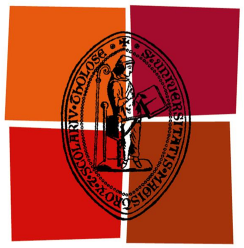


Université de Toulouse

CNRS - INPT - UPS - UT1 - UTM

THÈSE

en vue de l'obtention du

Doctorat de l'Université de Toulouse

délivré par

l'Université Toulouse 3 – Paul Sabatier

Discipline INFORMATIQUE

présentée par

GUILLAUME CABANAC

École doctorale :	Mathématiques, Informatique et Télécommunications de Toulouse
Unité de recherche :	Institut de Recherche en Informatique de Toulouse – IRIT UMR 5505 CNRS
Équipe d'accueil :	Systèmes d'Informations Généralisés

Fédération et amélioration des activités documentaires par la pratique d'annotation collective

soutenue le 5 décembre 2008 devant la commission d'examen :

JURY

Corine CAUVET	Professeur, Université Aix-Marseille 3	*rapporteure*
Max CHEVALIER	Maître de conférences, Université Toulouse 3	*co-encadrant*
Claude CHRISMENT	Professeur, Université Toulouse 3	*directeur de thèse*
Christine JULIEN	Maître de conférences, Université Toulouse 3	*co-encadrante*
Thérèse LIBOUREL	Professeur, Université Montpellier 2	*examinatrice*
Jean-Marie PINON	Professeur, INSA de Lyon	*rapporteur*
Chantal SOULÉ-DUPUY	Professeur, Université Toulouse 1	*invitée*

Mémoire composé avec LaTeX 2ε.
Version du 5 décembre 2008.

Guillaume CABANAC

Fédération et amélioration des activités documentaires par la pratique d'annotation collective

Directeur de thèse :
Claude CHRISMENT, professeur à l'université Toulouse 3 – Paul Sabatier

Résumé

Les activités documentaires couramment réalisées sur les documents papier sont aujourd'hui transposées sur leurs homologues électroniques. Ainsi, une kyrielle de systèmes permet de mener à bien les activités liées aux documents. Ils permettent notamment de rechercher de l'information utilisée pour rédiger un document qui peut être ensuite diffusé, exploité et organisé par ses lecteurs dans leur espace documentaire. Notre étude des systèmes existants a permis de révéler deux limites principales. Premièrement, un système ne répond généralement qu'à une seule, voire à deux activités. Ce cloisonnement des activités est préjudiciable à la fois pour les usagers (qui doivent maîtriser et jongler entre de nombreux outils) et pour les systèmes (qui ne possèdent qu'une représentation parcellaire des besoins des usagers). Deuxièmement, les systèmes n'exploitent pas les résultats des activités documentaires des membres organisationnels.

Notre contribution comprend deux volets. Premièrement, nous proposons un modèle fédérant les activités documentaires autour de la pratique d'annotation collective. Des processus collectifs y sont associés afin d'exploiter chaque activité documentaire pour enrichir les autres, apportant ainsi une assistance à chaque individu en tirant parti du groupe, et vice versa. Le but de cette approche originale est double : simplifier l'accès et l'appropriation des documents tout en anticipant les besoins de l'usager pour lui offrir une assistance non intrusive. Deuxièmement, nous proposons d'exploiter les espaces documentaires des membres organisationnels. Bien qu'ils contiennent des informations à haute valeur pour l'organisation, collectées au prix de coûteux efforts, ces espaces demeurent paradoxalement en sommeil. Afin de tirer parti de ces espaces documentaires, nous proposons une interface multi-facettes d'accès au capital documentaire d'une organisation. Cette interface permet l'exploration des documents et individus de l'organisation selon différents axes et niveaux de granularité. Nos propositions ont été validées par différentes expérimentations ainsi que par le développement du prototype TafAnnote qui souligne la faisabilité de notre approche fédérant les activités documentaires autour de l'annotation collective.

Institut de Recherche en Informatique de Toulouse – UMR 5505 CNRS
Université Toulouse 3 – Paul Sabatier, 118 route de Narbonne, F-31062 Toulouse cedex 9

Guillaume CABANAC

Federating and Improving Document-related Activities Through Collective Annotation Practice

Supervisor:
Claude CHRISMENT, Professor at Toulouse 3 University – Paul Sabatier

Abstract

Daily activities carried out with paper documents are nowadays transposed onto their digital counterparts. A plethora of software enable people to achieve document-related activities. In particular, these comprise information retrieval used while drafting new documents. Documents may later be disseminated, exploited and organized in readers' document repositories. Our study on current systems showed two main limitations. On the one hand, any system meets only one or at most two activities. The underlying activity compartmentalization is detrimental to users—who have to master and juggle several systems—as well as to systems—having partial knowledge of users' needs. On the other hand, systems do not harness the organizational members' document-related activities.

The proposed contribution is twofold. Firstly, we designed a model for federating the document-related activities through collective annotation practice. Associated with this model are collective processes intending to give each activity the benefit of the other ones. This also fosters inter-user benefit as people take advantage of the group and vice versa. Actually, the purpose of the proposed approach is twofold: simplifying document access and appropriation while anticipating individuals' needs to offer them unintrusive assistance. Secondly, our approach exploits the organizational members' document repositories. Although they do contain highly valuable information being collected with a lot of efforts, they paradoxically remain dormant. With the aim of harnessing these information sources, we designed a multi-faceted interface for accessing any organization's document resources. This interface allows the exploration of documents as well as users of these documents, according to various dimensions and granularity levels. Our proposals were validated through several experiments and the TafAnnote prototype development. They demonstrate the feasibility of our approach which federates document-related activities with collective annotation practice.

Institut de Recherche en Informatique de Toulouse – UMR 5505 CNRS
Université Toulouse 3 – Paul Sabatier, 118 route de Narbonne, F-31062 Toulouse cedex 9

Remerciements

À L'ISSUE de ce travail de doctorat, c'est avec grand plaisir que je me remémore mon expérience à l'université. Tout au long de cette aventure captivante, j'ai rencontré et travaillé avec des personnes aux qualités humaines et intellectuelles précieuses. Je désire ici leur témoigner ma reconnaissance et leur affirmer à quel point ils m'ont apporté.

Je tiens tout d'abord à remercier les membres extérieurs du jury. Ainsi, ma reconnaissance va à Madame Corine Cauvet, professeur à l'université Aix-Marseille 3 et à Monsieur Jean-Marie Pinon, professeur à l'INSA de Lyon qui m'ont fait l'honneur d'être rapporteurs de ce mémoire. Je remercie également Madame Thérèse Libourel, professeur à l'université Montpellier 2, d'en être l'examinatrice. Je les remercie pour leur évaluation scientifique et leur travail de synthèse. Qu'ils soient assurés de mon très grand respect.

Je dois beaucoup à Monsieur Claude Chrisment, professeur à l'université Toulouse 3, directeur adjoint de l'IRIT et co-responsable de l'équipe SIG. En qualité de directeur de recherche, il m'a fait bénéficier de son recul sur de nombreux domaines de l'informatique fondamentale comme appliquée. Disponible et rigoureux, il a permis l'aboutissement de ces travaux. J'ai particulièrement apprécié nos fréquents échanges dont j'ai pu tirer de nombreux enseignements. Je tiens ici à lui témoigner mon admiration, ma gratitude et mon profond respect.

Monsieur Max Chevalier, maître de conférences à l'université Toulouse 3, co-encadrant de cette thèse, a été pour moi un guide hors pair. Passionné par ses activités d'enseignant comme de chercheur, animé d'une motivation inaltérable, exigeant avec les autres comme avec lui-même, il a contribué grandement à la réalisation de ces travaux. Qu'il soit assuré de ma reconnaissance pour son soutien et ses nombreux encouragements, ainsi que du plaisir que j'ai à travailler avec lui.

Je suis reconnaissant envers Madame Christine Julien, maître de conférences à l'université Toulouse 3, co-encadrante de cette thèse, à plusieurs égards. Alors que j'étais étudiant à l'IUT informatique de Rangueil, c'est en grande partie à son contact que j'ai pris conscience de mon projet professionnel : devenir enseignant-chercheur. Sa pédagogie, sa disponibilité et sa gentillesse ont été pour moi sources de motivation. C'est également grâce à Christine que j'ai découvert l'IRIT, où j'ai effectué mon stage de DUT co-encadré par Max Chevalier, qui était alors doctorant...

Je souhaite remercier Madame Chantal Soulé-Dupuy, professeur à l'université Toulouse 1, co-responsable de la composante *Documents, données semi-structurées et usages* dans laquelle s'inscrit mon travail, pour l'intérêt et le regard critique qu'elle a toujours portés à mes recherches. À ses côtés, la rédaction d'un chapitre d'ouvrage a été propice à une réflexion approfondie sur notre domaine de recherche. Je tiens à lui exprimer mes remerciements pour l'honneur qu'elle me fait en participant à ce jury.

Ma gratitude va également à Claudette Cayrol, professeur à l'université Toulouse 3, et à Marie-Christine Lagasquie-Schiex, maître de conférences à l'université Toulouse 3. Alors que je cherchais à formaliser le concept de validation sociale d'annotation, j'ai bénéficié de leur expertise concernant la théorie de l'argumentation bipolaire. Je tiens à les remercier pour leur disponibilité et leurs conseils avisés.

Dans le cadre de mon doctorat, j'ai eu l'opportunité de collaborer sur le thème de l'annotation avec Franck Ravat, maître de conférences HDR à l'université Toulouse 1 et Olivier Teste, maître de conférences à l'université Toulouse 3. Ces travaux ont suscité des échanges constructifs, stimulants et fructueux, tout en resserrant un lien amical déjà fort. Concernant l'enseignement, je suis heureux d'avoir pu contribuer à l'environnement *CompAlgo* avec celles et ceux qui m'ont initié à l'informatique à l'IUT de Rangueil.

Un grand merci à ceux qui font du laboratoire un lieu de travail convivial : la très grande « famille » SIG où Claude Chrisment et Gilles Zurfluh m'ont accueilli. J'ai pu y nouer des liens avec Josiane Mothe, Florence Sèdes, Mohand Boughanem et Bernard Dousset, notamment au gré de nos « missions » en conférence. Je souhaite tout particulièrement remercier Gilles Hubert pour sa complicité, son humour et son regard éclairé. Au quotidien, l'ambiance studieuse et détendue du bureau ne serait pas la même sans Karim Djemal, Moultazem Ghazal, Ronan Tournier et les nouvelles recrues : Hamdi Chaker et Arlind Kopliku. Dans ce contexte masculin, les visites de Dana Al-Kukhun, Ilhem Ghalamallah, Éloïse Loubier, Karen Pinel-Sauvagnat et Bouchra Soukkarieh apportent une touche féminine salutaire ☺.

Je ne voudrais pas oublier le personnel de l'IRIT qui veille au bon fonctionnement du laboratoire. En particulier, je remercie Martine de Calmès et Brigitte Marchiset, mes Pythies pour lesquelles l'administration d'Oracle n'a plus aucun secret. Disponibles, compréhensives et toujours souriantes, elles m'ont toujours encouragé et je leur en suis très reconnaissant.

Au fil de ces quatre années passées à l'IRIT, j'ai eu la chance d'être entouré de personnes généreuses et passionnées. Je remercie Christine Maurel, Frédéric Migeon ainsi que Jean-Paul Arcangeli, Franck Morvan et Pascal Will pour les bons moments que nous passons au quotidien. Nos discussions du déjeuner, souvent désinvoltes mais parfois très sérieuses, sont autant d'escapades dont j'ai besoin. Contribuent à cette atmosphère mes amis, avec qui je partage le même *Radeau de la Méduse* depuis le DUT : Vincent Forest et son légendaire humour noir ainsi que Sylvain Rougemaille dont j'admire la répartie. Une pensée au quatrième mousquetaire, Olivier Rouhaud, qui a choisi la voie de la liberté et de l'aventure au Mexique. Tous mes vœux de réussite à Jérémy Philippeau qui incarne à mes yeux joie de vivre et indépendance.

Je souhaite également exprimer à mes amis à quel point leur présence m'est indispensable. Nous vivons des moments forts au fil du Canal du Midi à vélo avec Delphine et Cyrille, dans les caves de Cahors avec Sandrine et Cédric, en suivant le tour du monde du développement durable grâce à Arnaud. Je remercie également mes amis d'enfance, Charlène et Xavier, pour leur affection. Une pensée particulière va à mes voisins de cité universitaire Loïc et Stécy, mes premiers élèves de Turbo Pascal et Visual Basic ☺. Une pensée à Tafanor, évidemment.

Je remercie toute ma famille et ma belle-famille, notamment mon père à qui j'envie la curiosité, ma mère avec qui j'aime refaire le monde et mon oncle pour l'intuition qu'il a eue en m'offrant mon premier ordinateur. Quant à ma sœur, j'attends impatiemment l'ouverture de son propre salon de coiffure afin de lui développer le logiciel de gestion *Beauty Permanenty* !

Enfin et surtout, je remercie Claire à qui je dédie cette thèse. Elle a su me comprendre, m'épauler dans les moments difficiles, faire preuve de patience et d'un amour sans réserve. C'est à ses côtés que j'ai franchi cette étape, il me tarde à présent d'envisager les suivantes... TLNE !

Ces travaux de thèse ont été financés par l'allocation de recherche n° 16728–2005 et le contrat d'Attaché temporaire d'enseignement et de recherche n° 27 ATER 2118 de l'université Toulouse 3.

Table des matières

Introduction générale

> "This might not sound very radical but the mindset in Computing for the past 20 years has been on the value of storing information in large databases for selective retrieval. And because computers have opened up a new way to store vast amounts of "information" in a disembodied state and ship it around over faster and faster networks, people have come to believe that the more information you can store or ship, the better off you or your organisation are. We have confused what we can write down with what we usefully know and compounded the error by supposing that because computers can help us write down more they can obviously help us know more."
>
> *Alison Kidd (1994)*

Contexte de travail

Les travaux présentés dans ce mémoire s'inscrivent dans le contexte des Systèmes d'Informations sur lesquels s'appuient des organisations : entreprises, laboratoires de R&D ou communautés, au sens large. Toute organisation regroupe des individus qui œuvrent pour atteindre les objectifs fixés par son comité de pilotage. Pour ce faire, la plupart des membres organisationnels rassemblent, exploitent et créent au quotidien des documents, aussi bien sur support papier qu'électronique.

Bénéficiant des progrès continus dans divers domaines tels que l'électronique et les réseaux, l'informatique est désormais au cœur de chaque organisation. Or, l'informatisation des organisations puis l'avènement d'Internet et des nouvelles technologies de l'information et de la communication n'ont eu de cesse que de transformer les activités des individus. Ils ont alors pu en bénéficier et gagner en efficacité car de nombreuses limites des documents papier sont estompées lorsqu'on considère leurs contreparties sur support électronique. Grâce à cette « redocumentarisation du monde » (Pédauque, 2006, 2007) et à la dématérialisation des échanges, la rédaction et la diffusion planétaire des documents sont désormais une réalité accessible à tout un chacun.

Problématiques

En observant l'informatisation des organisations et la dématérialisation des documents, d'aucuns ont peu à peu élaboré le « mythe du bureau sans papier » critiqué par Sellen et Harper (2003), où les documents électroniques auraient totalement supplanté le papier. Or, l'observation des pratiques des individus par ces mêmes auteurs montre leur résistance au changement : ils persistent à imprimer les documents électroniques, notamment pour retrouver le contact physique du papier ! Certaines activités plus difficiles à réaliser sur écran telles que la lecture, la réflexion à l'aide d'annotations, etc. en seraient-elles la cause ? La transposition du papier au numérique, bien qu'apportant des avantages indéniables, n'aurait-elle pas dans le même temps dégradé certaines activités documentaires ?

Par ailleurs, le Système d'Informations (SI) de l'organisation profite pleinement des avancées technologiques : tous ses postes sont interconnectés par des réseaux aux débits croissants, l'adoption massive de terminaux miniaturisés (PDA, mobiles, etc.) fait de l'informatique pervasive une réalité, la standardisation des formats de données au travers de protocoles consensuels permet l'interopération entre organisations (B2B)… Toutefois, les informations extraites, filtrées, consolidées et organisées par et pour chaque membre organisationnel au niveau « microscopique » ne sont que peu valorisées au niveau « macroscopique » de l'organisation. De fait, les questions suivantes sont de réelles problématiques, des verrous pour les organisations, malgré l'environnement hautement numérique dont elles disposent de nos jours. Par rapport à la thématique du projet auquel je suis affecté, quels documents ont été capitalisés par mes collègues ? En fonction des activités de mon organisation, que puis-je lire en complément des documents que je possède déjà ? Quels sont les documents-clés de l'organisation ? Sur quelles références documentaires s'appuie telle équipe pour mener à bien ses activités ? Dans un contexte où le *turnover* s'accentue, quelles connaissances doit-on renouveler suite au départ de telles ou telles personnes ?

Contributions

La contribution de ce mémoire vise à fédérer et améliorer les activités documentaires électroniques au sein d'une organisation. Pour ce faire, nous définissons un modèle unifié des activités documentaires couvrant le cycle de vie du document (Sellen et Harper, 2003, p. 203). L'élément fédérateur au cœur de ce modèle est l'annotation collective, provenant de l'activité d'annotation papier et bénéficiant des capacités de traitement et de communication des systèmes informatiques. À ce modèle sont adjoints six processus visant à tirer parti du groupe pour l'individu, et vice versa. Ces processus exploitent le capital documentaire organisationnel formé par les espaces d'information des individus, que ces derniers gèrent et organisent au prix de coûteux efforts cognitifs. Cette exploitation repose sur le principe du donnant-donnant, tout en évitant la non-intrusion pour limiter la résistance au changement des usagers. Enfin, ces processus tirent parti de chaque activité documentaire (le classement de documents, par exemple) pour améliorer les autres activités documentaires de l'usager et de ses collègues (la recherche d'information, par exemple), apportant ainsi un enrichissement mutuel au sein de l'organisation.

Au niveau « macroscopique » le capital documentaire organisationnel est mis à profit par l'architecture reposant sur le modèle unifié des activités documentaires. L'objectif est de valoriser

tout document introduit dans l'organisation, alors que les espaces personnels d'information demeurent actuellement en sommeil. En effet, les usagers bénéficient rarement de telles sources d'informations à forte valeur ajoutée (qui sont pourtant internes à l'organisation) et se tournent à défaut vers d'autres ressources *a priori* moins pertinentes et structurées, telles que le Web. Cette situation implique de fait un retour sur investissement faible pour chaque usager et pour l'organisation en général. En réponse à cette méconnaissance des ressources et compétences voisines, nous concevons une interface multi-facettes, permettant la visualisation et l'exploration du capital organisationnel selon deux dimensions (les individus et les documents) et deux mesures de similarité complémentaires (sur le contenu et sur l'usage).

Dans une démarche globale de recherche nous validons les propositions présentées, notamment au travers d'une expérimentation « écologique » réalisée avec le concours de 121 volontaires qui ont participé en ligne, grâce à une plate-forme développée à cet effet. Enfin, nous démontrons la faisabilité technique de nos propositions par le développement du prototype de recherche TafAnnote, « preuve de concept » asseyant la contribution originale détaillée dans ce mémoire.

Organisation du mémoire

La partie I présente le contexte des activités documentaires réalisées par les individus au sein des organisations. Nous détaillons les caractéristiques de ces activités sur support papier comme sur support électronique, puis nous focalisons sur l'activité transversale d'annotation. Enfin, nous mettons en lumière les problématiques des activités documentaires sur support électronique au sein d'une organisation.

La partie II détaille notre proposition : une architecture de système pour fédérer et exploiter les activités du cycle de vie du document. Au cœur de notre proposition figure la pratique commune d'annotation, car elle s'insère dans chacune des activités et représente une réelle plus-value apportée aux documents par et pour les individus (Kidd, 1994). En effet, la trace d'une annotation sur un document reflète les efforts cognitifs de son créateur alors qu'il interagissait avec ce vecteur d'information : finalités d'apprentissage, d'argumentation, de correction, etc. Outre cette architecture reposant sur l'annotation, une assistance personnalisée est offerte aux membres organisationnels grâce à six processus définis sur les principes du donnant-donnant et de la non-intrusion. Nous exposons enfin la conception de l'interface multi-facettes d'accès au capital organisationnel.

La partie III présente les expérimentations réalisées afin de valider l'architecture et les processus proposés. Elle détaille également le prototype logiciel « preuve de concept » TafAnnote qui met en œuvre cette architecture pour en démontrer la faisabilité.

Enfin, nous discutons la contribution présentée dans ce mémoire, notamment les implications de la fédération pour l'usager, avant de conclure en évoquant les pistes de recherche que nous souhaitons considérer à l'avenir.

Première partie

Contexte : les activités documentaires au sein d'une organisation

1 Les activités documentaires *papier* au travers du cycle de vie du document

« Comme l'apprend vite tout bon rédacteur, c'est justement ce qui est évident qui doit être souligné — sinon on passera à côté. »

Peter Ferdinand Drucker (1909 — 2005)

PAR LE TERME « ORGANISATION » nous désignons tout groupe d'individus mettant en commun leurs connaissances et compétences pour atteindre un but commun lié à la production de savoirs au sens large. Ainsi, nous considérons aussi bien les entreprises, les laboratoires de R&D… que toute communauté, qu'elle soit réelle ou bien virtuelle. Selon Kidd (1994) trois catégories principales d'individus sont à l'œuvre dans ces organisations :

- les « travailleurs du savoir » (*knowledge workers*), concept forgé par Drucker (1959) pour désigner les individus employés pour assimiler et produire de la connaissance à partir des informations qu'ils consultent. En présence des mêmes éléments d'information, deux individus distincts produiront un résultat distinct en fonction de leurs connaissances et expériences différentes. Typiquement, ils travaillent dans le design, la publicité, le marketing, le management, la communication, la justice, la finance, la recherche…
- les « travailleurs communiquants » (*communications workers*) qui sont des amplificateurs d'information, employés pour la collecter et la diffuser au mieux. Contrairement aux *knowledge workers* ils ne modifient pas les informations qu'ils trouvent. Établir des relations et influencer d'autres personnes sont leurs principales motivations ;
- les « employés de bureau » (*clerical workers*) réalisent leurs activités à partir des informations qui ne proviennent pas d'eux-mêmes, des contrats d'assurance par exemple. Les résultats fournis pas deux individus varient très peu. Leur motivation personnelle consiste à être indispensables au fonctionnement efficace de l'organisation.

À l'heure de la société de l'information, Sellen et Harper (2003, p. 51) rapportent une étude indiquant que la catégorie des *knowledge workers* représentait déjà 31 % de la population active aux

États-Unis en 1995, cette proportion devant continuer de croître de manière significative au XXI^e siècle. En fait, la distinction entre ces trois catégories tendrait à s'atténuer avec la démocratisation de l'informatique, à tel point que Ballay (2002) suggère que « nous sommes tous des travailleurs du savoir ». Les travaux rapportés dans ce mémoire considèrent la relation entre les individus (perçus en tant que travailleurs du savoir) et les documents au sein de l'organisation. Pour illustrer les activités documentaires qu'ils réalisent au quotidien, ce chapitre s'appuie sur le cycle de vie du document proposé par Sellen et Harper (2003, p. 203) et reproduit en figure I.1.1. Il comprend les six activités que nous détaillons dans les sections suivantes ; par la suite, nous y ferons référence au travers des symboles de ① à ⑥.

