES ANNOTATIONS SÉMANTIQUES

Les annotations sémantiques sont persistantes (du moins si le document n'est pas modifié) et implicites, car faisant référence à une connaissance (habituellement une ontologie) séparée du document. Elles sont le plus souvent attachées au document et ne possèdent pas d'ancrage particulier. Elles sont des annotations opérationnelles, car elles sont destinées à être traitées par des machines (par opposition aux annotations libres en langage naturel ou composées de symboles souvent tacites). Leur objectif majeur est de *désambigüiser* le document pour un traitement automatique. Les annotations sémantiques ne sont ni publiques (pas consultables directement, il faut disposer d'un éditeur spécialisé ou regarder le code source du document) ni publiques (destinées des outils ayant accès à l'ontologie référencée) ce qui en fait un cas très particulier d'annotations.

Derrière le processus général d'annotation sémantique de documents par des ontologies se cache plusieurs phases qui font référence à des annotations de natures un peu différentes. Ces phases sont au nombre de trois :

Repérer : processus manuel (Handschuh, 2001; Maedche *et al.*, 2001; Heflin *et al.*, 2002)... ou automatique (Desmontils & Jacquin, 2002b) qui consiste à placer dans le document des références aux concepts de l'ontologie qu'il contient. Ces éléments sont considérés comme des méta-données⁵.

Instancier : processus manuel ou automatique permettant de valuer les attributs des concepts à l'aide des informations présentes dans le document (là encore, ce sont des méta-données).

Enrichir : processus manuel visant à ajouter des informations par l'intermédiaire des attributs de concepts qui n'ont pas pu être évalués à la phase précédente.

Les deux premières étapes sont des étapes d'insertion de méta-données : il n'y a pas ajout d'information mais plutôt *localisation et caractérisation de l'information* déjà présente de manière intrinsèque. Par contre, la dernière est plutôt une étape d'annotation plus classique, car il y a *ajout d'information* : le document est enrichi d'information qui n'est pas explicitement présente dans le document (il y a plus-value informationnelle). Cette annotation est directement "formalisée" par des méta-données.

Dans le cadre des annotations sémantiques, les outils sont généralement des éditeurs d'ontologies permettant de choisir une ontologie, les concepts représentant le documents et les instances des concepts présents dans le document. Les annotations sont directement insérés dans le code source du document. Ces concepts et ces instances sont soit directement parcourus par des moteurs de recherche soit utilisés pour indexer les documents.

5. Le terme "Méta-donnée" est pris dans le sens "donnée sur une donnée". Il n'y a pas ajout d'information mais plutôt désambiguïsation ou formalisation (puisque l'information est présente mais pas directement opérationnelle).

4 CONCLUSION

Les spécificités du Web en font un outil très difficile à exploiter. Les outils d'annotation sont une voie prometteuse pour l'échange et le partage d'informations. Ils permettent d'espérer atteindre un objectif primordial des concepteurs du Web à savoir un environnement collaboratif où chacun est aussi bien lecteur que rédacteur. Concrètement, les systèmes développés se sont quasiment tous soldés par des échecs. Nous pensons que ces échecs sont particulièrement dus à deux causes : une mauvaise connaissance des annotations et des utilisations que l'on peut en faire. De plus, ces annotations ont un potentiel considérable afin d'améliorer la compréhension et la manipulation des documents dans le cadre du partage d'information et de l'interopérabilité. Un cas particulier d'annotation se développe actuellement autour du Web Sémantique. Ces annotations un peu particulières, puisque plus orientées vers un traitement automatique des documents demandent encore à être spécifiées et nous pensons que les mettre dans le contexte plus général des annotations est bénéfique. Cela nous a permis de caractériser ces annotations et de décomposer le processus d'annotation en mettant en évidence des phases d'annotation différentes.

RÉFÉRENCES

- BREMER (2002). Web annotations. <http://www.db.cs.ucdavis.edu/bremer/annotations.html>.
- BUSH V. (1945). As we may think. *The Atlantic Monthly*.
- DENOUE L. (2000). *De la création à la capitalisation des annotations dans un espace personnel d'informations*. PhD thesis, Univ. De Savoie.
- DENOUE L. & VIGNOLLET L. (1999). Yawas : un outils d'annotation pour les navigateurs du web. In *IHM'99*, Montpellier, France.
- DESMONTILS E. & JACQUIN C. (2002a). Annotation et indexation de documents électroniques : application au web. Cours CNAM, Ingénierie des systèmes d'information, <http://www.sciences.univ-nantes.fr/info/perso/permanents/desmontils/CNAM/>.
- DESMONTILS E. & JACQUIN C. (2002b). *Indexing a Web Site with a Terminology Oriented Ontology*, In *The Emerging Semantic Web*, p. 181–197. IOS Press, i.f. cruz and s. decker and j. euzenat and d. l. mcguinness edition.
- GARFUNKEL J. (1999). Web annotation technologies. <http://look.boston.ma.us/garf/webdev/annotate/software.html>.
- GOLOVCHINSKY G. (1998). Emphasis on the relevant: Free-form digital ink as a mechanism for relevance feedback. In *ACM SIGIR*, Melbourne, Australie.
- HANDSCHUH S. (2001). Ontomat. <http://km.aifb.uni-karlsruhe.de/annotation/index.html>.
- HECK R. M., LUEBKE S. M. & OBERMARK C. H. (1999). *A Survey of Web Annotation Systems*. Rapport interne, Dep. Of Mathematics and Computer Science, Grinnell College, USA.
- HEFLIN J., HENDLER J., LUKE S., GASARCH C., ZHENDONG Q., SPECTOR L. & RAGER D. (2002). The shoe knowledge annotator. <http://www.cs.umd.edu/projects/plus/SHOE/KnowledgeAnnotator.html>.
- IMARKUP SOLUTIONS (2002). Imarkup. <http://www.imarkup.com/>.
- KOIVUNEN M.-R., KAHAN J., SWICK R. & PRUD'HOMMEAUX E. (2001). Annotea project. <http://www.w3.org/2001/Annotea/>. W3C.
- MAEDCHE A., MOTIK B. & VOLZ R. (2001). The karlsruhe ontology and semantic web tool suite. <http://kaon.semanticweb.org/>.
- MARSHALL C. C. (1998). Toward an ecology of hypertext annotation. In *ACM Hypertext*.
- PERRY P. (2001). Web annotations. <http://www.paulperry.net/notes/annotations.asp>.
- RÖSCHEISEN M., MOGENSEN C. & WONOGRA T. (1994). Commentor. <http://hci.stanford.edu/commentor/doc/>.
- VASUDEVAN V. & PALME M. (1999). On web annotations: Promises and pitfalls of current web infrastructure. In *HICSS*.
- XLIBRIS. FX Palo Alto Laboratory, Physical/Digital Documents research program, Fuji-Xerox Co. Ltd.
- YEE K.-P. (1997). Critlink. In *Fifth Foresight Conference on Molecular Nanotechnology*, <http://crit.org/critlink.html>.
- ZEILIGER R. (2001). Nestor : The web browser and cartographer. <http://www.gate.cnrs.fr/zeiliger/nestor.htm>. CNRS.
- ZOHAR R. (1999). *Web Annotation - an Overview*. Rapport interne, Dept. of Electrical Engineering, Israel Institute of Technology.