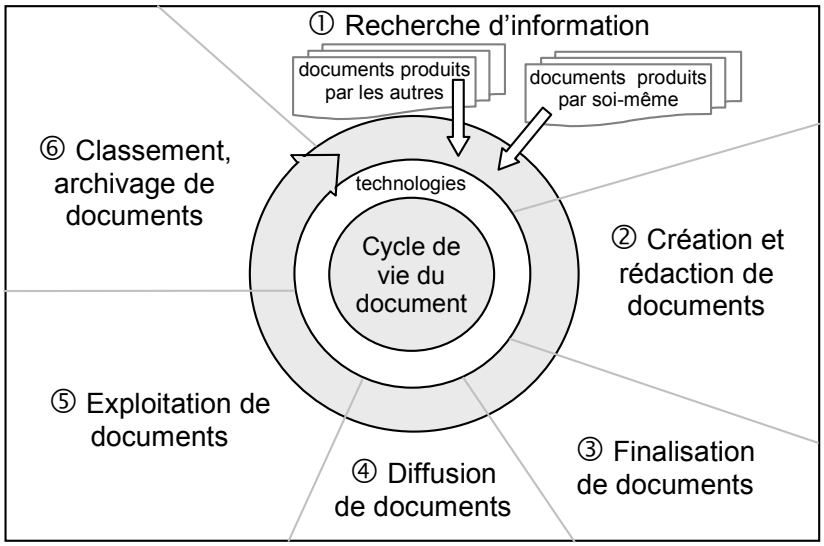


Figure I.1.1 – Les six activités du cycle de vie du document (Sellen et Harper, 2003, p. 203).

Malgré la fin annoncée du papier dans les organisations modernes, les individus consacraient encore récemment 86 % de leur temps de travail à des activités impliquant des documents papiers, dont 35 % en conjonction avec des documents électroniques (Sellen et Harper, 2003, p. 56). Une autre observation illustre la réticence précoce au « tout numérique » : Greengard (1999) rapporte que l'introduction du courrier électronique dans une organisation aux États-Unis entraînait une surconsommation de papier de l'ordre de 40 % en moyenne.

Au regard de ces chiffres, nous nous interrogeons sur les raisons qui poussent les individus à passer de l'électronique au papier. Autrement dit : quelles pratiques documentaires perdons-nous en présence d'un document électronique, nous forçant alors à l'imprimer ? Afin de comprendre les avantages et inconvénients des documents sur support papier (abrégé en « documents papier » où papier désigne le support) *versus* sur support électronique (abrégé en « documents électroniques »), nous exposons dans ce chapitre les activités documentaires *papier* avant d'aborder dans le chapitre I.2 le support *électronique*.

Les éléments rapportés dans ce chapitre proviennent principalement de l'étude ethnographique (Sellen et Harper, 1997) des pratiques documentaires de 138 employés pendant six mois au sein du Fonds Monétaire International (FMI), basé à Washington, aux États-Unis. À l'époque de cette étude, le FMI comprenait 3 000 employés dont 900 économistes. Tous les employés étaient dotés d'ordinateurs en réseau, chaque économiste bénéficiant d'un ordinateur portable associé

à une station d'accueil connectée à une imprimante personnelle. Cette étude est également rapportée dans le chapitre *"Paper in Knowledge Work"* de (Sellen et Harper, 2003, ch. 3). Nous n'avons pas été en mesure d'identifier d'étude plus récente — même de moindre ampleur — sur la relation individus-documents papier. Toutefois, les moyens dont disposait le personnel du FMI à l'époque plaçait cette organisation à l'avant-garde de la technologie et nous pensons que les éléments rapportés ont encore une valeur de nos jours pour la majorité des organisations — qui sont plus modestes que le FMI.

1.1 Acquisition d'informations ① à partir de documents *papier*

L'acquisition d'informations à partir de documents papier implique forcément une interaction physique qui repose sur les *affordances* du papier. Ce concept introduit par le psychologue James J. Gibson (1979) désigne les propriétés physiques d'un objet utilisées par l'individu qui les perçoit ; en fait ces propriétés déterminent les actions possibles sur ledit objet. Concrètement, les propriétés physiques du papier (fin, léger, poreux, opaque, flexible, etc.) suggèrent les actions humaines de saisie, de transport, de pliage, d'écriture…

Pour la tâche ① donc, l'individu peut rassembler et agencer plusieurs documents sur son bureau et les parcourir rapidement pour en obtenir une vision globale. Ayant passé en revue les documents, il peut les organiser physiquement sur la surface de travail : faire des piles selon un critère intrinsèque tel que la thématique ou la provenance. Lors d'une lecture approfondie, il peut passer d'un document à l'autre, ou même en consulter d'autres sans perdre la vision globale de cet espace documentaire. Enfin, il peut également identifier des passages intéressants et s'approprier les documents en y inscrivant une marque (astérisque, surlignement, etc.) ou en écrivant un commentaire dans la marge. Sa pratique d'annotation lui est profitable pour la tâche ②, de plus certains individus valorisent également les annotations des lecteurs qui les ont précédés, en privilégiant les exemplaires de documents qui contiennent le plus d'annotations (Marshall, 1998).

1.2 Création ② et finalisation ③ de documents *papier*

Une feuille de papier, un crayon et une gomme sont les outils fondamentaux suffisant à la rédaction sur papier. Ils permettent d'élaborer le plan d'un document en disposant les différents items sur la feuille, en mettant en exergue les liens entre les sections à l'aide de différentes marques, telles que des flèches notamment. Une fois le plan ébauché, la rédaction est communément réalisée à l'aide d'un traitement de texte. Ses fonctionnalités telles que les outils de mise en forme, le copier-coller, les correcteurs orthographique et grammatical, etc. facilitent grandement la tâche de création. Dans leur étude des pratiques de création de documents, Sellen et Harper (2003, p. 65) notent que les individus saisissent des documents au format électronique et recourent au papier en phase de relecture, pour pouvoir formuler leurs corrections sur papier. Cette stratégie est encore plus affirmée lorsque les individus collaborent à la révision d'un document qu'ils ont rédigé collectivement : ils passent alors 82 % de leur temps sur papier.

1.3 Diffusion ④ des documents *papier*

Les documents papier revêtent un rôle important au sein d'une organisation, en particulier concernant la communication entre ses différents membres. Sellen et Harper (2003, p. 53) ont observé qu'ils tendent à être imprimés et remis en main propre plutôt que d'être envoyés sous forme électronique. L'observation de la remise des rapports au sein du FMI) leur a permis d'identifier quatre raisons à ce comportement :

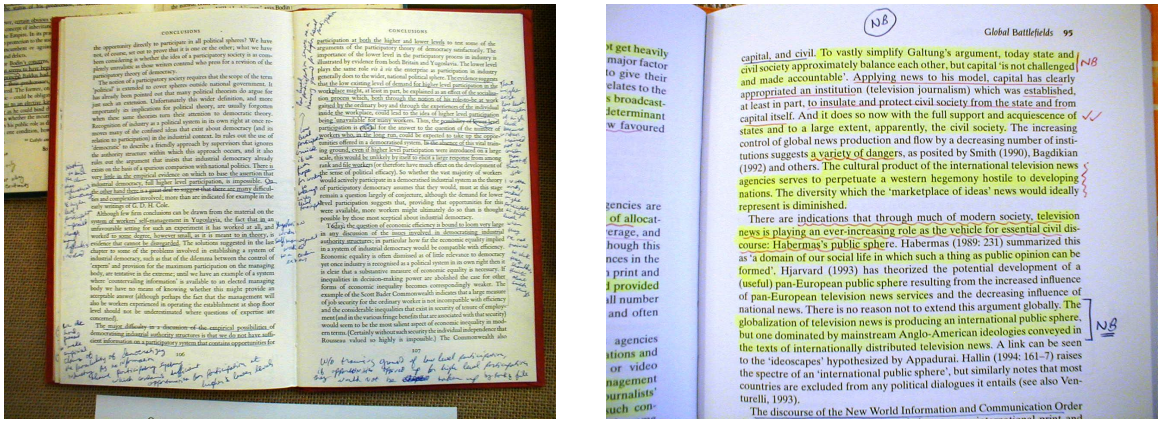
1. pour *accompagner le document de précisions* telles que la durée accordée pour réaliser une relecture, des problèmes inhabituels soulevés dans le rapport, son contexte général, etc.
2. par *preuve de déférence* envers le lecteur à qui l'on confie une relecture et pour souligner l'importance du rapport ;
3. pour *personnaliser les relations* entre individus, où la remise du rapport papier crée un contexte favorable à l'humanisation des interactions humaines ;
4. pour *s'assurer de la réception* d'un document et faire en sorte que sa présence physique sur le bureau du destinataire attire son attention et lui rappelle la tâche qu'il doit réaliser.

La diffusion sur support papier semble être adaptée pour la remise de rapports à des fins de relecture, où les destinataires sont clairement identifiés. Par contre, le moyen de diffusion papier paraît inefficace pour transmettre des documents en ciblant les personnes qui pourraient en avoir besoin, en fonction de leurs thématiques et de leurs besoins, par exemple. En effet, cela demanderait de connaître les besoins des membres organisationnels en temps réel, ainsi que des tâches coûteuses de manutention (photocopier, mettre sous pli, distribuer, etc.).

1.4 Exploitation ⑤ des documents *papier*

L'observation par Adler *et al.* (1998) des pratiques documentaires de quinze professionnels issus de domaines très variés (médecine, justice, architecture, protection civile, etc.) montre que la lecture représente 70 % de leurs activités documentaires. Pour la majorité des sujets, cette activité s'accompagne d'écriture à hauteur de 75 % à 91 % du temps de lecture. La mise en œuvre de ces activités complémentaires est motivée par la création et la mise à jour de documents (18 %) mais aussi et surtout par l'annotation et la prise de notes (48 %). Ce dernier cas s'apparente au concept de « lecture active » défini par Adler et van Doren (1972, p. 4) où le lecteur enrichit, filtre, met en exergue, résume et organise ce qu'il lit de façon critique à l'aide d'annotations (Price *et al.*, 1998). Ces dernières interrompent une lecture linéaire en connectant des passages d'un document, ce qui leur confère des caractéristiques hypertextuelles (Marshall, 1998). La figure I.1.2 illustre le résultat de l'activité d'annotation sur papier[1], où les lecteurs emploient une grande variété de marques et de symboles caractéristiques tels que « ✓ » et « ~~~ ». Kidd (1994) souligne également que les *knowledge workers* prennent de nombreuses notes à la fois pendant les réunions auxquelles ils participent mais aussi lorsqu'ils réfléchissent à un problème et structurent leurs idées une fois seuls. Notons que le chapitre I.3 est consacré à l'étude approfondie de l'activité d'annotation papier et électronique.

1. Exemples issus du site Web *"Marginalia and other crimes"* cf. `http://www.lib.cam.ac.uk/marginalia`.



(a) (b)

Figure I.1.2 – Deux livres annotés provenant de la bibliothèque universitaire de Cambridge.

Le support papier offre une grande souplesse pour l'exploitation des documents. En effet, le lecteur peut transporter ses documents et les consulter (même) en situation de mobilité, pouvant par exemple lire et annoter un livre dans les transports en commun. Par ailleurs, les affordances des documents papier permettent de les feuilleter à la recherche d'un passage précis, mais aussi de les disposer côte à côte si nécessaire. Enfin, la pratique d'annotation permet aux individus de s'approprier le contenu des documents par la formulation de diverses marques et de commentaires en contexte, au plus près du passage ciblé. Du reste, ces annotations peuvent être formulées pour soi mais aussi à destination d'autres personnes pour réaliser une synthèse collective notamment.

1.5 Classement et archivage ⑥ de documents *papier*

L'ouvrage édité par Jones et Teevan (2007b) synthétise les dernières recherches dans le domaine de la gestion personnelle d'information (*Personal Information Management*). D'après Jones et Teevan (2007a, p. 10) cette communauté s'accorde sur les définitions suivantes.

- **Espace Personnel d'Information (EPI).** Chaque individu possède un seul *Personal Space of Information* qui rassemble tout les items d'information qu'il possède. Concrètement, l'EPI d'une personne est constitué de ses livres et documents papier, ses courriels éventuellement issus de différents comptes de messagerie, ses documents électroniques et autres fichiers…
- **Collections Personnelles d'Information (CPI).** L'EPI d'un usager est partitionné en plusieurs *Personal Information Collections* définies par et pour ses activités. L'individu peut recourir à différentes organisations pour chaque CPI ; les informations d'une CPI partagent une même structure et cohérence. Des exemples de CPI incluent les différentes piles de documents posées sur un bureau (éventuellement classées selon des critères différents), les documents rangés dans des classeurs, ceux qui sont stockés dans les boîtes à archives, etc.

D'après la synthèse de Jones (2007) au sujet de l'organisation des CPI, face au choix de conserver ou pas un document les individus doivent anticiper son utilité au regard de leurs activités courantes et futures. Cette décision fait intervenir de multiples considérations, ce qui en fait une tâche difficile et sujette à erreurs. Une fois la décision prise, l'individu retient usuellement une des deux

stratégies suivantes : le classement ou l'empilement. Empiler est relativement rapide et facile, de plus les documents empilés sont visibles et accessibles. Par contre, le nombre de piles est limité par l'espace physique disponible (la surface du bureau, par exemple) et la mémoire de l'individu. D'autre part, classer demande un effort cognitif plus important car l'individu doit identifier le classeur le plus approprié pour le document considéré. Le nom du classeur, censé évoquer sa raison d'être, est souvent vague et évolue au fur et à mesure de son utilisation. Par contre, le nombre de documents classables est plus important et y accéder est plus facile que lorsqu'ils sont empilés. Au sujet des *knowledge workers*, Kidd (1994) remarque qu'ils peinent à classer les documents tant qu'ils n'ont pas pris connaissance de leur contenu. Paradoxalement ils n'ont plus besoin de les classer une fois qu'ils en connaissent le contenu car ils se réfèrent plutôt aux annotations qu'ils ont rédigées lors de leur lecture… En outre, certains apprécient de consulter les documents empilés il y a longtemps et oubliés depuis car ils perçoivent alors de nouvelles relations entre les éléments et les exploitent différemment, en fonction de leur expérience acquise entre-temps.

Concernant l'archivage des documents personnels, Kaye *et al.* (2006) ont réalisé une étude ethnographique avec 48 participants universitaires. Leur analyse montre qu'un individu conserve en général le même critère d'organisation pour toutes ses CPI : alphabétiquement, thématiquement, chronologiquement, par exemple. Ils dégagent cinq buts principaux de l'archivage :

1. *retrouver les documents plus tard.* Les individus ont tendance à empiler et à repousser le classement. Lorsqu'ils classent, aucune organisation préférée (alphabétiquement…) ne se dégage, tant elles sont personnelles et variées. Enfin, ils éprouvent des difficultés à atteindre les documents lorsqu'ils en ont besoin ;
2. *construire un héritage* (legacy). Certains individus archivent les documents car ils matérialisent pour eux « une vie de travail » qu'ils pourront transmettre à leurs successeurs. L'archive ainsi constituée a également pour but latent la matérialisation de la personnalité de leur propriétaire : travailleur, organisé, etc.
3. *partager les documents.* L'étude révèle que de nombreux sujets archivent les documents tout en donnant accès à un index électronique partagé afin que d'autres personnes puissent bénéficier de leur archive ;
4. *peur de perdre des documents.* L'archivage dans des endroits distincts et en plusieurs exemplaires permet d'éviter la perte de documents importants, en cas de feu notamment ;
5. *construction de l'identité.* La consultation de l'archive renforce la perception de soi, tout en imposant aux autres cette identité que l'on s'applique à construire incrémentalement.

1.6 Bilan des activités documentaires *papier*

Ce chapitre a mis en lumière les principaux avantages et inconvénients des documents papier. Leurs *affordances* permettent aux individus de les manipuler physiquement, de les lire n'importe où tout en les annotant, ce qui facilite l'appropriation de leur contenu. De telles annotations peuvent être destinées à soi mais aussi à d'autres personnes. Par contre, la création et la diffusion des documents papier est limitée, ainsi que la rédaction collective dans un contexte de travail à distance. Enfin, le classement et la recherche de documents entraînent également des efforts importants de la part des individus. Afin de répondre à ces derniers points, les documents des or-

ganisations, existant autrefois sur support papier sont désormais majoritairement dématérialisés. Le chapitre suivant étudie cette transposition du papier au numérique.

2 Les activités documentaires *électronique* au travers du cycle de vie du document

« Naguère, le document était un objet : passeport, pièce à conviction, livre, article, dossier d'archives, ou toute autre trace, solidaire de son contenu. La trace était validée par sa matérialité. Tout reposait sur la solidité de ce couple, aussi indissociable que le sont le signifiant et le signifié chez de Saussure. Voici que l'essor de la production numérique nous oblige à décadenasser ce trésor, à déboucher ce flacon d'où la potion, magique mais volatile, s'échappe. En pulvérisant les contenus et en traversant les supports, l'information numérique a fait valser cette croyance qui semble désormais naïve. Le document n'est plus enclos dans une enveloppe cachetée, pas plus que l'âme humaine ne s'enferme dans un tonneau. La question que pose Roger T. Pédauque est donc d'actualité : que devient la notion de *document* à l'heure du numérique ? Elle est d'autant plus pertinente que jamais la société n'en fit si grand usage. Nous sommes tous, partout, en toutes circonstances, *documentés*. »

Michel Melot (1943 —), Préface *de (Pédauque, 2006, p. 11)*

ÉTANT DONNÉS les progrès de l'informatique tout au long de la seconde moitié du XX^{e} siècle, les ordinateurs ont eu tôt fait de devenir la pierre angulaire de toute organisation. Les documents organisationnels papier qui nécessitaient jusqu'alors une manutention coûteuse tout en impliquant un faible rendement ont été progressivement transposés sur support électronique. Cette dématérialisation et l'exploitation des capacités des ordinateurs ont permis l'automatisation de traitements répétitifs sur des quantités de données croissantes : classement, interrogation, etc. De nos jours, les documents électroniques font désormais partie de notre quotidien.

Ce chapitre revisite le cycle de vie du document (figure I.1.1) dans le contexte électronique. Nous détaillons comment les applications (indifféremment appelées « systèmes » ou « logiciels »

par la suite) apportent des réponses pertinentes aux limites des activités documentaires papier identifiées dans le chapitre précédent, notamment la création, le classement et la recherche. À l'inverse, nous présentons aussi les limites imposées par la transposition du papier au numérique, dues en partie à la perte des *affordances* du support papier. En effet, certaines activités sont désormais plus difficilement réalisables, nécessitant alors une rematérialisation (l'impression du document). Enfin, nous mentionnons également les limites des activités documentaires électroniques dans le contexte organisationnel, où les documents sont gérés individuellement au prix de coûteux efforts mais que trop peu capitalisés au niveau « macroscopique » de l'organisation, entraînant un faible retour sur investissement.

2.1 Acquisition de documents ① électroniques

Les documents électroniques peuvent être regroupés dans des sources d'information internes ou externes à l'organisation. Par exemple, la base documentaire et l'intranet d'une organisation sont des sources internes alors que le Web est une source externe. Les Systèmes de Recherche d'Information (SRI) facilitent l'accès aux documents d'une source d'information ① en offrant aux usagers deux modalités qu'ils alternent inconsciemment (Agosti, 1996), leur permettant alors de satisfaire leurs besoins en information :

- *la recherche par interrogation* grâce à un moteur de recherche auquel l'usager soumet une requête. Le système recherche les mots de la requête dans les documents du corpus composant la source d'information interrogée. Puis, il restitue à l'usager la liste des documents dont le contenu correspond à la requête initiale, triée selon la similarité requête-document ;
- *la recherche par navigation* offre à l'usager une représentation visuelle du corpus documentaire qu'il peut explorer en interagissant avec le système. Par exemple, l'Open Directory Project [1] offre une vision thématique du Web où les documents sont organisés dans des catégories, telles que « Science/Biology ». D'autres approches permettent de naviguer dans la bibliothèque numérique d'un individu, au travers de sites de *social bookmarking* (Hammond *et al.*, 2005) tels que Delicious [2] ou Connotea [3] (Lund *et al.*, 2005).

Globalement, la recherche d'information est une tâche hautement cognitive pour l'individu. En effet, il doit savoir utiliser un moteur de recherche, transcrire la représentation mentale de son besoin en une requête (usuellement une liste de mots), naviguer dans la kyrielle de documents retournée par le système et enfin interpréter les résultats afin d'en extraire les documents pertinents pour satisfaire son besoin initial (Ciaccia, 2008). Afin de limiter les efforts de l'usager, une multitude d'approches étudiée par Cabanac *et al.* (2008c) vise à lui porter assistance en amont et en aval de cette activité : sélection du moteur le plus adapté à ses besoins, (re)formulation de ses requêtes, personnalisation des résultats obtenus, techniques avancées de visualisation d'information, etc. Malgré de tels assistants, cette activité ① qui représente selon Feldman (2004) entre 15 % et 35 % du temps de travail des individus dans un contexte organisationnel n'est que peu rentable car la moitié des recherches échouent.

1. cf. `http://dmoz.org` où une communauté de volontaires répertorie des sites Web depuis 1998.
2. cf. `http://delicious.com` permettant de publier sa collection de signets avec les descripteurs *(tags)* associés.
3. cf. `http://connotea.org` de l'éditeur *Nature* : plate-forme de *social bookmarking* destinée aux scientifiques.

2.2 Création ② et finalisation ③ de documents électroniques

Les activités de création, rédaction ② et finalisation ③ de documents électroniques sont principalement mises en œuvre grâce à des logiciels de traitement de texte, éventuellement utilisés pour la rédaction collaborative (Noël et Robert, 2004). Il existe aussi des approches complémentaires comme les wikis (Guzdial *et al.*, 2000) qui permettent la rédaction collaborative asynchrone. D'autres approches encore rendent la rédaction synchrone possible, où chaque rédacteur voit en temps réel les modifications des autres contributeurs (Swarts, 2004; Zheng *et al.*, 2006). Une évaluation quantitative de ces activités montre un faible rendement, en partie dû à l'inefficacité de l'activité ① : un nouveau rapport contiendrait en moyenne 90 % d'information recréée (Feldman, 2004).

2.3 Diffusion ④ des documents électroniques

Cette section décrit les moyens utilisés par les membres organisationnels pour partager et diffuser les documents électroniques qu'ils organisent dans leur EPI. Nous détaillons les approches manuelles puis automatiques, en soulignant leurs limites pour l'individu tant au niveau cognitif qu'au niveau motivationnel (Hinds et Pfeffer, 2003).

2.3.1 Partage et diffusion manuels de documents au sein de l'organisation

De nombreux moyens peuvent être mis en œuvre pour partager les documents au sein d'une organisation :

- en positionnant les *droits d'accès en lecture* sur les répertoires des arborescences de son EPI. Cette stratégie de partage est limitée car il faut identifier les personnes potentiellement intéressées et leur indiquer le chemin des répertoires partagés ;
- en créant un *répertoire partagé sur le réseau* de l'organisation. Chacun doit alors faire l'effort d'alimenter en documents dans cet espace partagé. La structuration de cet espace (en termes d'étiquetage des fichiers et répertoires et de découpage en sous-répertoires) impose une « pensée unique » lorsqu'elle est réalisée par une seule personne, chacun étant obligé d'adhérer à cette perception singulière des documents. Le problème demeure avec l'approche de classification non supervisée proposée dans (Wu et Gordon, 2004; Wu *et al.*, 2004). Lorsque cette tâche est laissée au groupe en général, en espérant assister à l'émergence d'une structure plus ou moins consensuelle, chaque usager reste tout de même contraint à adopter un point de vue qui peut ne pas être le sien, nécessitant de fait une adaptation de sa part, donc une surcharge cognitive ;
- en les publiant sur *l'intranet* de l'organisation grâce à des outils tels que Microsoft SharePoint Services ou Lotus Notes. Cette approche nécessite un effort de la part des individus qui doivent sélectionner la (les) rubrique(s) adaptée(s) pour un document donné, en se demandant où les autres membres chercheraient un tel document. La difficulté de cette tâche est amplifiée par la taille de l'intranet, Dmitriev *et al.* (2006) rapportent que celui d'IBM comprendrait au moins 5,5 millions de pages. De plus, Feldman (2004) estime que 40 % des recherches sur l'intranet de grands comptes échouent ;

– en utilisant des logiciels de *social bookmarking* (Hammond *et al.*, 2005) tels que le service Dogear (Millen *et al.*, 2006) chez IBM ou Connotea (Lund *et al.*, 2005) chez la maison d'édition Nature. Ces approches permettent à une personne de constituer sa collection de *bookmarks* et d'en partager tout ou partie. Chaque *bookmark* comprend l'URL du document, un commentaire libre et des *tags* qui sont des mots descriptifs fournis par l'usager. Par la suite, la navigation de *tag* en *tag* permet d'explorer ce corpus collectif. La principale limite de cette approche concerne les *tags* dont la sémantique est ambiguë : « BD » peut faire référence à « base de données » ou « bande dessinée », par exemple.

Comme alternative au *partage* de documents, les individus peuvent les *diffuser* par le biais de courriels ou de listes de diffusion. Cette démarche active consiste à sélectionner les documents à diffuser et à identifier les personnes potentiellement intéressées. Cela demande un effort à l'expéditeur qui doit anticiper les besoins de ses collègues, mais aussi aux destinataires qui peuvent être surchargés par de tels envois. Afin de limiter les efforts demandés aux individus par les approches manuelles, des stratégies automatiques présentées dans la section suivante ont été proposées.

2.3.2 Partage et diffusion automatiques de documents au sein de l'organisation

La mise en place d'un système de filtrage est une alternative à la recherche et au partage manuels de documents. Un tel système vise à recommander automatiquement des documents à des individus, en fonction de leurs besoins. Ce processus nécessite la construction de profils, à la fois pour représenter les documents et les besoins des usagers. Les critères de construction des profils sont très variables comme le montrent Montaner *et al.* (2003). Un choix possible consiste à représenter les thématiques des documents et les centres d'intérêt des individus. Le processus de recommandation repose alors sur une fonction d'appariement entre les profils des documents et des usagers. Les limites de cette approche concernent la difficulté à modéliser les profils et à les faire évoluer afin qu'ils représentent au mieux les attentes réelles de l'usager (Chevalier *et al.*, 2008). De plus, l'appariement usager-document souffre également de limites telles que la nécessité d'une masse critique d'usagers, le frein du démarrage à froid (difficulté d'émettre des recommandations à un nouvel usager) et le problème du vocabulaire (Furnas *et al.*, 1987) qui est récurrent en Recherche d'Information (RI) : identification et prise en compte de la synonymie, de l'homonymie, des figures de style, etc.

2.4 Exploitation ⑤ des documents électroniques

Nous discernons deux niveaux d'exploitation des documents. D'une part, le niveau « individuel » fait référence à l'individu qui prend connaissance d'un document grâce à la lecture active. D'autre part, le niveau « organisationnel » concerne l'individu qui parfait sa connaissance de son environnement et des compétences de ses collègues en accédant au capital documentaire de son organisation, constitué des documents provenant des divers EPI des membres organisationnels.

2.4.1 Niveau individuel : lecture et compréhension des documents

Pour un individu, exploiter un document requiert invariablement sa lecture. Murphy *et al.* (2003) rapportent une expérimentation réalisée avec 131 étudiants aux profils diversifiés (sexe,

âge, origine sociale, etc.) destinée à comparer la lecture d'un article du magazine *Time* avec la lecture du même document scanné et affiché sur un moniteur d'ordinateur. L'étude montre que le document scanné est significativement plus difficile à comprendre pour les participants qui le trouvent également moins intéressant et moins crédible. Par ailleurs, O'Hara et Sellen (1997) ont observé que lire un document papier et le résumer sur papier est plus simple que de lire une version électronique et produire un résumé avec un traitement de texte. En particulier, les *affordances* du papier permettent de disposer plusieurs documents sur la surface de travail, facilitant ainsi la lecture du texte et des annotations pendant l'écriture du résumé. Ces conclusions sont en accord avec Kidd (1994) qui indique que les *knowledge workers* griffonnent des annotations pour extérioriser leur réflexion, ces dernières étant par la suite utiles pour générer de l'information. Enfin, Sellen et Harper (2003, p. 63) rapportent également que la pratique d'annotation est importante pour les travailleurs du savoir, car elle leur permet de structurer et d'organiser leur pensée. Bien qu'elle soit habituelle et facile à mettre en œuvre sur papier, c'est une pratique pas ou peu supportée — toujours avec moins de souplesse que sur le papier — dans les environnements informatiques, suscitant de ce fait la frustration des lecteurs (Sellen et Harper, 2003, p. 96) lorsqu'ils sont privés de cet outil précieux.

2.4.2 Niveau organisationnel : visualisation et exploration du capital documentaire

Une kyrielle de techniques et d'outils de visualisation d'information a été proposée dans la littérature, comme l'attestent divers travaux de synthèse (Herman *et al.*, 2000; Chen, 2006; Yang *et al.*, 2008). Ces approches permettent la visualisation et l'exploration d'un corpus documentaire. De fait, elles peuvent être mises en œuvre dans le contexte organisationnel pour restituer le capital documentaire constitué des EPI des membres organisationnels. Sans avoir vocation à l'exhaustivité, cette section présente ces diverses approches exploitant les méta-données ainsi que le contenu des documents.

Parmi les approches reposant sur les méta-données, Fekete et Plaisant (2002) tirent parti de

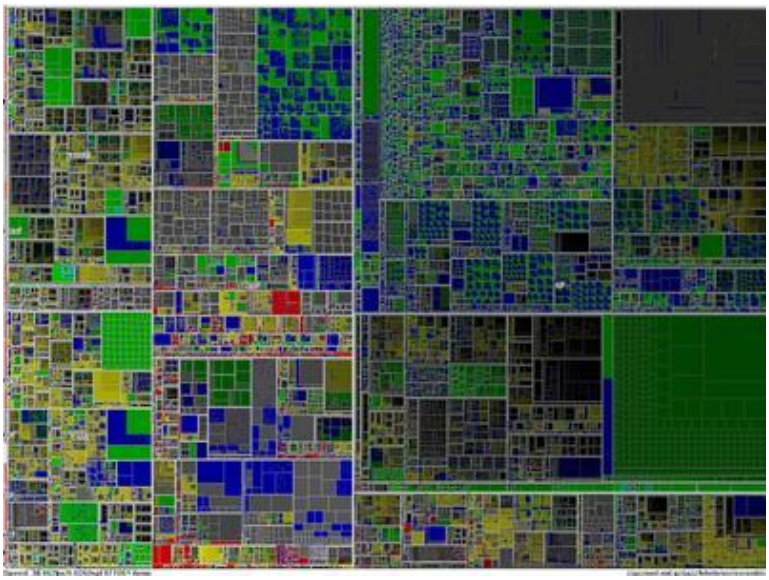


Figure I.2.1 – Visualisation de la taille des fichiers d'une hiérarchie (Fekete et Plaisant, 2002).

la visualisation *Tree-map* introduite par Johnson et Shneiderman (1991) pour représenter une arborescence de fichiers en fonction de leur taille. Sur la Figure I.2.1, on distingue des rectangles imbriqués, chacun représentant un fichier ou un répertoire. La couleur des rectangles correspond au type (extension) du fichier associé ; leur dimension est proportionnelle à la taille physique du répertoire ou fichier représenté. Cette visualisation permet également d'identifier le degré d'imbrication des répertoires correspondant aux projets : les répertoires les plus imbriqués sont présentés de manière plus sombre.

D'autres approches telles que les cartes auto-organisatrices de Kohonen (2001) représentent les thématiques des documents en se basant sur l'analyse de leur contenu. La carte est divisée en zones qui symbolisent des thématiques dont l'intitulé est affiché : on distingue « courses » au centre de la figure I.2.2, par exemple. Le dégradé de couleurs sur la carte représente le nombre de documents pour les différentes thématiques. Tout comme pour le *Tree-map*, l'usager peut consulter les détails d'une zone en la sélectionnant, obtenant alors une nouvelle carte de la zone sélectionnée. Appliquée aux EPI de l'organisation, cette visualisation offre aux membres organisationnels une vision globale des thématiques collectives. Boyer *et al.* (2007) en proposent une extension afin d'identifier les propriétaires des documents sélectionnés pour accéder à leurs documents, permettant ainsi une navigation alternative entre documents et personnes.

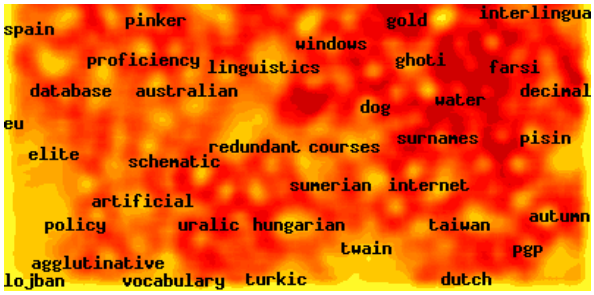


Figure I.2.2 – Carte auto-organisatrice générée par WEBSOM (Lagus *et al.*, 2004).

Par ailleurs, plutôt que de visualiser une seule caractéristique des documents — leur taille ou leurs thématiques dans les exemples précédents — des systèmes tels que DocCube (Mothe *et al.*, 2003) et Tétralogie (Dousset, 2003; Karouach, 2003) permettent l'étude du corpus documentaire selon différents axes d'analyses pouvant être spécifiés à partir de leurs méta-données (taille, date de création, auteurs) ou de leur contenu (thématiques). Enfin, ISIDOR (Chevalier et Verlhac, 2000) adossé à un moteur de recherche représente les résultats de recherche dans un cône 3D de façon à identifier leur affinité avec les mots-clés composant la requête.

2.5 Classement et archivage ⑥ de documents électroniques

L'activité de classement d'un document dans l'EPI ⑥ est une tâche hautement cognitive pour l'usager. Rucker et Polanco (1997) indiquent que le regroupement d'objets dans un répertoire reflète une cohérence sémantique entre ces objets. Ainsi, après avoir construit une représentation mentale du contenu d'un document et décidé de le conserver, l'individu doit identifier le répertoire le plus adapté dans son EPI, éventuellement en créer un nouveau, pour enfin y stocker le document considéré. Diverses études montrent que nous recourons au classement hiérarchique

pour planifier nos projets, en les décomposant en sous-projets (Jones *et al.*, 2005; Jones, 2007). Cette stratégie est confirmée par Khoo *et al.* (2007) qui soulignent la fréquence des arborescences par projet décomposées selon un à trois niveaux au moins. Des études ethnographiques reflètent la fréquence de cette activité ⑥. D'une part, Abrams *et al.* (1998) rapportent qu'un usager de signets Web en stockait de trois à quatre par session de navigation. D'autre part, l'étude des pratiques de 31 personnes réalisée par Boardman et Sasse (2004) rapporte les chiffres suivants : en moyenne une personne possède 56,6 répertoires (min = 5, max = 218) et en crée 0,35 par jour. La profondeur moyenne d'un répertoire de l'arborescence est de 3,3. Enfin, un individu classe en moyenne 5,92 fichiers par jour dans ses répertoires. En réalité, l'arborescence de documents de chaque membre organisationnel contient les documents qui lui sont utiles, organisés de façon à réaliser au mieux sa réflexion autour de ses activités. De ce fait, l'ensemble des EPI constitue un capital documentaire à forte valeur ajoutée pour l'organisation dans son ensemble.

2.6 Limites des activités documentaires sur support électronique

Notre étude des activités documentaires de ① à ⑥ (figure I.1.1) sur support électronique nous a permis d'identifier trois problématiques majeures que nous détaillons dans cette section.

2.6.1 Maîtrise d'un système par activité : surcharge cognitive pour l'usager

Concernant la mise en œuvre des six activités du cycle de vie du document nous remarquons que chaque activité requiert l'utilisation d'un système distinct. Par exemple, un système de gestion de fichiers ⑥ ne permet pas de rechercher de l'information ①. Ainsi, un usager doit maîtriser au moins six systèmes différents pour couvrir l'ensemble des activités documentaires, entraînant de fait une charge cognitive importante pour l'individu. Cette surcharge est également accentuée lorsque chaque système impose une nouvelle CPI telle que l'arborescence des courriels (logiciel de messagerie), celle des signets (navigateur Web), celle des fichiers (système d'exploitation)… Ceci entraîne une fragmentation des données dont les usagers se plaignent, comme l'indique Jones (2007) dans sa synthèse des recherches sur la mémorisation et l'organisation des informations personnelles. Par ailleurs, Kidd (1994) insiste sur le fait que, pour les *knowledge workers*, les annotations ont davantage d'importance que les documents initiaux une fois lus. De plus, les travaux de Sellen *et al.* (2002) relatifs à l'usage du Web par les *knowledge workers* montrent qu'ils ont besoin de conserver des parties de documents avec leurs notes, et pas uniquement leur URL. Or, très peu de systèmes permettant la lecture de documents proposent une fonctionnalité de création et de conservation d'annotations en contexte : les navigateurs Web en sont exempts alors que ce sont des outils indispensables à la recherche d'information ① de nos jours.

2.6.2 Représentation parcellaire des usagers : assistance offerte limitée

Les activités documentaires représentées dans la figure I.1.1 semblent linéaires, alors que l'observation des pratiques documentaires reflète plutôt leur entrelacement. Par exemple, un individu peut rechercher de l'information ①, commencer la rédaction d'un document ② puis continuer à naviguer dans les résultats de sa recherche ① afin d'approfondir sa connaissance de la question.

Par ailleurs, les activités documentaires sont également cloisonnées : chaque système est spécialisé dans la réalisation d'une seule activité. De fait, il ne peut construire qu'une représentation partielle des usagers réalisant l'activité pour laquelle il est conçu, ignorant de ce fait les cinq autres activités du cycle de vie. Pour autant, l'usager et ses besoins restent les mêmes eu égard aux six activités documentaires. Par conséquent, toute assistance apportée par un système sur cette base partielle se révélerait être sous-efficiente par essence.

2.6.3 Faible valorisation des EPI : capital documentaire organisationnel en sommeil

Le capital documentaire constitué par les EPI organisationnels est une mine d'informations à forte valeur ajoutée car elle résulte des efforts quotidiens des membres organisationnels. Paradoxalement, ce capital documentaire organisationnel n'est usuellement que peu valorisé au niveau organisationnel : par défaut, un EPI n'est accessible que par son propriétaire. De ce fait, les documents qu'un individu a trouvés au prix de coûteux efforts ne profitent pas aux autres membres, bien que certains aient des besoins informationnels proches voire similaires. De fait, ces documents qui sommeillent dans les EPI feront l'objet d'efforts de recherche répétés, parfois en vain car une recherche sur deux échouerait (Feldman, 2004). Cette méconnaissance des ressources et compétences voisines aboutit à une recréation inutile d'information : un nouveau rapport serait constitué de 90 % d'informations préexistantes selon Feldman (2004).

Nous avons souligné l'importance de la pratique d'annotation pour diverses activités documentaires, que ce soit sur support papier comme électronique. Le chapitre suivant expose plus en détail les caractéristiques de cette pratique ainsi que sa transposition du papier au numérique.

3 Zoom sur la pratique d'annotation : transposition du papier au numérique

"I have a Trick of writing in the Margins of my Books, it is not a good Trick, but one longs to say something."

Hester Thrale Piozzi (1741 — 1821)
citation rapportée par Jackson (2002, p. 74)

QUE CE SOIT sur support papier comme électronique, l'étude des activités documentaires au travers du cycle de vie du document (figure I.1.1) révèle à quel point la pratique d'annotation est médiatrice de la relation individu-document. En outre, les lecteurs recourent aux annotations pour un usage individuel comme collectif. De par son caractère intemporel et transversal aux activités documentaires, elle figure au cœur de notre contribution visant à fédérer et améliorer ces activités sur support électronique (partie II). C'est pourquoi le présent chapitre est consacré à la pratique d'annotation dont nous présentons la mise en œuvre sur papier dans la section I.3.1. Suivant un fil conducteur chronologique, nous exposons ensuite la transposition de cette activité sur support électronique (section I.3.2) en soulignant les apports et limites de son informatisation. Enfin, la section I.3.3 discute les limites de l'annotation électronique au regard de l'étude de 64 systèmes développés par industriels et universitaires durant les vingt dernières années.

3.1 L'annotation papier : une pratique séculaire toujours d'actualité

Annoter est une pratique séculaire dont on retrouve les traces sur des documents du Moyen Âge. De ces temps anciens nous sont notamment parvenues les annotations réputées du rabbin français Rashi (1040 — 1105) qu'il avait formulées sur des exemplaires de la Bible et du Talmud,

dont Fraenkel et Klein (1999) soulignent la concision et la pertinence. Lortsch (1910) rapporte l'intérêt que suscitaient les annotations provenant d'érudits : l'imprimeur et savant français Robert Estienne écrivait « En l'an 1541, j'imprimai le Nouveau Testament avec brèves annotations en marge, lesquelles j'avais eues de gens bien savants ». Wolfe et Neuwirth (2001) notent que ces documents anciens contiennent même des strates d'annotations *(layers)* formées par les générations successives de lecteurs qui complètent et répondent aux annotateurs qui les ont précédés… La pratique d'annotation a traversé les siècles, exercée par les lettrés de toutes disciplines. Jackson (2002, p. 32) rapporte qu'en mathématiques par exemple, Pierre de Fermat annote vers 1637 un exemplaire de l'*Arithmetica* de Diophante, formulant en marge de la huitième conjecture (soit en notation mathématique moderne $\forall n \in \mathbb{N}_* \ \nexists (x,y,z) \in \mathbb{N}_*^+ \quad n > 2 \ \wedge \ x^n + y^n = z^n$) l'annotation :

> « J'ai trouvé une merveilleuse démonstration de cette proposition, mais la marge est trop étroite pour la contenir. »

Il ne publia jamais la preuve évoquée. Au fil des siècles, une multitude de mathématiciens tenta alors de prouver cette conjecture, en vain. Ce n'est qu'au XX$^{\text{e}}$ siècle qu'Andrew Wiles (1995) fournit une démonstration de la conjecture de Fermat, désormais appelée « théorème de Fermat-Wiles » (Kleiner, 2000). C'est dire à quel point une annotation qui peut sembler anodine a pu susciter plus de trois siècles et demi de recherches en mathématiques ! En littérature, à une époque où le livre était un objet précieux et personnel, les lecteurs annotaient leurs exemplaires tout en lisant, ce qui les aidait à mener une réflexion critique. Cette lecture active (Adler et van Doren, 1972, p. 4) était notamment employée par les lecteurs des poésies de William Blake ou de John Keats, comme le révèle l'étude des annotations conduite par Jackson (2002) sur les livres imprimés entre 1790 et 1830. On remarque également que les auteurs eux-mêmes recourraient à l'annotation pour relire et corriger leurs documents, cf. le manuscrit des *Misérables*[1] (figure I.3.1).

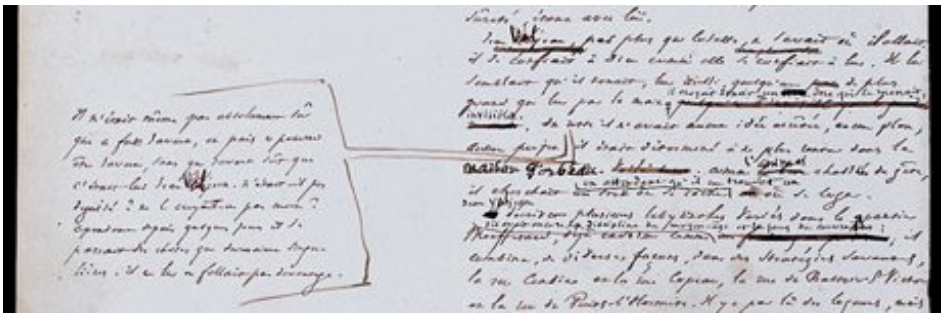


Figure I.3.1 – Extrait d'un manuscrit annoté par Victor Hugo en 1881.

Similairement, Durand Degranges (2003) rapporte que « Stendhal avait coutume d'annoter ses manuscrits (jusqu'à se perdre dans ses notes), trouvant ainsi l'occasion de placer ses propres réflexions. Sa création littéraire s'opère dans un double jeu de miroir : tout d'abord dans son texte où il intervient et surtout dans ses notes. Celles-ci constituent un véritable "journal du roman" ». De nos jours encore, l'activité d'annotation sur documents papier est quotidiennement pratiquée. En particulier, Kidd (1994) et Sellen et Harper (1997, 2003) l'ont étudiée chez les *knowledge workers* professionnels alors que Marshall (1997, 1998) en montre l'emploi par les étudiants du supérieur.

1. Volume 2 *(Cosette)*, chapitre « Les zigzags de la stratégie », page 259 de l'édition de 1881 (Hetzel-Quantin). L'image relative à ce texte est issue du site de la BNF ; le manuscrit est extrait de *Brouillons d'écrivains* (sous la direction de Marie-Odile Germain et de Danièle Thibault), Bibliothèque nationale de France, 2001, p. 66.

3.1.1 Définition de l'annotation

Dans la littérature et selon Azouaou *et al.* (2003), il n'existe pas une définition consensuelle pour l'annotation, mais plutôt plusieurs définitions générales (provenant de divers dictionnaires) ou bien spécifiques (variant selon les domaines de recherche : conception d'interfaces homme machine, psycholinguistique, documentation...). Les travaux de recherche présentés dans (Mille, 2005, ch. 2) complètent cette étude bibliographique et mettent en exergue pour sa complétude la définition suivante de Bringay *et al.* (2004) que nous adoptons.

> « Une **annotation** est une note particulière attachée à une **cible**. La cible peut être une collection de documents, un document, un segment de document (paragraphe, groupe de mots, mot, image ou partie d'image, etc.), une autre annotation. À une annotation correspond un contenu, matérialisé par une inscription, qui est une trace de la représentation mentale que l'annotateur se fait de la cible. [...] Nous appelons l'**ancre** ce qui lie l'annotation à la cible (un trait, un passage entouré, etc.). »

Les livres représentés en figure I.1.2 (p. 11) tout comme le manuscrit de Victor Hugo illustré en figure I.3.1 présentent de nombreuses variétés d'annotation. À partir de l'exemple prototypique de la figure I.3.2 (issu de la même source que la figure I.1.2) nous exposons les différentes formes (textuelles comme non-textuelles) et fonctions associées aux annotations dans les sections suivantes.

3.1.2 Formes textuelles : notes de lecture, remarques, corrections...

Un individu peut formuler une annotation textuelle à différents endroits d'un document, sa fonction dépend principalement de cette localisation qui est choisie par l'annotateur. Une étude de l'activité d'annotation dans le milieu de l'enseignement et de la recherche universitaires conduite par Ovsiannikov *et al.* (1999) indique les fréquences de localisation suivantes :

- 50 % dans la *marge* : les individus notent leurs idées et commentaires qui accompagnent ainsi le texte sans le surcharger. Ces commentaires permettent de paraphraser ou de reformuler certains passages pour mieux les comprendre ;
- 22 % dans l'*en-tête* du document : cet emplacement est privilégié par les annotateurs lorsqu'ils résument le document. Cette activité requiert un effort cognitif considérable car elle nécessite de remanier le contenu du document dans son propre vocabulaire. C'est peut-être pour cela que ce type d'annotation est moins fréquent ;
- 19 % *en dehors* du document : annoter en dehors du document vise, tout comme l'annotation dans l'en-tête, à résumer un document. Toutefois, cette forme est moins courante que la précédente ;
- moins de 10 % *entre les lignes* du document : bien que le fait d'annoter dans l'interligne soit une pratique courante dans le milieu de l'édition, elle est peu utilisée par les individus sondés. C'est peut-être parce qu'elle est surtout mise en œuvre lors des corrections de documents et rarement dans le cas d'une simple lecture.

Les annotations textuelles décrites ci-dessus sont clairement visibles sur la figure I.3.2. Cet exemplaire de livre annoté contient également des formes non-textuelles que nous examinons dans la section suivante.

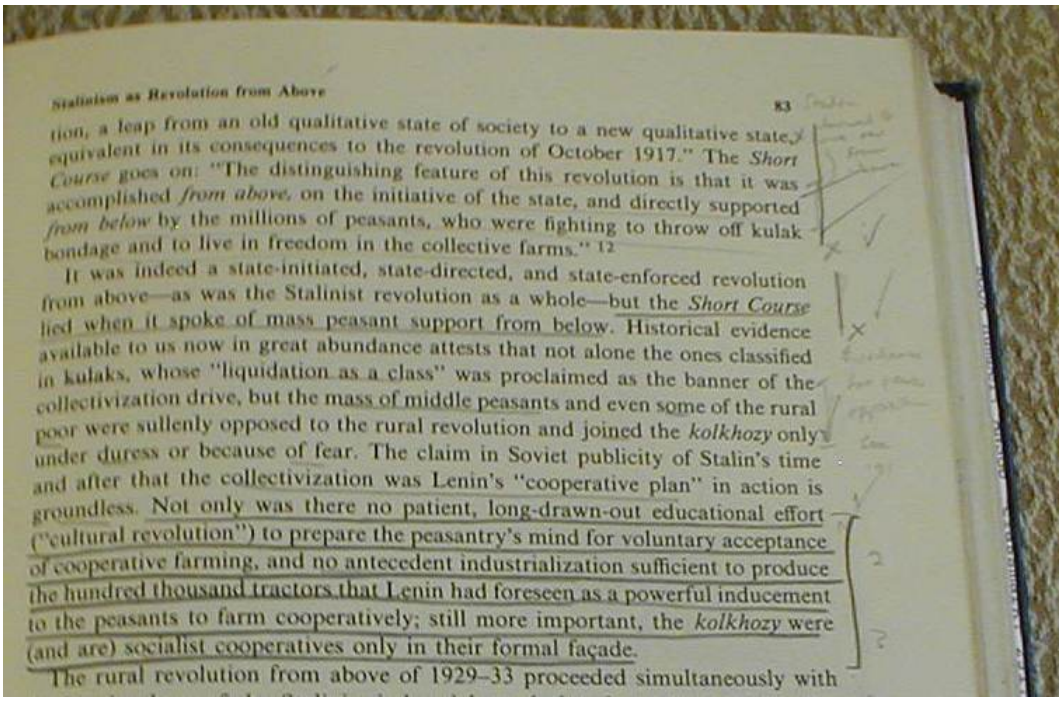


Figure I.3.2 – Diverses formes textuelles et non-textuelles d'annotations.

3.1.3 Formes non-textuelles : mise en emphase, apprentissage, catégorisation…

L'étude d'une large collection de livres annotés par des étudiants conduite par Marshall (1997, 1998) montre qu'ils emploient divers « outils » : ils utilisent surligneurs, stylos, crayons à papier et autres instruments pour créer les annotations. Il existe une grande fluidité dans la forme : notations symboliques, dessins sur et autour du texte ; les annotateurs soulignent, entourent, encadrent et surlignent tout type d'élément textuel. Cette étude met en évidence la très grande créativité des lecteurs dans leur engagement avec les documents pour atteindre divers objectifs :

1. **mettre en valeur un passage du document annoté.** Parmi les techniques de marquage possibles, les annotateurs choisissent neuf fois sur dix de surligner ou de souligner (Ovsiannikov *et al.*, 1999; Denoue, 2000). Une autre pratique omniprésente, la mise en emphase, consiste à ajouter des marques près du texte : « ✓ » et « ✗ » sont notamment visibles sur la figure I.3.2. Le passage adjacent à la marque est alors visuellement identifiable. Les marques d'emphase sont le plus souvent des étoiles ou astérisques bien qu'un annotateur inventif puisse employer une grande variété de symboles. Elles peuvent aussi caractériser des niveaux d'importance, par exemple la répétition de symboles permet de marquer un passage clé du document. Enfin, il existe d'autres techniques de mise en emphase qui consistent à varier l'épaisseur du trait avec un surligneur, surligner de deux couleurs différentes un même passage, etc. En phase de relecture, ce type de mise en valeur permet à l'individu de faire un tri de ses annotations pour rassembler en un clin d'œil les idées principales en se focalisant sur les passages clés d'un document. Il est notamment mis en œuvre lorsqu'un effort de mémorisation est nécessaire : la mise en valeur des définitions d'un cours avec un surligneur fluorescent, par exemple.
2. **typer et catégoriser des passages du document annoté.** L'utilisation de différentes couleurs ou symboles peut servir à différencier les annotations afin :

- d'identifier des passages traitant du même thème, dans l'objectif de les rassembler au sein d'une synthèse par exemple,
- d'associer un jugement à un passage, en convenant d'une sémantique. Par exemple, la couleur verte ou le symbole « ✓ » peuvent exprimer l'accord alors que la couleur rouge ou le symbole « ✗ » le désaccord. Ce type de jugement sert notamment dans le cadre de l'élaboration d'articles scientifiques, en particulier en phase de correction.

3. **re-segmenter le document annoté.** Lorsque la structure imposée par l'auteur ne convient pas au lecteur, ce dernier recourt aux annotations non-textuelles afin de re-segmenter le document. Cette restructuration du texte est souvent réalisée en énumérant des parties du texte (cf. les chiffres 1, 2 et 3 en bas de la figure I.3.2) ou en surlignant d'une couleur différente les éléments du document. Ici, une couleur n'équivaut pas à un type mais c'est le changement de couleur qui indique la restructuration du document. Marshall (1998) précise qu'il ne faut tout de même pas surestimer l'utilisation de ce type d'annotation car il est bien moins commun que d'autres (commentaire, surlignement…). De plus, un étudiant qu'elle a interrogé pour son étude a confessé qu'il changeait de couleur pour maintenir un intérêt dans sa lecture !

En synthétisant les différentes fonctions associées aux formes constatées et exposées jusqu'ici, nous détaillons dans les sections suivantes les finalités recherchées par l'individu qui annote une ressource, à des fins personnelles ou collectives.

3.1.4 Finalités de l'activité d'annotation pour un usage personnel

Jackson (2002, ch. 3) consacre le chapitre intitulé *"Motives for Marginalia"* aux finalités des annotations papier. Sans avoir prétention à l'exhaustivité, nous présentons dans cette section une synthèse des finalités identifiées dans la littérature :

1. **favoriser l'apprentissage grâce à la lecture active.** De nombreuses activités intellectuelles humaines sont basées sur un cycle de lecture-écriture des documents, c'est en particulier le cas des activités des *knowledge workers*. Dans ce cycle, les annotations permettent aux lecteurs de devenir instantanément rédacteurs. L'écriture d'annotations facilite l'appropriation du texte grâce à la reformulation : l'ajout de commentaires permet d'identifier un passage difficile à comprendre, de le synthétiser en quelques mots, de le relier ou au contraire de l'opposer à une autre partie du même document. Cette pratique se développe avec le temps : l'expérience et les attentes des lecteurs modifient la façon dont ils créent leurs annotations. Par exemple, Marshall (1998) observe que les étudiants de 1re année ont une idée insuffisante de comment annoter alors que les étudiants de 3e et 4e année sont plus instruits sur cette pratique. D'après la littérature, il est indéniable que formuler des annotations sur des documents est une pratique indispensable à la lecture active (Adler et van Doren, 1972; Sellen et Harper, 1997; Marshall, 1997, 1998; Wolfe et Neuwirth, 2001; Jackson, 2002) ;
2. **catégoriser des passages du document.** Les lecteurs immergés dans le texte sont rarement plus explicites que nécessaire lorsqu'ils annotent. Les annotations personnelles qui en résultent sont par nature télégraphiques, incomplètes et tacites (Marshall, 1998). Une phrase surlignée, une remarque lapidaire dans la marge (« Non ! »), un lien entre deux paragraphes non commentés sont difficiles à interpréter pour quiconque autre que l'annotateur original. Toutefois, ces marques sont suffisantes pour qu'il les comprenne lors d'une relecture ;

3. **matérialiser physiquement l'état d'avancement d'une tâche.** En s'appuyant sur une étude de l'utilisation des annotations, Denoue (2000) remarque que les individus les emploient pour repérer l'état d'avancement dans la lecture d'un document. Similairement, Marshall (1998) décrit un phénomène qui se produit en présence de textes particulièrement denses : l'annotation devient une trace visible de l'attention humaine. C'est notamment le cas sur les livres de philosophie étudiés : ils sont intégralement surlignés page après page par le lecteur. Dans ce cas, ces marques sont clairement importantes pour l'activité physique de lecture ;

4. **se remémorer les points clés du document.** Les lecteurs ont tendance à oublier le contenu d'un document et ont besoin de se le remettre en tête de temps en temps (Ovsiannikov *et al.*, 1999). En parcourant un document annoté, la seule lecture des commentaires et du texte mis en valeur permet de se rappeler son contenu.

Globalement, Ovsiannikov *et al.* (1999) rapportent les trois finalités suivantes associées à leur importance relative, eu égard aux annotations pour un usager individuel : se souvenir (41 %), réfléchir (32 %) et clarifier (23 %). Nous relatons dans la section suivante les finalités des annotations pour un usage collectif : créées par un lecteur à destination des futurs autres lecteurs.

3.1.5 Finalité de l'activité d'annotation pour un usage collectif

On retrouve les traces d'annotations rédigées à destination de futurs lecteurs sur des documents datant du Moyen Âge. Wolfe et Neuwirth (2001) expliquent qu'à cette époque de multiples lecteurs avaient typiquement accès au même exemplaire d'un texte, ce qui en faisait une ressource publique de choix pour le partage d'information. De fait, les annotations formulées en marge de ces textes revêtaient un rôle de communication et d'échange. De nos jours encore, elles sont employées pour des raisons similaires lorsqu'on considère les annotations des *knowledge workers* qui relisent et corrigent les rapports du FMI), dans l'étude de Sellen et Harper (2003, p. 61).

Dans le cadre plus général des livres annotés, Marshall (1998) signale que les individus n'accordent pas tous de la valeur aux annotations d'autrui. En effet, certains étudiants de son étude achètent le livre d'occasion le moins annoté, le plus immaculé. *A contrario,* les annotations d'un lecteur sont parfois considérées par d'autres comme une valeur ajoutée. Wolfe (2000) montre par exemple qu'elles influencent la perception individuelle des arguments exprimés dans le texte annoté. Dans une étude complémentaire, Wolfe et Neuwirth (2001) considèrent l'activité de lecture comme support à l'écriture d'un résumé de texte ; les annotations ont une influence sur le lecteur car il produit des écrits de moins bonne qualité lorsque les annotations associées au texte sont du même avis que lui, comme s'il ne voyait pas le besoin de persuader le futur lecteur. Par contre, la production des individus est meilleure lorsque le texte qu'ils lisent est annoté avec des points de vue conflictuels. En effet, Wolfe (2008) note que leur production est moins descriptive, davantage critique et exprime une réflexion personnelle dans cette situation. En résumé, la consultation des annotations de lecteurs précédents semble favoriser la réflexion critique du lecteur, notamment lorsqu'elles expriment des opinions diversifiées.

3.2 Transposition de la pratique d'annotation sur support électronique

L'étude de la littérature révèle l'utilité et l'importance de la pratique d'annotation au sein des activités documentaires des individus. De fait, les concepteurs de logiciels ont tôt fait de l'intégrer dans leurs applications en cherchant à reproduire la commodité des annotations papier tout en bénéficiant des capacités de traitement et de communication offertes par l'informatique moderne. L'annotation électronique est alors la cheville ouvrière des « documents pour l'action » de Zacklad (2007), soutenant les activités coopératives des individus dans de nombreux contextes professionnels. Ainsi, les annotations électroniques sont mises à contribution sur le dossier patient (Bringay *et al.*, 2007), sur les plans des architectes (Boulanger *et al.*, 2007) comme des concepteurs en mécanique (Boujut *et al.*, 2007), dans les projets informatiques *Open Source* (Barcellini *et al.*, 2007), sur les relevés topographiques des glaciologues (Fogli *et al.*, 2005) et archéologues (Baber *et al.*, 2005) et même sur les partitions des musiciens (Donin et Theureau, 2007)...

Dans cette section, le terme annotation fait référence à l'annotation « informelle » au sens de Marshall (1998), également qualifiée de « cognitivement sémantique » par Zacklad *et al.* (2003). C'est exactement le reflet de la pratique d'annotation papier dont le contenu est libre [2]. Les sections suivantes présentent les caractéristiques des « systèmes d'annotation » permettant d'annoter les documents électroniques textuels. Nous exposons leur architecture générale puis leur mise en œuvre avant de présenter un panorama de 64 systèmes développés durant les vingt dernières années par industriels et universitaires. Enfin, une discussion de leurs limites conclut cette partie I.

3.2.1 Catégories et architecture générale d'un système d'annotation

Les systèmes d'annotation — par la suite abrégés en « SA » — ont été développés dès les années 1990 pour transposer sur document électronique la pratique séculaire d'annotation. Puis, ces systèmes ont progressivement profité des capacités de traitement et de communication des ordinateurs modernes pour enrichir la pratique d'annotation désormais électronique. En effet, certains SA permettent à des lecteurs distants de visualiser les mêmes annotations lorsqu'elles sont partagées, par exemple. Nous avons identifié dans la littérature les catégories de SA suivantes :

1. les SA *personnels* tels que ScreenCrayons (Olsen *et al.*, 2004) et OAS (Harmon, 2007) permettent d'annoter des documents sans avoir vocation à partager les commentaires avec d'autres personnes. Concrètement, le lecteur dispose d'outils de dessin (stylo, gomme, zone de texte, etc.) grâce auxquels il ajoute des marques sur une capture écran du document à annoter. De fait, ce type de SA permet d'annoter n'importe quel format de document, mais aussi des documents placés côte à côte similairement à la pratique sur support papier ;
2. les SA *hybrides* tels que Microsoft Word et Adobe Acrobat Reader sont des outils qui offrent une fonctionnalité d'annotation personnelle mais également partageable. Les annotations peuvent être ancrées sur une partie du texte ; elles sont stockées au sein du document annoté,

2. Uren *et al.* (2006) présentent une synthèse des travaux sur l'annotation « formelle » ou « computationnellement sémantique » qui est utilisée dans le cadre du Web sémantique (Berners-Lee *et al.*, 2001). Ce type d'annotation vise à associer un élément de sémantique (éventuellement issu d'une ontologie, ou d'un vocabulaire contrôlé) à tout ou partie d'une ressource. Bien que la finalité des annotations informelles diffère de celle des annotations formelles, toutes deux reposent sur des principes communs : point d'ancrage, stockage, etc. qui sont traités dans cette section.

il faut donc disposer d'un droit en écriture sur le fichier en question. Un lecteur transmet ses annotations à une tierce personne en lui transférant le document annoté ;

3. les SA *collectifs* tels que ThirdVoice (Margolis et Resnick, 1999) ou celui développé par le W3C nommé Annotea/Amaya (Kahan *et al.*, 2002) sont conçus pour créer des annotations en contexte et les partager avec de futurs lecteurs. Ils sont généralement destinés aux documents HTML du Web. Comme il s'agit de documents non modifiables, les SA de cette catégorie stockent les annotations à part, en conservant le point d'ancrage de l'annotation sur le document. À la restitution d'un document, les annotations associées y sont intégrées afin que le lecteur les voie en contexte.

Dans le cadre de ce mémoire relatif aux activités documentaires dans un contexte organisationnel, les SA *collectifs* sont les plus adaptés pour couvrir la pratique d'annotation d'un groupe. Cette catégorie requiert un développement plus important que les deux autres, notamment pour la spécification du point d'ancrage des annotations. Nous détaillons dans les sections suivantes la mise en œuvre d'un tel système d'annotation.

3.2.2 Mise en œuvre d'un système d'annotation

Nous exposons dans cette section les principaux éléments relatifs à l'implantation d'un SA : spécification de l'ancrage des annotations, stockage de leur contenu, restitution à l'usager *via* une visualisation adaptée, ainsi que variantes d'intégration dans l'environnement de l'usager.

3.2.2.1 Techniques d'ancrage robustes pour localiser les annotations sur les documents

Deux stratégies sont envisageables quant à la mémorisation d'une annotation relative à un document donné. La première consiste à intégrer (un lien vers) l'annotation dans le document, dont le contenu est alors modifié. Lorsqu'on ne peut pas intégrer les annotations dans les documents parce qu'ils sont en lecture seule (c'est le cas sur le Web, par exemple) il faut recourir à une technique d'ancrage pour conserver le lien annotation-document sans modifier le document.

Nous considérons ici le cas des documents HTML car c'est le format retenu par la majorité des SA étudiés (section I.3.2.3). Une première technique d'ancrage implantée dans Yawas (Denoue, 2000) consiste à mémoriser le passage sélectionné ainsi que son « rang » dans le document. Cela permet de le distinguer parmi les éventuelles occurrences du même texte dans le document. Cette technique est intuitive, simple à mettre en œuvre mais peu robuste car peu résistante aux modifications apportées au document. Ovsiannikov *et al.* (1999) proposent une technique plus robuste pour les documents semi-structurés, où le point d'ancrage du passage sélectionné est le chemin dans la structure logique du document. Un tel ancrage correspond alors à une expression de type « Dans le document X, dans le chapitre 1, second paragraphe, démarrant par "externally-guided" jusqu'à la fin de la phrase ». Ce type d'ancrage a été standardisé par le W3C dans le cadre de documents XML où le chemin d'un élément peut être exprimé grâce au langage XPointer (DeRose *et al.*, 2002). Ce dernier permet de formuler des points d'ancrage au sein même d'un élément XML, offrant ainsi une granularité plus fine que XPath (Clark et DeRose, 1999). Concrètement, XPointer est utilisé dans Annotea/Amaya (Kahan *et al.*, 2002) qui crée de préférence des points d'ancrages relatifs, afin qu'ils soient moins sensibles aux modifications en amont dans la structure du document. L'algorithme suivant est employé à cet effet : à partir du passage sélectionné, remonter

dans l'arborescence du document jusqu'au premier élément possédant un attribut `id` qui l'identifie, ou bien jusqu'à la racine si un tel attribut n'existe pas (ancrage absolu). Par exemple, l'expression XPointer $\mathscr{P}$ = `xpointer(string-range(id("Issues")/p[2],"",0,5))` désigne le mot « Amaya » du second paragraphe dans le fragment de document HTML de la figure I.3.3.

```
...
<div id="Issues">
    <h1>Issues with...</h1>
    <p>If you are using...</p>
    <p>Amaya uses <strong>XPointer</strong>...</p>
</div>
...
```

Figure I.3.3 – Bribe d'un document HTML sur lequel on peut créer un point d'ancrage XPointer.

La spécification d'un point d'ancrage en XPointer permet d'identifier sans ambiguïté un passage dans un document XML qui n'évolue pas. Dans le cas contraire et lorsque les modifications impactent la restitution du point d'ancrage, Kahan *et al.* (2002) appellent une telle annotation :

- *orpheline (orphan)* lorsque le point d'ancrage ne peut plus être résolu. Par exemple, après la suppression du second paragraphe dans la figure I.3.3 pour le point d'ancrage $\mathscr{P}$;
- *trompeuse (misleading)* lorsque le contenu du passage a été modifié sans altérer la structure du document. Par exemple, après modification des cinq premiers caractères du second paragraphe (changement « Amaya » → « erreur ») toujours considérant $\mathscr{P}$. Pour détecter une annotation trompeuse, le système d'annotation peut s'assurer de l'égalité entre le contenu originellement sélectionné et mémorisé (« Amaya ») et le contenu obtenu après résolution du point d'ancrage (« erreur »). MS Office Web Discussions (Cadiz *et al.*, 2000) procède de cette façon en présentant la particularité de ne stocker que la valeur de hachage des chaînes de caractères, par économie de stockage et de temps de transfert.

Afin d'améliorer la robustesse du point d'ancrage des annotations, Phelps et Wilensky (2000) utilisent la structure du document si disponible (XML) ou des informations contextuelles sinon. Sur des documents semi-structurés, Bouvin *et al.* (2002) définissent plusieurs XPointer pour un même passage. En fait, les expérimentations par Abe et Hori (2003) sur un corpus de page Web fréquemment modifiées indiquent que l'ancrage absolu est bien plus stable que l'ancrage relatif. Ceci est dû au fait que l'ancrage relatif (sur la balise père la plus proche) dépend de cette balise, qui se révèle dans les faits souvent instable. D'autres approches (Heck et Luebke, 1999) emploient un algorithme approximatif de comparaison de chaînes de caractères *(approximative string matching)* pour retrouver le passage annoté même s'il a été modifié. De même, Bargeron *et al.* (2001) considèrent des techniques d'ancrage robustes *(robust text anchoring)*. Enfin, l'algorithme *"Keyword Anchoring"* proposé par Brush (2002) se focalise sur les mots-clés de l'ancre plutôt que sur sa localisation dans la structure du document, ce qui permet de l'utiliser pour une grande variété de formats électroniques.

3.2.2.2 Stockage des annotations électroniques

Une fois généré, le point d'ancrage sélectionné ainsi que les informations composant l'annotation (typiquement, un titre et un commentaire, accompagnés d'une indication de visibilité : pri-

vée ou public) sont stockés dans un serveur d'annotations. Plusieurs personnes utilisant le même serveur peuvent alors partager leurs annotations et visualiser celles des autres individus. Nous n'avons pas identifié dans la littérature un modèle d'annotation ou un format de représentation consensuels et rapportons ici un échantillon représentatif de la diversité observée. Yawas (Denoue et Vignollet, 2000b) stocke les annotations dans un fichier texte structuré en champs, éventuellement partagé entre plusieurs usagers au travers d'un dossier accessible par le réseau. WebAnn (Bargeron *et al.*, 2001) repose sur le modèle *"Common Annotation Framework"* exprimé en AML — basé sur RDF — et interprété par le serveur associé. L'emploi de RDF pour décrire le modèle d'annotation est également retenu pour le SA Amaya/Annotea (Kahan *et al.*, 2002) du W3C. Enfin, IPSA (Agosti *et al.*, 2005) repose sur un modèle entité-association alors que Agosti et Ferro (2007) exposent un modèle formel pour l'annotation de documents électroniques.

3.2.2.3 Visualisation des annotations électroniques

Nous avons observé dans la littérature deux stratégies pour restituer les annotations d'un document au lecteur :

1. *Représentation hors contexte.* Contrairement au résultat de certaines annotations papier, certains SA ne modifient pas du tout la mise en page originale des documents. Par exemple, Third Voice (Margolis et Resnick, 1999) affiche la liste des annotations dans un cadre à côté du document, reprenant alors la métaphore de la marge dans une feuille de papier. D'autres systèmes comme JotBot (Vasudevan et Palmer, 1999) et Pharos (Bouthors et Dedieu, 1999) restituent les annotations dans une fenêtre annexe, sans les incorporer dans leur document d'origine. Ce type de visualisation demande un effort cognitif à l'utilisateur qui doit fusionner mentalement les annotations et le document. De plus, il est difficile de gérer les deux fenêtres (navigateur et assistant). Ces limites ont conduit les SA plus récents à incorporer les annotations en contexte, au sein même des documents ;
2. *Représentation en contexte.* C'est l'alternative la plus retenue par les SA, consistant à intégrer les annotations au sein du document original. Ceci donne la possibilité au lecteur d'interrompre à sa guise une lecture linéaire. Nous exposons par la suite un aperçu des différentes techniques employées pour fusionner les annotations aux documents, de la plus « envahissante » à la plus discrète et flexible :

   (a) *Incorporation du contenu de l'annotation.* À l'extrême opposé des systèmes qui ne modifient pas le document visualisé, certains y incorporent la totalité de l'annotation. Ainsi, MS Office Web Discussions (Cadiz *et al.*, 2000) ou Annotator (Ovsiannikov *et al.*, 1999) mettent en évidence les annotations dites « *inline* » en insérant les commentaires dans le document, dans un style (police et couleur de fonte) choisi par l'auteur et paramétrable par le lecteur. Ceci permet de bien faire la différence entre une annotation et le texte original. Lorsque le contenu d'une annotation dépasse une certaine taille, elle est qualifiée de « *memo* » et est affichée dans une fenêtre auxiliaire qui est automatiquement ouverte. En plus de l'annotation, CoNote (Davis et Huttenlocher, 1994) va jusqu'à insérer le nom de l'auteur ainsi que le jour et l'heure de création. Ce type de représentation altère fortement les documents et ne permet pas d'afficher un grand nombre d'annotations sans distraire voire déranger le lecteur. D'autre part, Vasudevan et Palmer (1999)

soulignent la difficulté d'afficher des annotations dans la marge d'un document HTML alors que cette représentation est naturelle sur du papier. C'est pourquoi des techniques alternatives moins intrusives ont été étudiées ;

(b) *Incorporation d'un pictogramme hypertextuel.* La majorité des systèmes n'insère dans le document qu'une icône qui matérialise le début du point d'ancrage, comme par exemple ComMentor (Röscheisen *et al.*, 1994), Amaya (Kahan *et al.*, 2002) avec « », Annozilla (Mozdev, 2003). Les icônes servent aussi d'hyperliens qui permettent de voir le contenu de l'annotation en plaçant le pointeur de la souris au-dessus : le texte d'affiche dans une infobulle. En cliquant sur le lien, l'utilisateur accède à des fonctionnalités supplémentaires : ajouter un commentaire, copier le contenu, voir le profil de l'annotateur, etc. Techniquement, le simple fait de rajouter une balise HTML d'ancre dans le document, par ex. `<A alt="texte de la bulle">passage annoté</A>` permet d'afficher dans le navigateur une bulle d'aide contenant le texte spécifié par l'attribut `alt`. Bien qu'elle puisse être mise en œuvre simplement, cette technique induit des limitations : plusieurs bulles ne peuvent pas être affichées simultanément, la bulle peut occulter le texte original et les utilisateurs ne peuvent pas modifier directement l'annotation ni copier le commentaire. Pour éviter les problèmes des infobulles de HTML, Bouvin *et al.* (2002) introduisent le concept d'annotation « fluide ». Les ancres de ces annotations sont représentées dans leur contexte grâce à une mise en forme particulière définie par l'utilisateur, un soulignement pointillé par exemple. Pour les visualiser, l'utilisateur clique dessus : le contenu (formulé en HTML : texte, image, tableau, etc. sont exprimables) de l'annotation est progressivement inséré dans le document. Cette animation permet de visualiser les modifications que subit la mise en page du document, ce qui aide le lecteur à localiser l'annotation ouverte car les annotations passent fréquemment inaperçues sans animation. Quatre étapes d'une telle animation sont présentées dans la figure I.3.4.

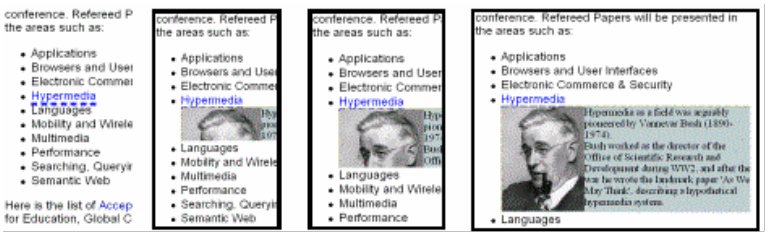


Figure I.3.4 – Étapes de l'ouverture animée d'une annotation « fluide » (Bouvin *et al.*, 2002).

Lors de l'évaluation du système Yawas, Denoue (2000) remarque que la quasi-totalité des annotations créées avec ce système ne comportent pas de commentaire. C'est pourquoi Yawas utilise un surlignage fluorescent, qui rappelle la métaphore papier tout en minimisant l'impact sur la mise en page originale.

Par ailleurs, certains SA enrichissent la pratique d'annotation électronique en faisant de chaque annotation un point d'entrée pour un débat en contexte. En fait, tout lecteur peut réagir à une annotation en y associant une réponse, formant ainsi un fil de discussion *(discussion thread)*. Par la suite, les réponses peuvent également susciter de nouvelles réponses. Ainsi, chaque annotation peut initier un forum en contexte, là où ceux de Usenet nécessitent de situer le contexte

de sa question ou remarque. Similairement à Usenet, un fil de discussion suscité est couramment représenté par une arborescence de réponses : la figure I.3.5 montre la visualisation proposée par Amaya (Kahan *et al.*, 2002).

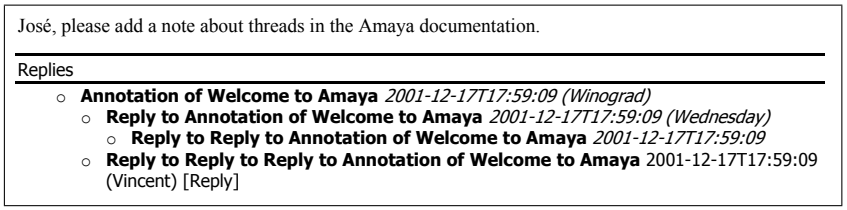


Figure I.3.5 – Représentation d'un fil de discussion dans Amaya (Kahan *et al.*, 2002).

La technique d'ancrage, le stockage des annotations et leur restitution sur ou autour des documents représentent les briques de base de tout système d'annotation. À partir de ces éléments, les concepteurs de SA ont recouru à différentes approches pour intégrer leur application dans l'environnement de l'usager, détaillées dans la section suivante.

3.2.2.4 Variantes d'intégration dans l'environnement de l'usager : 3 générations de systèmes

L'intégration d'un système d'annotation dans l'environnement des individus est une question critique quant à son adoption. Nous avons identifié dans la littérature trois générations de SA, en partant des systèmes les plus anciens et les moins intégrés vers des systèmes récents qui sont totalement incorporés dans le navigateur de l'usager :

1. les SA *autonomes* imposent à l'usager d'abandonner ses propres outils à leur profit. Un SA représentatif est ComMentor (Röscheisen *et al.*, 1994) qui fut développé lors des balbutiements du Web, preuve que le besoin d'annoter les documents consultables sur ce nouveau médium s'est tôt fait sentir. C'est en fait une version étendue du navigateur Web NCSA Mosaic. Quant au système Annotea/Amaya (Kahan *et al.*, 2002), c'est un navigateur à part entière ;
2. les SA « *parasites* » (Marais et Bharat, 1997) se comportent comme des *proxies* HTTP : à l'écoute du chargement des pages du navigateur, ils détournent le flux de données provenant du Web pour y insérer les annotations avant l'affichage des documents. Un tel système requiert la configuration du navigateur afin qu'il utilise le SA comme *proxy*. WebTagger (Keller *et al.*, 1997), Annotator (Ovsiannikov *et al.*, 1999) et Pharos (Bouthors et Dedieu, 1999) sont prototypiques de cette seconde génération ;
3. les SA *intégrés* sont des composants additionnels insérés dans le navigateur de l'usager, qui garde ses habitudes de travail acquises avec son outil. Deux représentants de cette famille sont Yawas (Denoue et Vignollet, 2000b) et Annozilla (Mozdev, 2003).

3.2.3 1989 – 2008 : panorama de 64 systèmes d'annotation informelle

Nous avons recensé 64 systèmes d'annotation développés par industriels et universitaires durant les vingt dernières années (tableaux I.3.1 et I.3.2). Les caractéristiques des systèmes sont groupées selon les en-têtes suivants :

- **système.** L'année de développement du système, son nom et une référence sont mentionnés. Une astérisque à droite du nom indique que la référence citée ne provient pas des créateurs du SA, mais d'auteurs tiers qui y font référence dans leur article ;
- **format.** Le format des annotations et celui des documents que le SA permet d'annoter sont mentionnés dans cet en-tête ;
- **auteur.** Nous mentionnons si les annotations sont présentées avec le nom et le courriel de leur créateur, permettant alors la mise en relation des individus, par exemple ;
- **ancrage.** Concerne la possibilité d'ancrer une annotation sur tout ou partie d'un document, ainsi que la possibilité de définir plusieurs points d'ancrage pour une même annotation ;
- **dates.** Spécifie le stockage des dates de création et de modification éventuelle des annotations ;
- **contenu de l'annotation.** Identifie la présence d'un titre, d'un contenu textuel et la possibilité de formuler les annotations avec de l'encre numérique (utilisation éventuelle d'un stylet). De plus, nous notons la présence de *tags* (mots descriptifs librement choisis par l'annotateur), de références vers d'autres documents (éventuellement des hyperliens), de « types » d'annotation (indicateurs de la sémantique du commentaire : question, exemple, etc.). Enfin le champ « visibilité » indique le niveau de partage que l'usager peut associer à son annotation : pour l'Individu créateur uniquement, Public, Groupe ou Sélection de certains usagers ;
- **activités.** Nous indiquons quelles activités du cycle de vie du document (figure I.1.1) sont supportées par le système considéré ;
- **notification.** Cette colonne indique si le lecteur peut distinguer les nouvelles annotations de celles qu'il a déjà vues au sein d'un document.

Dans les tableaux suivants, le symbole « + » désigne la présence de la caractéristique considérée, alors que « − » reflète son absence. Enfin « ? » désigne une caractéristique dont nous n'avons pu évaluer ni la présence ni l'absence.

Les données synthétisées dans ces tableaux montrent un développement continu depuis 1993, où prédominaient les initiatives du milieu académique. De nos jours, elles semblent s'estomper au profit de l'industrie. Nous supposons que ce phénomène résulte d'un transfert de technologie. Les documents annotables sont majoritairement au format HTML, notamment parce que ce format permet l'accès au texte brut et la modification dynamique de la structure logique (Document Object Model) du document pour y insérer les annotations. L'annotation de tout ou partie du document est supportée par quasiment tous les SA ; par contre, la sélection de plusieurs points d'ancrage pour une annotation est très rarement supportée. Le contenu minimal d'une annotation semble être un titre et un texte pour formuler un commentaire. Notons qu'une minorité de systèmes adapte l'affichage des annotations pour clairement indiquer au lecteur celles qui sont nouvelles depuis sa dernière visite (« notification »).

De nombreux SA permettent le partage des annotations (visibilité différente de « I »), ils sont toutefois de moins en moins nombreux à offrir une fonctionnalité de fil de discussion. Au regard des activités documentaires supportées, c'est naturellement l'exploitation ⑤, au travers de la lecture active, qui figure en tête. Notons que la possibilité de classer ses annotations au sein d'un espace personnel (tel qu'une arborescence) ne figurait pas dans les premiers systèmes et tend à se généraliser dans les SA les plus récents. L'exploitation de telles ressources à des fins de diffusion ④ d'information est également une fonctionnalité plus fréquente dans les systèmes récents.

système			format		auteur	ancrage		dates	contenu de l'annotation									
année	nom	référence	rech/indus	annotations	documents	identité courriel	intégralité	sélection	*n* sélections	création	modification	titre	contenu textuel	encre digitale	tags	références	types	marques diverses	fils de discussion	visibilité	activités	notification
1989	Internote	(Catlin *et al.*, 1989)	R	Texte	Texte	+ –	+	+ –		+	+	–	+	–	–	–	–	–	+	P	⑤	–
1993	Mosaic v 1.0	(NCSA, 1993)	I	Texte	HTML	– +	+	– –		+	– +	+	+	–	–	–	–	–	–	IPG	⑤	+
1993	Word		I	prop.	DOC	+ –	+	+ –		+	–	–	+	–	–	–	–	–	–	P	②③⑤	–
1994	ComMentor	(Röscheisen *et al.*, 1994)	R	prop.	HTML	+ +	+	+ –		+	?	+	+	–	–	–	–	–	+	IPG	⑤	–
1994	CoNote	(Davis et Huttenlocher, 1995)	R	?	HTML	+ –	+	+ –		+	–	+	+	–	–	–	–	–	+	P	①⑤	–
1995	Futplex	(Holtman, 1996)	R	HTML	HTML	+ –	+	+ –		?	+	+	+	–	–	+	–	–	+	PG	⑤	–
1995	Hypernews	(LaLiberte et Braverman, 1995)	R	HTML	aucun (forum)	+ +	–	– –		+	–	+	+	–	–	–	+	–	+	P	⑤	–
1996	GrAnt	(Schickler *et al.*, 1996)	R	HTML	HTML	+ +	+	– –		+	–	+	+	–	–	–	–	–	+	IPG	⑤	+
1996	Internet Explorer v. 3		I	Texte	tout format	– –	+	– –		+	+	+	–	–	–	–	–	–	–	I	⑥	–
1996	Multivalent Annotations	(Phelps et Wilensky, 2000)	R	?	HTML, PDF, DVI, Image etc.	– –	+	+ –		–	–	+	+	–	–	–	–	–	+	I	⑤	–
1997	DocReview	(Hendricksen, 1997)	R	HTML	HTML	+ +	+	+ –		?	?	?	+	–	–	–	–	–	–	P	③⑤	+
1997	JotBot*	(Vasudevan et Palmer, 1999)	I	Texte	HTML	+ –	+	– –		+	–	–	+	–	–	–	+	–	+	P	⑤	–
1998	CritLink*	(Heck *et al.*, 1999)	R	prop.	HTML	+ +	+	+ –		+	+	+	+	–	–	–	+	–	–	IPS	⑤	+
1998	PageSeeder*	(Brush, 2002)	I	Texte	HTML, PDF	+ +	+	+ –		+	–	+	+	–	–	–	+	–	+	PG	⑤	–
1998	Xlibris	(Price *et al.*, 1998)	R	XML	Image	– –	+	+ +		–	–	–	+	–	+	–	–	–	–	I	③⑤	–
1999	Annotation Engine	(Seltzer, 1999)	R	DB	HTML	+ +	+	+ –		+	–	+	+	–	–	–	–	–	+	P	⑤	–
1999	Annotator	(Ovsiannikov *et al.*, 1999)	R	DB	Texte brut	+ –	+	+ +		–	–	+	+	–	+	–	–	–	+	IPS	①⑤	–
1999	AnnoteImage	(Lober *et al.*, 2001)	R	Texte	IMG	– –	–	+ –		–	–	–	+	+	–	–	–	–	–	I	②	–
1999	Arakne	(Bouvin, 1999)	R	HTML	HTML	? ?	+	+ +		?	?	–	+	–	–	–	–	–	+	IP	⑤	–
1999	HyperPass	(Heck et Luebke, 1999)	R	HTML	HTML	+ +	+	+ –		?	–	+	+	–	–	–	–	–	+	IPG	①⑤	–
1999	Pharos	(Bouthors et Dedieu, 1999)	R	Texte	URL	+ +	+	– –		+	–	+	+	–	–	+	–	–	–	IPG	①④⑤⑥	+
1999	ThirdVoice*	(Margolis et Resnick, 1999)	I	?	HTML	+ +	+	– –		?	–	+	+	–	–	–	–	–	–	IPG	⑤	–
1999	WebVise	(Grønbæk *et al.*, 1999)	R	?	HTML, DOC, XLS	+ +	+	+ +		+	–	–	+	–	–	–	–	–	–	P	⑤⑥	–
2000	Anchored Conversations	(Churchill *et al.*, 2000)	R	Texte	DOC, HTML	+ –	+	+ –		+	–	–	+	–	–	–	–	–	+	IPGS	①②③⑤	–
2000	Collate	(Thiel *et al.*, 2004)	R	RDF	TIFF, JPG	+ +	+	+ ?		+	–	–	+	–	+	–	+	–	+	P	①⑤	–
2000	D3E / JIME	(Shum et Sumner, 2001)	R	Texte	HTML	+ +	+	+ –		+	?	+	+	–	–	–	+	–	+	PG	③⑤	–
2000	eNotate	`ischian.com/informal`	I	prop.	HTML, DOC, PPT, XLS	– –	+	+ +		–	–	–	–	+	–	–	–	+	–	I	③⑤	–
2000	E-Quill*	(Brush, 2002)	I	?	HTML, PS, VISIO	? ?	?	+ +		+	?	?	?	+	?	?	?	+	–	I	⑤	?
2000	iMarkup	(iMarkup Solutions Inc., 2000)	I	prop.	HTML, PDF	– –	+	+ –		+	+	–	+	+	+	+	–	+	–	I	①③④⑤⑥	+
2000	MS Office Web Discussions	(Cadiz *et al.*, 2000)	I	DB	HTML	+ –	+	+ –		+	+	+	+	–	–	–	–	–	+	IP	③④⑤	+
2000	WebAnn	(Brush *et al.*, 2002)	I	CAF	HTML, PPT, DOC, XLS	+ –	+	+ –		–	–	+	+	–	–	–	–	–	–	IP	⑤	+

Tableau I.3.1 – Comparaison des systèmes d'annotations (1/2)

système				format		auteur		ancrage			dates		contenu de l'annotation										
année	nom	référence	rech/indus	annotations	documents	identité	courriel	intégralité	sélection	*n* sélections	création	modification	titre	contenu textuel	encre digitale	tags	références	types	marques diverses	fils de discussion	visibilité	activités	notification
2000	Yawas	(Denoue et Vignollet, 2000a)	R	TXT	HTML	–	+	+	+	–	–	–	–	+	–	–	–	+	–	–	IP	①③⑤	–
2001	CAF	(Bargeron *et al.*, 2001)	I	XML-S	HTML	+	–	+	+	?	+	–	–	+	+	–	–	–	+	+	IP	③⑤	–
2002	Ann. Sys. for Sem. Web	(S et RKVS, 2002)	R	HTML	Texte brut	+	–	+	+	–	+	+	–	+	–	–	–	+	–	+	IP	①⑤⑥	–
2002	Annotea/Amaya	(Kahan *et al.*, 2002)	R	RDF	HTML	+	–	+	+	–	+	+	+	+	–	–	–	+	–	+	IP	③⑤	–
2002	UCAT	(Bottoni *et al.*, 2003)	R	Texte	HTML	+	–	+	+	–	+	+	+	+	–	–	–	+	–	+	IP	①③⑤	–
2003	ALT	(Gabrielli et Law, 2003)	R	?	?	–	–	–	+	–	–	–	–	–	+	–	–	–	+	–	S	⑤	–
2003	Annozilla	(Mozdev, 2003)	I	RDF	HTML	+	–	+	+	–	+	–	–	+	–	–	–	+	–	+	IP	⑤	–
2003	IPSA	(Agosti *et al.*, 2005)	R	DB	Image	+	–	+	+	–	+	+	–	+	–	–	+	+	–	+	IPG	④⑤	–
2004	Connotea	(Lund *et al.*, 2005)	I	RDF	HTML	+	+	+	–	?	+	–	+	+	–	+	–	–	–	–	IPG	①④⑤⑥	+
2004	Digital Graffiti	(Carter *et al.*, 2004)	R	prop.	Image	+	+	–	?	–	+	–	–	+	+	–	–	–	+	+	P	⑤	–
2004	Dinosys	(Desmontils *et al.*, 2004)	R	Texte	HTML	+	+	+	+	–	+	–	+	+	–	–	–	+	–	–	IP	⑤	–
2004	Flick'r	`flickr.com`	I	prop.	Image	+	+	+	+	–	+	–	+	+	–	+	–	–	–	+	IPG	①④⑤⑥	+
2004	Madcow	(Bottoni *et al.*, 2004)	R	XML	HTML, Image, Vidéo	+	–	+	+	+	+	+	+	+	–	–	+	+	–	–	IPG	①③⑤	–
2004	Microsoft OneNote*	(Mock, 2004)	R	prop.	Image	–	–	–	+	+	–	–	–	–	+	–	–	–	+	–	P	③⑤	–
2004	PDF Annotator	(GRAHL software, 2004)	I	prop.	PDF	–	–	–	–	–	–	–	–	–	+	–	–	–	+	–	I	⑤	–
2004	Scrapbook	(Ma et Murota, 2006)	R	Texte	HTML	–	–	+	+	–	+	–	+	+	–	–	–	–	–	–	I	①③⑤⑥	–
2004	ScreenCrayons	(Olsen *et al.*, 2004)	R	?	Image	–	–	+	+	–	–	–	–	+	+	+	–	–	+	–	I	③⑤⑥	–
2005	AnT&Cow	(Lortal *et al.*, 2006)	R	RDF	HTML	+	–	+	+	+	+	–	–	+	–	–	–	+	–	+	IP	⑤	–
2005	B-Glaciologist	(Fogli *et al.*, 2004)	R	XML	Images SVG	+	–	+	+	–	–	–	+	+	–	–	+	–	–	–	GS	⑤	–
2005	DocAnnot	(Bringay *et al.*, 2005)	R	Texte+	HTML	+	–	+	+	–	+	+	+	+	–	–	+	–	–	–	IPGS	①⑤⑥	–
2005	TafAnnote	(Cabanac, 2005)	R	prop.	HTML	+	+	+	+	–	+	+	+	+	–	+	+	+	–	+	IP	①③④⑤⑥	+
2005	Yawas for Firefox	(Denoue, 2005)	R	Texte	HTML	+	+	+	+	–	–	–	–	+	–	–	–	+	–	–	IP	①③⑤	–
2006	Diigo	`diigo.com`	I	prop.	HTML	+	+	+	+	–	+	–	+	+	–	+	–	–	–	–	IPG	①④⑤⑥	?
2006	OAS	(Harmon, 2007)	R	HTML	Image	+	+	+	+	+	+	–	–	+	+	–	–	–	+	–	IPGS	③⑤	–
2006	Adobe Reader		I	prop.	PDF	+	–	–	+	–	+	–	–	+	+	–	–	–	+	–	I	⑤	–
2006	SportsAnno	(Lanagan et Smeaton, 2007)	R	Texte	HTML	+	–	+	+	–	+	–	–	+	–	–	–	–	–	+	P	⑤	–
2006	Stickis	`stickis.com`	I	prop.	HTML	+	+	+	–	–	+	?	–	+	–	+	–	–	–	–	IPGS	④⑤⑥	+
2007	CWS "Augmented CACM"	(Freyne *et al.*, 2007)	R	Texte	HTML	+	–	+	–	–	–	–	–	+	–	–	–	–	–	–	IPG	①⑤	–
2007	Digg	`digg.com`	I	Texte	URL	+	+	+	–	–	+	–	–	+	–	–	–	–	–	+	P	①	–
2007	Notate	`a.nnotate.com`	I	HTML	HTML	+	–	+	+	–	+	–	–	+	–	+	+	–	–	+	IPG	①③④⑤⑥	–
2007	Notebook	`google.fr/notebook`	I	Texte	HTML	+	+	+	+	–	–	–	+	+	–	+	–	–	–	–	IPGS	①②④⑥	–
2008	Armarius	(Doumat *et al.*, 2008)	R	Texte	Image	+	–	–	+	–	–	–	–	+	–	+	–	+	–	–	IPG	①⑤	–

Tableau I.3.2 – Comparaison des systèmes d'annotations (2/2)

3.2.4 Limites de l'activité d'annotation collective

Bien que faisant l'objet de développements depuis vingt ans déjà et malgré les nombreuses preuves de l'utilité des annotations (section I.3.1), les systèmes d'annotation ne sont pas adoptés par les individus. Pourtant, on peut constater qu'ils recherchent les finalités de l'annotation électronique sans toutefois employer de SA à proprement parler. Ils créent par exemple des *(social) bookmarks*, ils s'envoient des courriels à eux-mêmes pour conserver une bribe de document, en collectent d'autres au sein d'un proto-document qu'ils créent avec un logiciel de traitement de texte… Sur la base de ce constat, de nombreux travaux dont (Ovsiannikov *et al.*, 1999; Vasudevan et Palmer, 1999; Wolfe et Neuwirth, 2001; Brust et Rothkugel, 2007) élicitent les limites des SA et les obstacles auxquels ils sont confrontés, freinant de fait leur adoption à plus grande échelle. À partir de ces études et de notre propre expérience, nous mettons en lumière dans cette section deux limites majeures auxquelles nous proposons une solution dans la partie II :

1. pour favoriser leur adoption, nous pensons que les SA devraient couvrir l'ensemble des activités documentaires (figure I.1.1) de façon à éviter à l'usager de recourir à un système par activité, ce qui entraîne une charge cognitive élevée. De plus, les résultats des activités de chaque membre d'un groupe devraient être capitalisés et réinjectés a) au niveau de chaque individu et b) au niveau des autres activités documentaires sur le principe du donnant-donnant, afin de favoriser un enrichissement mutuel ;
2. la représentation des annotations en contexte, bien que pertinente pour afficher peu d'annotations, peut déranger le lecteur de documents massivement annotés. Par exemple, la figure I.3.6 illustre ce problème de passage à l'échelle avec Amaya/Annotea (Kahan *et al.*, 2002) du W3C, où les annotations qui sont intégrées dans le texte parasitent littéralement la lecture. Nous illustrons plus amplement ce problème à l'aide d'une vidéo montrant le système en action : « `http://www.irit.fr/~Guillaume.Cabanac/annotation/demoAmaya.wmv` ».

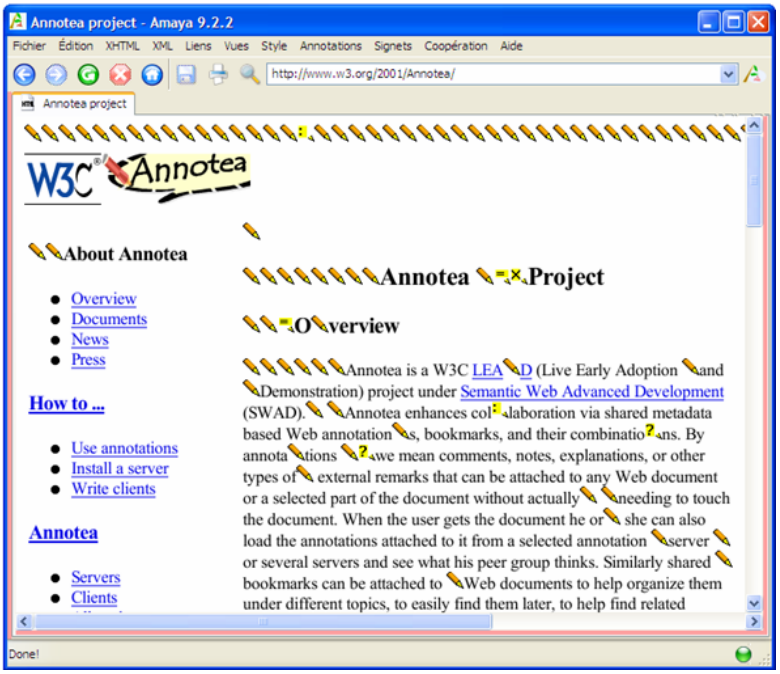


Figure I.3.6 – Problème du passage à l'échelle de la représentation des annotations en contexte.

De plus, la figure I.3.6 montre qu'aucune indication ne permet au lecteur de sélectionner une annotation plutôt qu'une autre : il lui est impossible de discerner l'étendue d'une annotation (seul le début du point d'ancrage étant représenté), de distinguer les corrections (coquilles, fautes de grammaire, etc.) des véritables questions, de voir quelles annotations ont suscité une discussion... Par ailleurs, la visualisation d'un débat (figure I.3.5) ne présente pas les différentes opinions de chaque réponse, alors que Wolfe (2000) souligne que c'est une nécessité pour l'individu qui réalise une lecture active. Ce manque d'informations « périphériques » (l'opinion n'est accessible que lorsqu'on « ouvre » une réponse en cliquant dessus) complique la compréhension du débat. En effet, le lecteur doit consulter chaque réponse, identifier l'opinion des arguments correspondants et en faire une synthèse pour savoir si l'annotation initiale est validée ou non par le groupe qui a débattu. De fait, cette tâche demande une charge cognitive importante. Or, tout effort mental en marge de la lecture elle-même devrait être évité à tout prix selon O'Hara et Sellen (1997).

3.3 Vers l'annotation collective de documents électroniques

Actuellement, dans le cadre de l'évolution du Web qui a été baptisé « Web 2.0 » par O'Reilly (2005) ou qualifié plus généralement de « social » et « participatif », les systèmes d'annotations tendent à considérer et à exploiter les annotations collectives en tant que contributions spontanées des individus. Elles sont perçues comme une valeur ajoutée aux ressources sur lesquelles elles sont ancrées et sont alors mises à contribution pour améliorer les fonctionnalités existantes des SA et en proposer de nouvelles. En fait, Marshall et Brush (2004) montrent que les lecteurs modifient en général leurs annotations privées lorsqu'ils les rendent publiques, en reformulant leur contenu pour le rendre plus intelligible, notamment.

Nous rapportons dans cette section les principales approches visant à tirer parti des annotations électroniques. Par exemple, une expérience de Golovchinsky *et al.* (1999) montre que les passages annotés par un lecteur sont de meilleurs indicateurs du retour de pertinence (*relevance feedback*) que leurs jugements explicites. S'appuyant sur ce résultat, XLibris (Price *et al.*, 1998) affiche des recommandations de lecture (hyperliens vers les documents jugés pertinents par le système) à côté des annotations formulées par le lecteur. Dans un contexte de partage des annotations, Marshall (1998) propose la fonctionnalité du « consensus des lecteurs », permettant de ne visualiser que les parties d'un document qui ont été surlignées par le plus grand nombre de lecteurs. Appliquée aux livres des bibliothèques universitaires par exemple, cette fonctionnalité permettrait d'obtenir un résumé des points clés selon le point de vue des lecteurs successifs.

Par ailleurs, Fraenkel et Klein (1999) montrent que la prise en compte des annotations dans le processus de RI permet d'en améliorer le rappel et la précision. Cette amélioration est due au fait qu'un annotateur reformule généralement le passage annoté avec ses propres mots, qui sont alors considérés par le système comme une description alternative du passage annoté. Par la suite, le système FAST (Agosti et Ferro, 2005, 2007) enrichit le résultat d'une recherche en y intégrant les document trouvés indirectement, à partir des fils de discussion qui y sont ancrés. Frommholz et Fuhr (2006) améliorent cette proposition en prenant en compte la polarité (positive ou négative, mais pas graduelle) des arguments exprimés dans une annotation, de façon à ne pas favoriser des documents critiqués dans les fils de discussion associés. Enfin, Hansen (2006) décrit l'avènement

des annotations pervasives : réalisées à l'aide de dispositifs mobiles, tels qu'un téléphone, en prenant en photo un bâtiment puis en y associant un commentaire et en postant cette annotation sur un serveur partagé avec les membres de sa communauté...

La première partie de ce mémoire a détaillé les activités documentaires formant le cycle de vie du document (figure I.1.1). Nous avons considéré le support papier puis le support électronique et focalisé notre étude sur la pratique d'annotation et sa transposition du papier au numérique. À partir des limites que nous avons mis en lumière, nous exposons dans la deuxième partie du présent mémoire notre approche visant à améliorer les activités documentaires de l'organisation par la pratique d'annotation collective.

Deuxième partie

Fédérer et améliorer les activités documentaires de l'organisation

1 Aperçu synthétique de la contribution

"Yesterday's solutions are today's problems."

Bruce E. Spivey (1934 —)

CETTE thèse s'inscrit dans le cadre général de la relation individus-documents au sein d'une organisation. Nous nous intéressons en particulier aux « travailleurs du savoir », ces individus qui travaillent principalement avec de l'information et qui en produisent au sein d'organisations. L'étude de leurs activités documentaires, schématisées par le cycle de vie du document de Sellen et Harper (2003, p. 203) représenté en figure I.1.1 (p. 8), a mis en lumière les trois problématiques principales suivantes :

1. les tâches réalisées au quotidien requièrent la maîtrise de plusieurs systèmes. En effet, la section I.2 a montré qu'un individu met en œuvre au moins six systèmes distincts pour mener à bien l'ensemble des activités documentaires. Cette contrainte implique principalement une charge cognitive importante pour l'usager ;
2. les systèmes sont très spécialisés et ne communiquent pas entre eux. De ce fait, chaque système perçoit ses usagers de manière parcellaire, sans pouvoir en compléter sa représentation en échangeant des données avec les autres systèmes. Par conséquent, toute assistance proposée par un tel système se révèle être sous-optimale ;
3. les EPI des individus sont de véritables mines d'informations pertinentes eu égard aux activités de l'organisation. Ils résultent des efforts de recherche, de filtrage, de consolidation… mis en œuvre par les individus au prix de coûteux efforts. Paradoxalement ils ne sont pas valorisés au niveau de l'organisation, entraînant un retour sur investissement très limité.

Cette deuxième partie du mémoire expose notre contribution : la fédération et l'amélioration des activités documentaires par l'annotation collective. Cette contribution originale vise à limiter les trois problématiques identifiées, tout en enrichissant la relation usagers-documents, toujours dans le contexte organisationnel. De ce fait, elle comprend deux volets, chacun étant étayé par

diverses propositions. Le premier volet concerne la modélisation unifiée des activités documentaires des membres organisationnels. De façon complémentaire, le second volet traite de l'exploitation du capital documentaire organisationnel ainsi constitué, qui est actuellement en sommeil.

1.1 L'annotation collective pour fédérer les activités documentaires

Afin d'apporter une réponse aux problématiques précédemment identifiées, nous proposons un modèle unifié pour couvrir les activités documentaires du cycle de vie du document. L'annotation collective est au cœur de ce modèle, en tant que vecteur d'information transversal aux activités documentaires. Sur ce modèle unifié repose alors l'architecture logicielle détaillée dans cette deuxième partie, visant à répondre à deux types de besoins complémentaires suivants : fournir une assistance personnalisée ainsi que collective.

1.1.1 Fournir une assistance personnalisée

L'architecture proposée repose sur un modèle unifié fédérant les usagers et les six activités documentaires, de façon à remédier aux représentations (profils) partielles et éparpillées dans plusieurs applications. Adossée à des processus intégrés, cette architecture apporte une assistance à l'usager en situations de recherche ①, de rédaction ②③, de distribution ④, d'exploitation ⑤ et d'organisation ⑥ d'information. L'architecture proposée capitalise également les résultats produits par un usager réalisant une de ces activités, afin de les réinjecter dans les autres activités qui s'enrichissent alors mutuellement. Enfin, l'usager est aidé dans son activité globale, ce qui profite notamment à la construction et à l'évolution quotidienne de son espace documentaire.

1.1.2 Fournir une assistance collective

Nous faisons l'hypothèse que les membres organisationnels partagent des intérêts, des connaissances, des tâches et des activités car ils évoluent dans cette même organisation. Au cœur de nos travaux, nous considérons leurs activités afin de leur porter assistance. L'architecture multi-utilisateurs proposée intègre donc des processus sur le principe du donnant-donnant, afin d'accroître le retour sur investissement relatif aux EPI des usagers. Jusqu'alors manifestement considérés comme un capital documentaire en sommeil, nous les exploitons en tant que vecteurs d'une forte valeur ajoutée focalisée sur les activités de l'organisation. En retour, les usagers bénéficient du système en obtenant de l'information relative à leurs activités, cette information provenant alors du groupe dans son ensemble. De ce fait, toute information introduite dans l'organisation — extraite de sources externes, comme produite par ses membres — devient profitable pour l'ensemble du groupe, alors qu'elle végète bien souvent dans les EPI des membres organisationnels.

L'architecture multi-utilisateurs proposée a notamment pour but d'aider chaque individu à partir des activités documentaires du groupe et vice versa. L'assistance qui est apportée au niveau microscopique (chaque usager est aidé) profite également au niveau macroscopique lorsqu'on considère le capital documentaire introduit, filtré, consolidé et structuré constituable sur la base des EPI. En complément, le second volet de notre proposition concerne la conception d'une interface de visualisation et d'exploration de capital organisationnel.

1.2 Exploitation du capital documentaire organisationnel en sommeil

Le second volet de notre contribution porte sur l'exploitation non-intrusive du résultat des activités documentaires réalisées par les membres organisationnels. Nous tirons parti de l'EPI de chaque individu afin de visualiser le capital documentaire acquis par l'organisation dans son ensemble. La visualisation proposée au travers d'une interface multi-facettes représente à la fois les documents et les individus, selon deux dimensions complémentaires : leurs thématiques et leurs contextes d'usage. L'interaction usager-interface permet aux individus d'explorer le contenu du capital documentaire, afin de répondre aux besoins du pilotage de l'organisation comme à ceux des membres organisationnels.

Cette deuxième partie du mémoire est structurée comme suit. Le chapitre II.2 définit les concepts d'annotation collective et d'Espace Personnel d'Annotations (EPA). Sur la base de l'état de l'art présenté dans la partie I, nous élicitons trois catégories d'annotations destinées à *commenter*, *mémoriser* et *débattre*. Ces divers éléments sont rassemblés dans un modèle unifié à partir duquel six processus intégrés opèrent. Parmi ces processus, certains exploitent la « validité sociale » des annotations argumentatives, reflétant à quel point le groupe social qui s'exprime dans leur fil de discussion est d'accord avec l'annotation initiale. Ce concept et les algorithmes associés font l'objet du chapitre II.3. Similairement, d'autres processus tirent profit de la mesure de similarité d'usage entre documents, établie en tirant parti de l'exploitation des EPA des membres organisationnels. Cette mesure complémentaire aux mesures de similarité sur le contenu des documents est formalisée dans le chapitre II.4.

Étant données ces deux mesures auxiliaires (validation sociale et similarité d'usage) le chapitre II.5 expose les processus intégrés au modèle unifié, destinés à améliorer chacune des six activités documentaires, sur les principes du donnant-donnant et de la non-intrusion. Puis, le chapitre II.6 détaille l'interface multi-facettes conçue pour visualiser et explorer le capital documentaire formé par les EPA des membres organisationnels. Enfin, le chapitre II.7 discute les limites relatives à l'ensemble des propositions de cette partie, avant de la conclure par une synthèse de notre contribution.

2 Modélisation unifiée des six activités documentaires

> "Concentrate on capturing and *reproducing* the appearance of marks made by knowledge workers rather than interpreting them. These marks made on paper, screen (or indeed any other physical surface from cave wall to whiteboard) is *how* people change their environment in order to carry information from place to place or time to time. They are also used to externalise their own thinking — a type of scaffolding whilst they are in the process of informing themselves."
>
> *Alison Kidd (1994)*

L'ÉTAT DE L'ART présenté dans la première partie du mémoire a mis en lumière le contexte de nos travaux : la relation usagers-documents dans le cadre d'une organisation. Nous y avons détaillé les activités documentaires formant le cycle de vie du document, en considérant les documents papiers puis les documents électroniques. Enfin, nous avons focalisé notre étude sur la pratique d'annotation qui est transversale aux activités documentaires, tant elle est partie intégrante de la rédaction ②, de la finalisation ③ et de l'exploitation ⑤ des documents.

Sur la base des éléments identifiés dans l'état de l'art et afin de répondre aux problématiques identifiées — l'usager ne bénéficie que trop peu des nombreux systèmes qui ne communiquent pas entre eux — ce chapitre expose un modèle unifié pour les activités documentaires. Il permet la mise en œuvre de processus visant à décloisonner les activités documentaires de façon à privilégier leur enrichissement mutuel. Le modèle proposé représente les usagers ainsi que les documents qu'ils exploitent et gèrent. Ces deux entités sont liées grâce au concept d'annotation collective, retenu en tant qu'élément fédérateur des activités documentaires. C'est sur ce modèle unifié que reposent nos propositions visant à enrichir la relation usagers-documents, détaillées dans les chapitres suivants.

2.1 Définition des éléments constituant le modèle unifié

Cette section détaille les concepts sur lesquels nous établissons notre contribution : l'amélioration des activités documentaires et l'enrichissement du rapport usagers-documents. Pour ce faire, nous définissons dans un premier temps les entités correspondant aux usagers et à leur Espace Personnel d'Annotations (EPA). Puis, nous établissons une typologie des annotations collectives, où l'annotation explicite un lien d'utilité établi entre usagers et documents.

2.1.1 Individus, documents et espaces personnels d'annotations

Dans le but d'améliorer les activités documentaires des membres organisationnels, nous devons tout d'abord modéliser les éléments fondamentaux manipulés par l'architecture logicielle :

Les individus. Ce sont les membres organisationnels qui réalisent les activités documentaires. Au sein de l'organisation, ils appartiennent à un ou plusieurs groupes. Ces groupes peuvent être formels ou informels, selon qu'ils sont issus de l'organigramme de l'organisation ou créés *ad hoc*, pour les besoins d'un projet par exemple ;

Les documents. Les individus consultent des documents électroniques, également appelés ressources, atteignables par leur URL (Uniform Resource Locator), tels que « `file:///C:/MesDocuments/these.pdf` » ou « `http://www.irit.fr` ». Un tel document est dématérialisé et exprimé dans un format particulier : HTML, PDF, DOC...

Les espaces personnels d'annotations. Au quotidien, chaque individu conserve des documents ainsi que des signets vers des documents dans son Espace Personnel d'Information. La section I.2.5 en mentionne les multiples objectifs, le principal étant l'accès ultérieur facilité. Or, Kidd (1994) indique que de nombreuses situations de travail nécessitent davantage la mémorisation des passages d'intérêt (quelques phrases, une définition, une photo, un horaire de train, par exemple) avec les notes du lecteur en contexte, que le simple stockage du document dans son intégralité (section I.2.6). Par conséquent, nous modélisons dans notre approche un espace personnel d'annotations par usager où il peut y stocker et organiser les annotations qu'il crée.

D'après l'étude de la pratique d'annotation réalisée en partie I, nous avons constitué trois catégories d'annotations collectives utiles dans le cadre des activités documentaires que les individus réalisent au sein de l'organisation.

2.1.2 Typologie des annotations collectives selon leur objectif

Cette section définit le concept d'annotation collective qui est au cœur de notre proposition. Nous détaillons ses caractéristiques générales avant de le raffiner en trois catégories, selon l'objectif de l'annotation : *commenter*, *mémoriser* et *débattre*.

Définition 1. Le concept d'*annotation collective* présenté dans (Cabanac *et al.*, 2007b) fait référence à une annotation partageable entre les différents usagers du système. Ces derniers peuvent alors consulter une telle annotation et y répondre, formant ainsi un fil de discussion (définition 2). Une annotation collective est définie par le couple $\langle DO, IS \rangle$ explicité ci-dessous.

1. La composante *DO* représente les Données Objectives créées par le système d'annotation. Elle regroupe les attributs obligatoires (non nuls) suivants :
   – son *identification* grâce à un identifiant unique tel qu'une URL ;
   – l'*identité de son auteur* qui regroupe les informations telles que son nom et son courriel ;
   – son *estampille temporelle*, c'est-à-dire sa date de création permettant au lecteur d'identifier les nouvelles annotations créées depuis sa dernière visite de la ressource, ainsi que d'organiser les arguments du fil de discussion chronologiquement ;
   – son *emplacement* éventuel dans un répertoire de l'EPA de son créateur ;
   – ses *points d'ancrage* qui spécifient de manière non ambiguë sa localisation au sein de la ressource annotée. De nombreuses techniques d'ancrage ont été proposées dans la littérature (section I.3.2.2.1), notamment XPointer (DeRose *et al.*, 2002) pour le format HTML.

2. La composante *IS* représente les Informations Subjectives formulées par l'individu qui crée l'annotation. Elle regroupe les attributs optionnels suivants :
   – son *contenu*, tel qu'un commentaire textuel ou un enregistrement audio ;
   – sa *visibilité* (privée, publique, accessible uniquement à certains usagers) qui permet de restreindre sa diffusion et sa consultation ;
   – l'*expertise* de son créateur, qu'il estime lui-même subjectivement. Marshall (1998) rapporte que c'est une indication utile aux futurs lecteurs qui tendent à davantage avoir confiance aux experts qu'aux débutants ;
   – le *jugement* de l'annotateur quant au passage annoté (négatif, neutre ou positif) ;
   – une *liste de références* fournie par l'annotateur qui désire justifier ses arguments, par exemple. Ainsi, la cote d'un livre, une citation, une URL… peuvent y être associées ;
   – des *tags*, c'est-à-dire des termes descriptifs choisis par l'annotateur afin de décrire son annotation. Par la suite, l'usager peut accéder à ses annotations en sélectionnant un de ses tags, de façon similaire aux approches de *social bookmarking* (Hammond *et al.*, 2005) ;
   – divers *types d'annotation* permettant à un annotateur de fournir un aperçu de la sémantique de son annotation. Ils sont basés sur les types proposés par Kahan *et al.* (2002), auxquels nous avons adjoint des types d'opinion afin de répondre aux besoins des lecteurs identifiés par Wolfe (2000). Nous avons divisé ces types d'annotation représentés dans le tableau II.2.1 en deux classes : « commentaire » et « opinion ».

Classe	Commentaire			Opinion (types exclusifs)		
Type	question	modification	exemple	réfutation	neutre	confirmation
Notation	$\mathscr{Q}$	$\mathscr{M}$	$\mathscr{E}$	$\mathscr{R}$	$\mathscr{N}$	$\mathscr{C}$

Tableau II.2.1 – Types d'annotation disponibles pour une annotation collective.

D'une part, la classe « commentaire » reflète la sémantique de l'annotation par rapport à la ressource annotée : question, modification ou exemple. D'autre part, la classe « opinion » offre une sémantique de l'opinion qu'a voulu exprimer l'annotateur : réfutation, neutre ou confirmation. En combinant les types de ces deux classes, les annotateurs décrivent des points de vue graduels subjectifs. Par exemple, on peut considérer une annotation typée $\mathscr{R}$ *versus* une annotation typée $\mathscr{RE}$. Concrètement, ces types sont fournis par les annotateurs eux-mêmes, ou bien inférés à partir du contenu de l'annotation *via* des techniques d'identification d'opinion (Pang *et al.*, 2002; Liu, 2007) puis validés par leurs auteurs.

Toute annotation collective peut être consultée par des lecteurs et susciter des réactions, exprimées sous la forme de réponses qui constituent alors un fil de discussion (définition 2).

Définition 2. Un *fil de discussion* est la représentation d'un débat sous la forme d'une hiérarchie d'arguments. La racine de l'arbre est l'annotation qui suscite la discussion. Ce nœud spécifique peut faire l'objet de réponses. Récursivement, les réponses suscitent éventuellement à leur tour d'autres réponses. Elles sont ordonnées chronologiquement grâce à leurs estampilles temporelles. La figure II.2.1 représente une annotation ancrée sur un passage d'un document. Cette annotation fait l'objet de trois réponses dont les types sont indiqués par des pictogrammes visuels. À leur tour, certaines réponses de cet exemple ont suscité d'autres réponses.

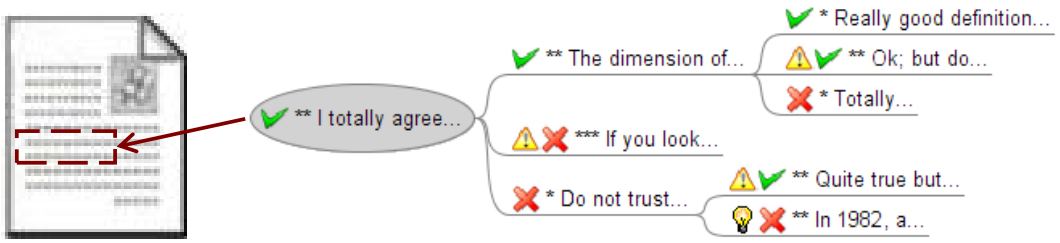


Figure II.2.1 – Exemple d'un document dont une annotation en contexte a suscité un débat.

Dans le cadre des activités documentaires, nous considérons que le concept d'annotation collective est trop « large » : ses attributs ne sont pas tous valués selon l'objectif de l'annotateur. Par exemple, un individu qui indique une coquille dans un document n'a certainement pas besoin de conserver cette correction dans son EPA. Afin de mieux cadrer avec les objectifs des individus, nous proposons dans les sections suivantes trois catégories d'annotation. Ces catégories reflètent les objectifs de *commenter*, *mémoriser* et *débattre*.

2.1.2.1 Annotation collective *remarque* : pour corriger ou commenter

L'*annotation-remarque* permet de commenter un passage de document sans pour autant obliger l'annotateur à la classer dans son EPA. L'utilité de cette catégorie d'annotation est visible lors de l'activité de finalisation ③ qui génère de nombreuses annotations (corrections de coquilles, indications de typographie, etc.) que l'annotateur n'a pas besoin de conserver dans son EPA. Par contre, il est essentiel que ses collègues et lui-même les voient lorsqu'ils visualisent le document annoté, par la suite.

2.1.2.2 Annotation collective *stockage* : pour mémoriser

L'*annotation-stockage* étend la notion de signet Web en permettant la mémorisation de passages de documents, au lieu de leur intégralité. Le créateur d'une telle annotation l'organise obligatoirement dans son EPA, qui reprend la structure des signets : une arborescence de répertoires. Une raison motivant le choix d'une hiérarchie fait écho à l'étude de Jones *et al.* (2005) soulignant le fait que les individus ont besoin de classer leurs informations dans une hiérarchie pour réaliser leurs activités, notamment leurs projets.

2.1.2.3 Annotation collective *argumentative* : discuter et débattre

L'*annotation-argumentative* est ancrée sur une partie d'un document et peut faire l'objet de réponses au sein de son fil de discussion. Cette catégorie d'annotation permet de créer des espaces de discussion similaires aux forums du Web ou de Usenet, tout en conservant la discussion sur le document annoté. Le fait de ne pas avoir à préciser le contexte de la discussion (qui est défini par le point d'ancrage dans le document, représenté en pointillés dans la figure II.2.1) est un avantage indéniable de cette approche par rapport aux forums. Par ailleurs, les lecteurs accèdent aux discussions relatives à un document donné sans avoir à connaître l'URL de la discussion, alors que cela est nécessaire actuellement avec une approche à base d'un forum distinct du document discuté.

La section suivante présente le modèle unifié correspondant aux éléments décrits jusqu'alors : individus, documents, catégories d'annotations et EPA.

2.2 Modèle unifié des activités documentaires

Les concepts manipulés afin de mettre en œuvre l'architecture proposée sont exposés en deux temps. La section II.2.2.1 modélise l'annotation collective ancrée sur les ressources et éventuellement stockée dans les EPA (Cabanac *et al.*, 2006a, 2007c, 2008a,b). Cette modélisation est ensuite complétée dans la section II.2.2.2 en ajoutant la notion de groupes d'usagers, ainsi que des éléments nécessaires aux processus intégrés qui sont détaillés dans le chapitre II.5. Les diagrammes de classes UML auxquels nous recourrons conservent un niveau élevé d'abstraction : nous cachons intentionnellement les compartiments des classes relatifs aux attributs et opérations, qui relèvent davantage de l'implantation et qui ne serviraient en rien le discours. Toutefois, nous mentionnerons dans le texte les éléments requis pour la compréhension du modèle.

2.2.1 Modélisation de l'annotation collective et des EPA

Nous commenterons le diagramme de classes UML de la figure II.2.2 à l'aide du scénario suivant. Un Usager visualise une Ressource telle qu'une page Web, un planning sur l'Intranet de son entreprise, une photo stockée dans son arborescence personnelle… Pour conserver tout ou partie de cette ressource il crée une instance dérivée d'une *AnnotationAncrée*. Selon la catégorie de l'annotation créée, son stockage dans un Répertoire de l'EPA de l'*Usager* est requis (AnnotationStockage) ou facultatif (AnnotationRemarque et AnnotationArgumentative). Dans un cas comme dans l'autre, l'annotation créée est visible sur la ressource annotée, le fait de l'insérer dans un Répertoire permet en plus d'y accéder directement depuis son EPA.

Concernant l'*AnnotationAncrée* créée, la notion d'ancre fait référence au passage du document que l'usager désire conserver. Deux situations peuvent alors survenir. D'une part, pour conserver l'intégralité de la ressource, le système emploie un AncrageGlobal qui correspond à un signet classique en stockant l'URL de la ressource, par exemple. Dans le cas contraire, l'usager ne désire stocker qu'une partie de la ressource : deux phrases non contiguës, par exemple. Pour ce faire, une alternative consiste à modifier la ressource en y intégrant des marqueurs spécifiques référençant le début et la fin de la sélection de l'usager. Cette solution n'est bien entendu possible

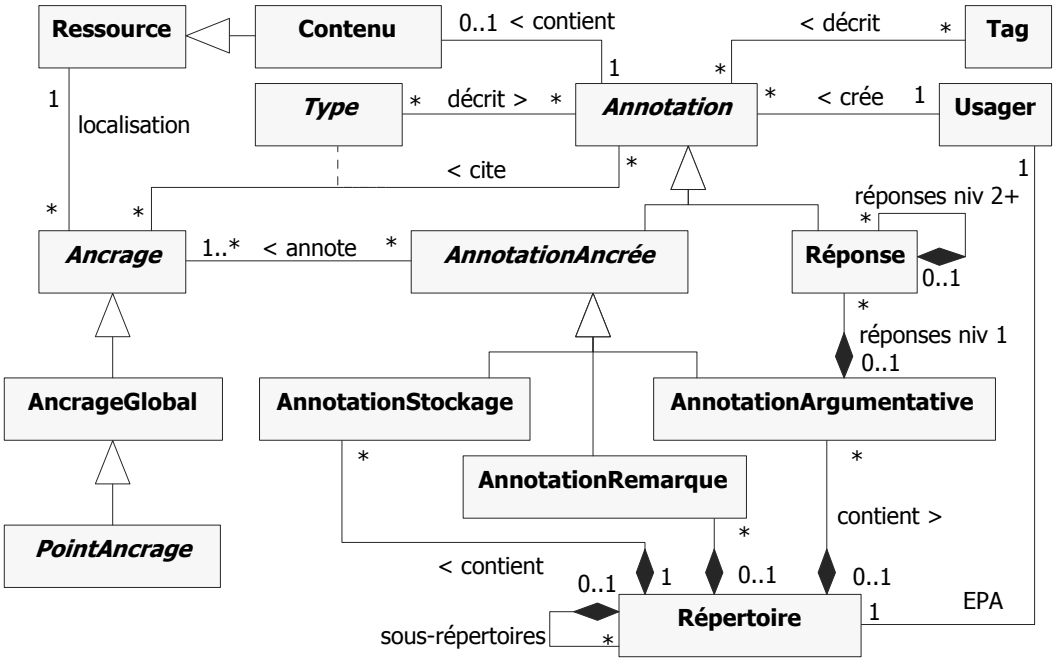


Figure II.2.2 – Diagramme de classes UML du concept d'annotation collective et des EPA.

qu'avec des documents modifiables, situation marginale dans le cas de la consultation de documents publics, notamment sur le Web. C'est pourquoi nous avons opté pour une seconde alternative qui consiste à stocker un point d'ancrage (*PointAncrage*) qui localise la sélection de l'usager de façon univoque. Afin de prendre en compte divers formats de documents (HTML, SVG, RTF, PDF, etc.), la classe abstraite *PointAncrage* doit être raffinée, c'est-à-dire sous-classée. Pour le format HTML par exemple, la syntaxe XPointer (DeRose *et al.*, 2002) exprimant la sélection de l'usager au sein de la structure logique du document (Document Object Model) peut être implantée dans une sous-classe de *PointAncrage*. À la restitution d'un tel point d'ancrage, il convient de s'assurer que l'annotation n'est ni *orpheline* ni *trompeuse*, comme indiqué en section I.3.2.2.1 (p. 30).

Selon la définition 1, le concept d'annotation collective s'articule autour des notions de Données Objectives (*DO*) et d'Informations subjectives (*IS*). Les classes mentionnées jusqu'à présent représentent les Données Objectives que le système d'annotation déduit du passage sélectionné dans le document. Par la suite, les classes restantes représentent les Informations Subjectives introduites par l'usager qui crée l'annotation. Toute *Annotation* possède un Contenu pour stocker les commentaires des individus, sans restriction quant au médium ou au format employés : du texte formaté ou un enregistrement audio, par exemple. En tant que sous-classe d'une Ressource, tout ou partie d'un Contenu peut être annoté à son tour, rendant ainsi possible la création d'annotations imbriquées (Bouvin *et al.*, 2002) utiles pour la rédaction ou l'interprétation collectives de document. Pour répondre à une AnnotationArgumentative l'usager crée une Réponse ; en tant que sous-classe d'*Annotation* son contenu est aussi riche (tags, références, etc.) tout en ne nécessitant pas l'*Ancrage* sur une ressource car elle est rattachée à l'AnnotationArgumentative associée.

Outre un commentaire, le créateur de l'annotation peut indiquer des références à d'autres (parties de) ressources. De plus, il peut associer à ces références et à l'annotation elle-même des *Types* par l'intermédiaire des relations « cite » et « décrit ». Cette classe abstraite *Type* décrit l'in-

tention de l'annotateur en fournissant un aperçu de sa sémantique : les sous-classes peuvent modéliser des taxonomies d'objectifs (par ex. : commentaire, exemple, question), d'actions (par ex. : à faire, à lire), d'opinions (par ex. : réfutation, neutre, confirmation) ou de concepts spécifiques au domaine d'application (par ex. : partenaire, produit, concurrence dans le domaine de la veille stratégique et technologique). Doter les membres organisationnels de telles taxonomies pourra les aider à décrire l'information qu'ils extraient à l'aide d'un référentiel commun, afin d'en améliorer la compréhension par exemple. De façon complémentaire au sous-classement de la classe *Type*, la classe Tag permet aux usagers de décrire une annotation à l'aide de leurs propres termes, dans la même optique que les systèmes de *social bookmarking* (Hammond *et al.*, 2005).

2.2.2 Modélisation des éléments requis par les processus intégrés

Le second diagramme de classes représenté en figure II.2.3 correspond à la prise en compte de l'aspect multi-utilisateurs afin d'enrichir leurs relations avec les documents introduits dans l'organisation. De plus, certains de ses éléments sont requis par les processus exposés dans le chapitre II.5. Concrètement, ce diagramme complète le premier diagramme (figure II.2.2) dont les classes apparaissent avec un fond blanc, contrairement aux nouveaux éléments ayant un fond gris. Les Usagers peuvent appartenir à des Groupes, décomposables en sous-groupes. Ils accordent à d'autres *Entités* de l'organisation — aussi bien Usagers que Groupes — des Droits d'accès aux Répertoires de leur EPA. Cette opportunité permet alors de partager manuellement les ressources d'intérêt au travers des annotations.

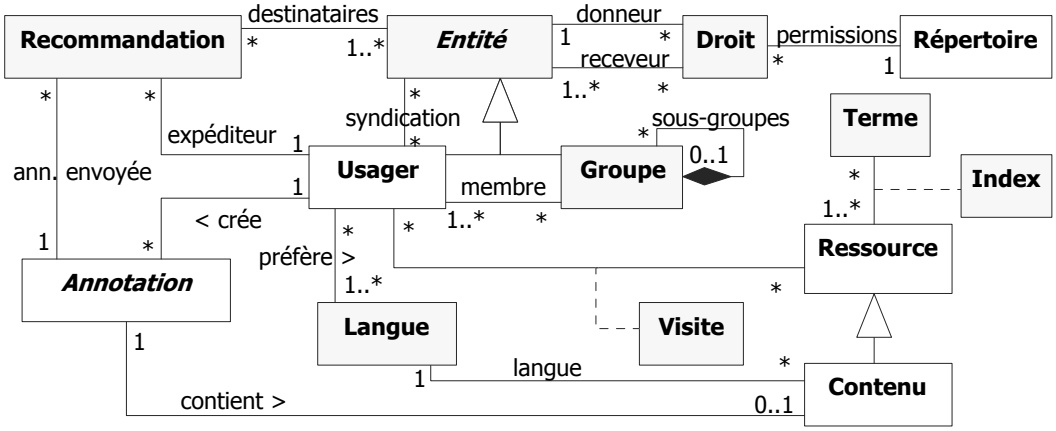


Figure II.2.3 – Diagramme de classes UML complétant la modélisation de la figure II.2.2.

Pour mettre en œuvre les processus intégrés, les données suivantes sont modélisées. Lorsqu'un Usager consulte une Ressource, le système conserve une trace de cette Visite en stockant la date associée. La date de création (ou de dernière modification, le cas échéant) est également conservée pour les Contenus des annotations, qui peuvent faire l'objet de Recommandations selon les préférences de l'usager en termes de Langue, notamment. De plus, les Usagers peuvent « syndiquer » des *Entités*, ce terme originellement associé aux flux RSS (*Really Simple Syndication*) désigne une notification suite à l'ajout d'un nouveau contenu (Hammond *et al.*, 2004).

Par ailleurs, les Ressources — documents consultés comme *Annotations* formulées — font l'objet d'une indexation sur le contenu, classique dans le domaine de la Recherche d'Information,

cf. (Baeza-Yates et Ribeiro-Neto, 1999, ch. 2) et (Manning *et al.*, 2008, ch. 2). Elle permet notamment d'identifier les Termes des documents et leur importance en terme de fréquence d'apparition, pour éventuellement déduire les thématiques des documents. Pour chaque document à indexer, les quatre étapes suivantes sont généralement mises en œuvre :

1. la segmentation permet de découper le contenu d'un document en unités lexicales, il existe des algorithmes de segmentation spécifiques à chaque format de document ;
2. l'élimination des « mots vides » est spécifique à la langue du document, elle permet de rejeter les unités lexicales qui ne permettraient pas de discriminer le document lors de recherches futures : articles, déterminant et autres mots-outils ;
3. la lemmatisation consiste à transformer un mot (éventuellement conjugué ou accordé) en sa forme canonique, à l'aide de l'algorithme de Porter (1980) pour l'anglais ou en le tronquant à sept caractères pour le français (Tuffery, 1984; Denjean, 1989), par exemple ;
4. le dénombrement des termes consiste à compter le nombre d'occurrences de chaque terme distinct pour le document indexé. Le résultat de ce processus est stocké au niveau de la classe Index reliée à l'association entre les classes Ressource et Terme, l'attribut Index.nb représentant le nombre d'occurrences du terme dans le document.

Résultant de la fusion des deux diagrammes de classes présentés dans cette section, le modèle unifié proposé est exploité par les processus intégrés faisant l'objet du chapitre II.5. Avant de détailler ces processus, nous introduisons deux propositions sur lesquelles ils reposent en partie : la validation sociale d'annotations collectives (chapitre II.3) et la mesure de similarité d'usage (chapitre II.4).

3 De l'identification à l'agrégation d'opinions : mesurer la « validation sociale » d'annotations collectives argumentatives

> "Even in a medium that allowed for perfect interactivity for all participants (something we have a reasonable approximation of today), the limits of human cognition will mean that scale alone will kill conversation."
>
> *Clay Shirky (2008, p. 98)*

L'ANNOTATION de ressources électroniques est utile dans de nombreuses situations décrites dans le chapitre I.3. Considérons donc le cas d'un système d'annotation utilisé par un nombre d'utilisateurs croissant qui annotent quotidiennement : au fil du temps, les ressources contiennent de plus en plus d'annotations. Alors que le lecteur peut bénéficier de quelques annotations sans qu'elles ne le dérangent, une dizaine de pictogrammes additionnels rendent la consultation de la ressource inconfortable. Un plus grand nombre d'annotations submerge le lecteur — il suffit de visiter la page Web d'Annotea (figure I.3.6) avec le système d'annotation Amaya (Kahan *et al.*, 2002) pour s'en convaincre.

Ainsi, bénéficier des annotations présentes sur une ressource est d'autant plus difficile qu'elles sont nombreuses. De plus, chaque annotation suscite potentiellement un débat sous la forme d'un fil de discussion. Pour savoir si le groupe est d'accord ou pas avec une annotation donnée, le lecteur doit en consulter chaque réponse, comprendre l'opinion qui y est exprimée puis les synthétiser récursivement. Il évalue alors mentalement la « validité sociale » de l'annotation. Cela lui demande un effort cognitif non négligeable qui le distrait de sa tâche principale : la lecture. Paradoxalement, une telle surcharge cognitive devrait être réduite à tout prix (O'Hara et Sellen, 1997).

Nous proposons dans ce chapitre d'évaluer la « validité sociale » des annotations collectives, en se focalisant sur les annotations argumentatives (section II.2.1.2.3) qui peuvent être débattues au sein d'un fil de discussion. Cet indicateur agrège les opinions exprimées dans le fil de discussion de façon à expliciter l'opinion globale du groupe qui a débattu. La validité sociale $v(a) \in [-1;1]$ d'une annotation a est une valeur continue que l'on peut représenter sur l'axe de la figure II.3.1.

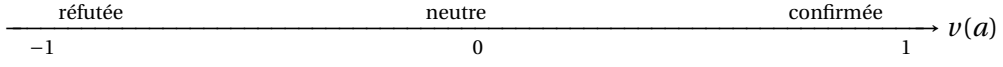


Figure II.3.1 – Valeur continue de la validité sociale $v(a)$ d'une annotation a.

Grâce à notre proposition, le lecteur peut obtenir la validité sociale d'une annotation sans avoir à éplucher chaque réponse dans le fil de discussion et à les synthétiser. Il consacre alors moins de temps et d'efforts à ces tâches que nous automatisons, lui permettant ainsi de se focaliser sur les annotations consensuelles ou, au contraire, controversées. Le chapitre III.2 (p. 99) détaille l'évaluation expérimentale de cette proposition, réalisée avec 121 participants.

3.1 Algorithmes pour mesurer la validation sociale

Cette section décrit trois approches que nous avons proposées pour calculer la validation sociale $v(a) \in [-1;1]$ d'une annotation a. Cette fonction continue combine l'opinion intrinsèque de a et l'opinion globale exprimée dans les réponses issues de son fil de discussion. Ces opinions sont définies dans le tableau II.2.1. Concrètement, $v(a) \to 0$ signifie que a n'a pas de réponse, ou que ses réponses sont équilibrées entre confirmations et réfutations. De plus, $v(a) \to 1$ indique que a est totalement confirmée dans le fil de discussion. Enfin, $v(a) \to -1$ signifie que a est totalement réfutée dans le fil de discussion.

À partir de l'évaluation $v(a)$ d'une annotation, on peut conclure qu'elle fait l'objet d'un consensus positif (resp. négatif) lorsqu'elle est totalement confirmée, soit $v(a) \to 1$ (resp. totalement réfutée, soit $v(a) \to -1$) par son fil de discussion. Plus généralement, une annotation a fait l'objet d'un consensus (positif ou négatif) lorsque $|v(a)| \to 1$.

Afin de définir de quelle façon $v(a)$ évolue, le tableau II.3.1 détaille les quatre combinaisons obtenues à partir de l'opinion de l'annotation et de l'opinion globale observée dans le fil de discussion. Par exemple, le cas 2 représente une annotation qui confirme un texte ($\mathscr{C}$). Cette annotation est globalement réfutée ($\mathscr{R}$) par son fil de discussion, c'est pourquoi sa validité sociale est diminuée $v(a) \to 0$. Les annotations neutres ($\mathscr{N}$) ne sont pas prises en compte car elles n'expriment pas une opinion à proprement parler.

	cas 1	cas 2	cas 3	cas 4
Opinion de l'annotation a	$\mathscr{C}$	$\mathscr{C}$	$\mathscr{R}$	$\mathscr{R}$
Opinion globale des réponses de a	$\mathscr{C}$	$\mathscr{R}$	$\mathscr{C}$	$\mathscr{R}$
Validité sociale $v(a)$	$v(a) \to 1$	$v(a) \to 0$	$v(a) \to -1$	$v(a) \to 0$

Tableau II.3.1 – Validité sociale d'une annotation selon l'opinion de ses réponses.

Les sections suivantes détaillent les trois approches que nous avons explorées pour calculer la validité sociale $v(a)$. La première approche considère le coefficient κ (kappa) de Cohen (1960),

introduit dans le cadre des Sciences Sociales. Montrant l'inadéquation de ce coefficient pour notre contexte, nous présentons ensuite un algorithme récursif d'agrégation de scores (Cabanac *et al.*, 2005, 2006b). La troisième approche (Cabanac *et al.*, 2006b, 2007b) assoit la validation sociale sur un cadre théorique, en étendant le système d'argumentation bipolaire par Cayrol et Lagasquie-Schiex (2005a,b) dans le domaine de l'Intelligence Artificielle.

3.1.1 Approche 1 : mesure du degré d'accord entre annotateurs

Le coefficient kappa de Cohen (1960) $\kappa \in [-1;1]$ mesure le degré d'accord entre $n = 2$ juges (personnes) qui répartissent N objets dans k catégories mutuellement exclusives. Il a été originalement proposé pour mesurer l'accord de n médecins diagnostiquant une maladie parmi k chez N malades. Le coefficient κ de Fleiss (1971) applicable à $n \geqslant 2$ juges est une généralisation du κ de Cohen (1960). La valeur $\kappa = \frac{P(A)-P(E)}{1-P(E)}$ de ces coefficients dépend de l'accord observé $P(A)$ entre les juges, ainsi que de $P(E)$ qui figure la probabilité d'obtenir un accord par hasard (coïncidence). Landis et Koch (1977) et Fleiss *et al.* (2003, p. 604) proposent une grille d'interprétation du coefficient κ utilisable pour la plupart des applications ("for most purposes"), présentée dans le tableau II.3.2.

Valeur du κ de Fleiss (1971)	$[-1;0,40[$	$[0,40;0,75[$	$[0,75;1]$
Degré d'accord dépassant la coïncidence	faible	moyen à bon	excellent

Tableau II.3.2 – Interprétation du coefficient κ selon (Landis et Koch, 1977; Fleiss *et al.*, 2003).

Par rapport à notre objectif relatif au calcul de la validation sociale, le coefficient κ est inadapté. En effet, il ne prend pas en compte le fait qu'une évaluation puisse être contestée par d'autres juges, au sein d'un fil de discussion dans notre cas. Étant donnée la structure arborescente d'un fil de discussion, nous avons développée l'algorithme d'agrégation récursive de scores d'arguments présenté dans la section suivante.

3.1.2 Approche 2 : agrégation récursive de scores d'arguments

L'algorithme de validation sociale proposé dans (Cabanac *et al.*, 2005, 2006b) opère sur une annotation et son fil de discussion. Dans cette section, nous regroupons sous le terme « annotation » aussi bien l'annotation initiale que les réponses qu'elle a suscitées dans le fil de discussion. L'algorithme décrit dans cette section comprend les deux étapes suivantes, notées 1. et 2.

1. Le calcul de l'*agrément* (3.3) de chaque annotation du fil de discussion dépend de deux paramètres, notés (a) et (b).
   (a) La valeur de *confirmation* $c(a) \in [-1;1]$ d'une annotation a est basée sur ses types opinion. Ainsi, une annotation typée $\mathscr{C}$ aura une valeur de confirmation positive, alors que le type de réfutation $\mathscr{R}$ implique une valeur de confirmation négative. La figure II.3.2 présente l'évaluation graduelle associée à chaque combinaison de types.
   (b) Le second paramètre considéré permet d'augmenter la valeur d'*agrément* en fonction de l'implication de son auteur. Par exemple, une personne qui fournit des références (information optionnelles) en plus de son commentaire montre une implication importante visant à justifier son opinion. Ceci lui demande un effort cognitif supplémentaire

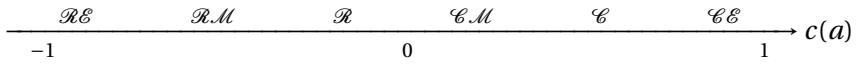


Figure II.3.2 – Valeur de *confirmation* d'une annotation a en fonction de ses types.

par rapport à la seule création d'un commentaire. L'adjonction de références apporte une valeur ajoutée à l'annotation, il suffit de comparer une réfutation (par ex. : « C'est faux ! ») avec une réfutation étayée (par ex. : « Selon X et comme on peut le constater dans Y, cette affirmation est fausse car… »).

Par la suite, A représente l'ensemble des annotations. De plus, nous utilisons la notation pointée en référence au modèle d'annotation (définition 1, p. 48). Ainsi *a.commentaire* signifie « le commentaire associé à l'annotation a ». La fonction (3.1) de signature i_c : $A \to [0;1]$ reflète la présence d'un commentaire associé à l'annotation a.

$$i_c(a) = \begin{cases} 0 & \text{si } \mathit{estVide}(a.\mathit{commentaire}) \\ 1 & \text{sinon} \end{cases} \tag{3.1}$$

De plus, la fonction (3.2) de signature $i_r : A \to [0;1]$ croît en fonction du nombre de références adjointes à l'annotation a.

$$i_r(a) = \frac{|a.\mathit{références}|}{1 + \max_{x \in A} |x.\mathit{références}|} \tag{3.2}$$

Sur cette base, l'*agrément* d'une annotation a est évalué par la fonction (3.3) de signature $a : A \to [0;1]$ qui prend en compte aussi bien son contenu (commentaire et références) que sa valeur de confirmation $c(a)$. Les paramètres $\alpha, \beta \in [0;1]$ permettent l'ajustement des poids relatifs de deux fonctions i_r et i_c.

$$a(a) = \frac{c(a)\,(1 + \alpha \cdot i_c(a))\,(1 + \beta \cdot i_r(a))}{(1+\alpha)\,(1+\beta)} \tag{3.3}$$

2. La seconde étape de l'algorithme consiste à combiner l'*agrément* intrinsèque à une annotation avec l'*agrément* global exprimé dans ses réponses. Cette combinaison appelée *validité sociale* est calculée par la fonction (3.4) de signature $v : A \to [-1;1]$.

$$v(a) = \begin{cases} 0 & \text{si } a.\mathit{père} = \lambda \wedge |a.\mathit{réponses}| = 0 \\ \frac{1}{2} \cdot a(a) \cdot \big(1 + \gamma \cdot s(a)\big) & \text{sinon} \end{cases} \tag{3.4}$$

L'évaluation de (3.4) repose sur les règles suivantes. On obtient $v(a) = 0$ lorsque l'annotation a n'est ni confirmée ni réfutée : c'est la racine ($a.\mathit{père} = \lambda$) d'un fil de discussion vide ($|a.\mathit{réponses}| = 0$). Sinon, $v(a)$ est fonction de l'*agrément* de a ainsi que de la *synthèse* des agréments de ses réponses. Pour ce faire, la fonction (3.5) de signature $s : A \to [-1;1]$ retourne une valeur négative lorsque a est réfutée par ses réponses, ou positive dans le cas contraire. Le paramètre $\gamma \in [0;1]$ permet d'ajuster l'impact du fil de discussion sur la *validité sociale* de a.

$$s(a) = \begin{cases} 1/\gamma & \text{si } |a.\mathit{réponses}| = 0 \\ \dfrac{\sum\limits_{r \in a.\mathit{réponses}} v(r) \cdot r.\mathit{expertise}}{\sum\limits_{r \in a.\mathit{réponses}} r.\mathit{expertise}} \left[1 + \ln\left(1 + \frac{|a.\mathit{réponses}|}{m(1+n(a))}\right) - \ln(2)\right] & \text{sinon} \end{cases} \tag{3.5}$$

La fonction (3.5) réalise la *synthèse* des réponses associées à une annotation donnée. Elle comprend une moyenne de leurs *validités sociales*, pondérée par l'*expertise* de leurs auteurs afin de renforcer l'impact des réponses formulées par des experts (cet attribut étant strictement positif). La valeur de *synthèse* est également renforcée en fonction du nombre de réponses : plus une annotation possède de réponses, plus sa *validité sociale* est renforcée. Dans (3.5), la fonction de signature $n : A \rightarrow \mathbb{N}_+$ retourne le niveau d'une annotation donnée dans le fil de discussion, la racine ayant un niveau égal à zéro. Enfin, la fonction $m : \mathbb{N}_+ \rightarrow \mathbb{N}_+$ retourne le nombre maximum de réponses $m(x)$ existant au niveau x du fil de discussion. En résumé, une annotation a n'ayant pas eu de réponse a pour *validité sociale* la valeur de son *agrément* : $|a.réponses| = 0 \Rightarrow v(a) = a(a)$. Dans le cas contraire, lorsque a possède des réponses, sa valeur de *synthèse* repose sur la moyenne pondérée de leurs *validités sociales* (A) multipliée par une expression (B) qui croît selon son nombre de réponses. L'expression B prend en compte le nombre maximum de réponses existant au même niveau que le niveau de a. Par exemple, si a (niveau n) possède m réponses et si le nombre maximum de réponses au même n^{e} niveau est égal à N, alors $B = 1+\ln\left(1+\frac{m}{N}\right)-\ln(2)$. Notons que la valeur maximale de B est atteinte lorsque $m = N \Rightarrow B = 1$. Par conséquent $B \in \left[\ln\left(\frac{e\cdot(N+1)}{2N}\right), 1\right]$, le logarithme népérien permettant de réduire les différences entre de faibles et de grandes valeurs de N. Comme $B \leqslant 1$, la valeur de A n'est pas augmentée par B.

La valeur $v(a)$ de l'annotation a évalue sa *validité sociale* en fonction des valeurs d'*agrément* et de *synthèse*. Concrètement, $|v(a)| \rightarrow 1$ reflète un consensus global au sein du fil de discussion de l'annotation a. En fonction du signe de $v(a)$, on peut conclure que le fil de discussion valide une annotation de type $\mathscr{C}$ confirmation (resp. $\mathscr{R}$ réfutation) lorsque $v(a) \rightarrow 1$ (resp. $v(a) \rightarrow -1$).

L'algorithme de validation sociale présenté dans cette section repose en partie sur des heuristiques et des paramètres initialisés empiriquement. Dans le but d'améliorer cette première approche empirique, la section suivante présente une alternative pour calculer la validation sociale. Elle repose sur une approche formelle du domaine de l'Intelligence Artificielle (Cayrol et Lagasquie-Schiex, 2005a,b) : la théorie de l'argumentation qui a été employée dans de nombreux contextes, tel que la prise de décision médicale.

3.1.3 Approche 3 : extension d'un système d'argumentation bipolaire

Dung (1995) modélise une argumentation par le couple $\langle A, R\rangle$ où A est un ensemble d'arguments et R est une relation binaire sur A^2 appelée « relation d'attaque ». Le système d'argumentation ainsi constitué peut être représenté par un graphe direct dont les nœuds sont les arguments et les arcs relient les arguments sources aux arguments cibles. L'identification des branches d'attaque et de défense dans ce système permet de décider de l'acceptabilité d'un argument. C'est une valeur binaire (acceptable ou non acceptable) dépendant de l'identification de groupes d'arguments sans conflit ou exprimant une défense collective.

Dans ce contexte, Cayrol et Lagasquie-Schiex (2005a) remarquent que les travaux sur l'argumentation les plus récents ne considèrent qu'un seul type d'interaction entre deux arguments : l'attaque. Pourtant, de nombreux travaux dont (Karacapilidis et Papadias, 2001) suggèrent qu'un autre type d'interaction doit être considéré afin de représenter complètement la connaissance, dans de nombreux contextes concrets. Ce second type d'interaction est l'*appui*. Par conséquent,

Cayrol et Lagasquie-Schiex (2005a) considèrent les arguments de type *attaque* et de type *appui* pour étendre le système d'argumentation de Dung (1995) en définissant un système d'argumentation bipolaire (définition 3).

Définition 3. Un Système d'Argumentation BiPolaire (Cayrol et Lagasquie-Schiex, 2005a) noté SABP est représenté par le triplet $\langle A, R_{app}, R_{att}\rangle$ où :

- A est un ensemble d'arguments, par exemple : $A = \{a_1, \ldots, a_n\}$,
- R_{app} est une relation d'*appui* sur A^2. Le couple $(a_i, a_j) \in R_{app}$ est représenté par $a_i \rightarrow a_j$.
- R_{att} est une relation d'*attaque* sur A^2. Le couple $(a_i, a_j) \in R_{att}$ est représenté par $a_i \not\rightarrow a_j$.

À partir de la définition 3, l'évaluation graduelle du SABP est définie de façon à respecter les trois principes suivants :

- **P1** l'évaluation d'un argument est fonction de l'évaluation de tous ses attaquants directs et de tous ses appuis directs ;
- **P2** si la qualité de l'appui (resp. de l'attaque) augmente alors la valeur de l'argument ainsi appuyé (resp. attaqué) augmente ;
- **P3** si on ajoute des appuis (resp. des attaques) alors la qualité de l'appui (resp. de l'attaque) augmente.

Par la suite, $R^-_{app}(a)$ désigne les appuis directs de l'argument a, de façon similaire $R^-_{att}(a)$ désigne ses attaques directes. De plus, V est un ensemble totalement ordonné admettant un plus petit élément v_{min} et un plus grand élément v_{max}. Enfin, V^* désigne l'ensemble des suites finies d'éléments de V. Considérant ces trois principes, les auteurs posent $a \in A$ avec $R^-_{app}(a) = \{b_1, \ldots, b_p\}$ et $R^-_{att}(a) = \{c_1, \ldots, c_q\}$ et définissent une évaluation graduelle comme l'application (3.6) de signature $v : A \rightarrow V$ telle que :

$$v(a) = g\left(h_{app}\left(v(b_1), \ldots, v(b_p)\right), h_{att}\left(v(c_1), \ldots, v(c_q)\right)\right) \tag{3.6}$$

Dans (3.6), la fonction $h_{app} : V^* \rightarrow H_{app}$ évalue la qualité de l'appui sur un argument. Sur le même principe, la fonction $h_{att} : V^* \rightarrow H_{att}$ évalue la qualité de l'attaque sur un argument. Par ailleurs, $g : H_{app} \times H_{att} \rightarrow V$ est définie telle que $g(x, y)$ est croissante en x et décroissante en y. Enfin, la fonction h ($h = h_{app}$ ou $h = h_{att}$) doit satisfaire les trois conditions suivantes :

- **C1** si $x_i \geqslant x'_i$ alors $h(x_1, \ldots, x_i, \ldots, x_n) \geqslant h(x_1, \ldots, x'_i, \ldots, x_n)$;
- **C2** $h(x_1, \ldots, x_i, \ldots, x_n, x_{n+1}) \geqslant h(x_1, \ldots, x_i, \ldots, x_n)$;
- **C3** $h() = \alpha \leqslant h(x_1, \ldots, x_i, \ldots, x_n) \leqslant \beta$ pour tous les $x_1, \ldots, x_i, \ldots, x_n$.

Cayrol et Lagasquie-Schiex (2005a) proposent deux instances de cette évaluation générique. La première agrège les valeurs des arguments en conservant le maximum des attaques et des appuis directs, c'est-à-dire $h_{att} = h_{app} = \max$. Cette première approche n'est pas acceptable dans notre contexte d'application car elle ne prend pas en compte l'ensemble de tous les arguments exprimés dans le fil de discussion.

Une seconde instance est proposée avec :

- $V = [-1; 1]$,
- $H_{app} = H_{att} = [0; \infty]$,
- $h_{app} = h_{att} = \sum_{i=1}^{n} \frac{x_i+1}{2}$,
- $g(x, y) = \frac{1}{1+y} - \frac{1}{1+x}$.

Afin de calculer la validité sociale d'une annotation a, nous la modélisons par un SABP créé à partir du fil de discussion de a (exemple 1). Ainsi, l'ensemble A contient les nœuds du fil de discussion, les couples des relations R_{app} et R_{att} étant respectivement définis à partir des réponses de type $\mathscr{C}$ et de type $\mathscr{R}$. De ce fait, l'application (3.6) associée à la seconde instance d'évaluation permet de calculer la validité sociale de l'annotation.

Exemple 1. *La figure II.3.3 représente une discussion au sujet d'un passage de document mathématique contenant l'expression « $\sqrt{x^2} = x$ ». Une personne a créé l'annotation a, qui a obtenu les réponses r_i constituant le fil de discussion associé. Le type des arguments est représenté par la notation fléchée (définition 3), leur contenu est reproduit dans le tableau II.3.3. Nous modélisons cette discussion par le SABP $\langle A, R_{app}, R_{att}\rangle$ où $A = \{a, r_1, r_2, r_3, r_{21}, r_{22}, r_{211}\}$ contient l'annotation a et ses réponses r_i. Les relations entre l'annotation et ses réponses sont exprimées par $R_{app} = \{(r_1, a), (r_3, a), (r_{21}, r_2), (r_{211}, r_{21})\}$ pour les réponses de type $\mathscr{C}$, ainsi que par $R_{att} = \{(r_2, a), (r_{22}, r_2)\}$ pour les réponses de type $\mathscr{R}$. La validité sociale de l'annotation a est $v(a) = 0,152$.*

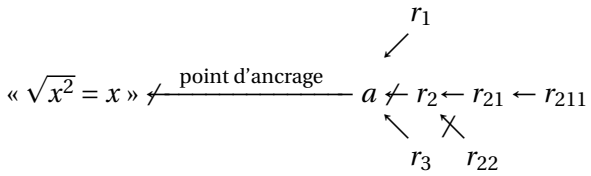


Figure II.3.3 – Discussion au sujet de l'expression « $\sqrt{x^2} = x$ ».

Nom	Type	Commentaire associé à l'argument
a	$\mathscr{RME}$	Cette formule est fausse, regardez ce contre-exemple : $\sqrt{(-2)^2} \neq -2$. Considérez donc la modification suivante : $\sqrt{x^2} = \lvert x\rvert$.
r_1	$\mathscr{CE}$	Remarque judicieuse car $\sqrt{(-4)^2} = \lvert -4\rvert = 4$.
r_2	$\mathscr{R}$	Cette leçon de collège concerne *uniquement* les nombres positifs…
r_3	$\mathscr{C}$	Ce n'est qu'un cas particulier de la formule $\forall (x, n) \in \mathbb{R} \times \mathbb{R}^* \quad \sqrt[n]{x^n} = \lvert x\rvert$.
r_{21}	$\mathscr{CM}$	Alors il faudrait préciser la restriction sur $\mathbb{R}_+$, par exemple : $\forall x \in \mathbb{R}_+ \quad \sqrt{x^2} = x$.
r_{22}	$\mathscr{RE}$	Particulièrement déroutant lorsqu'on ne se rend pas compte du niveau ciblé !
r_{211}	$\mathscr{CM}$	$\mathbb{R}$ est inconnu au collège, utilisez plutôt « nombre positif ».

Tableau II.3.3 – Arguments associés à la discussion représentée en figure II.3.3.

L'évaluation graduelle $v(a)$ de cet exemple ne prend pas en compte certaines informations disponibles relatives aux arguments du fil de discussion. En effet, des informations subjectives (IS, cf. définition 1) associées aux nœuds du fil de discussion ne sont pas considérées, telles que les types de la classe « commentaire », l'expertise, le contenu et les références de l'annotation.

Définition 4. Afin d'exploiter les informations subjectives portées par les annotations, nous avons proposé dans (Cabanac *et al.*, 2006b, 2007b) d'étendre le SABP de Cayrol et Lagasquie-Schiex (2005a) en redéfinissant dans (3.7) l'application d'évaluation v, que nous notons $v' : A \to V$.

$$v'(a) = g\left(h_{app}\left(i(b_1) \cdot v'(b_1), \ldots, i(b_p) \cdot v'(b_p)\right), h_{att}\left(i(c_1) \cdot v'(c_1), \ldots, i(c_q) \cdot v'(c_q)\right)\right) \tag{3.7}$$

Nous introduisons la fonction (3.8) de signature $i : A \to I$ en paramètre de $h_{app} : V^* \to H_{app}$ et de $h_{att} : V^* \to H_{att}$ pour mesurer la valeur intrinsèque d'un argument en prenant en compte n critères. Ayant au préalable retenu la seconde instance d'évaluation, nous devons respecter $V = [-1;1]$. Par conséquent, nous posons $I = [0;1]$ de façon à respecter $\forall a \in A \quad i(a) \cdot v'(a) \in V$. Par ailleurs, les n coefficients $\pi_i \in [0;1]$ définis tels que $\sum_{i=1}^{n} \pi_i = 1$ permettent d'ajuster l'importance relative des n critères évalués par les fonctions $f_i : A \to F_i \subseteq \mathbb{R}_+$. La fonction « sup » retourne la borne supérieure du domaine de définition des f_i. Le coefficient $\delta \in [0;1]$ permet de moduler globalement l'impact des n critères sur l'évaluation v', notons que $\delta = 0 \Longrightarrow v'(A) = v(A)$.

$$i(x) = \delta \cdot \sum_{i=1}^{n} \frac{\pi_i \cdot f_i(x)}{\sup(F_i)} \tag{3.8}$$

Pour notre cadre applicatif, nous avons identifié $n = 4$ critères à prendre en compte pour évaluer un argument : l'*expertise* et l'*agrément* de ses attaques et appuis ainsi que l'implication de l'annotateur en termes de *commentaires* et *références*. Ainsi, la fonction f_1 croît avec l'*expertise* associée à l'argument évalué, nous proposons une échelle d'expertises : béotien $\prec$ débutant $\prec$ intermédiaire $\prec$ confirmé $\prec$ expert. La fonction f_2 associe à chaque combinaison de *types* de la classe « commentaire » une valeur réelle qui représente l'agrément de l'annotateur, par ex. un individu qui donne un exemple ($\mathscr{E}$) est davantage d'accord qu'un individu proposant une modification ($\mathscr{M}$).

$$\xrightarrow[F_2]{\mathscr{M}2\mathscr{E} \quad \mathscr{M}2 \quad \mathscr{M}\mathscr{E} \quad 2\mathscr{E} \quad \mathscr{M} \quad 2 \quad \mathscr{E}} f_2(x)$$

La fonction f_3 évalue l'implication de l'annotateur en considérant la présence d'un *commentaire* pour une annotation. Enfin, la fonction f_4 croît selon le nombre de *références* citées dans une annotation : $f_4(x)$ calcule le ratio entre le nombre de références de x et le nombre maximum de références par annotation. Dans le cadre de l'implantation de notre proposition, nous avons fixé $\delta = 1$ afin d'obtenir un impact maximal des n critères sur l'évaluation des arguments. Tout en prévoyant d'expérimenter d'autres pondérations, nous avons fixé intuitivement $\pi_2 = \pi_3 = \frac{1}{3}$ et $\pi_1 = \pi_4 = \frac{1}{6}$ pour prendre principalement en compte le *commentaire* et l'*agrément* des appuis et attaques. Ainsi, à partir de (3.7) on peut déduire qu'une annotation fait l'objet d'un consensus social, c'est-à-dire elle est *validée socialement* lorsque ses réactions expriment globalement une même opinion (réfutation ou confirmation) :

$$v'(A) \to 1 \vee v'(A) \to -1 \Longleftrightarrow \left|v'(A)\right| \to 1 \tag{3.9}$$

3.2 Limites de la validation sociale

La validation sociale d'une annotation reflète l'opinion globale du groupe qui a débattu dans le fil de discussion associé. Par contre, elle n'évalue en aucun cas à quel point une annotation est *vraie*. En effet, la validation sociale, comme tout système démocratique, est biaisée par la « tyrannie de la majorité » (de Tocqueville, 1835). L'Histoire regorge d'exemples montrant que le nombre (de personnes) ne suffit pas à établir une vérité. Par exemple, considérons une annotation (fictive) de Galilée : « Ce n'est pas la Terre qui est au centre de l'Univers, mais le Soleil » et les arguments de

ses détracteurs contemporains. L'évaluation de cette annotation aboutirait à un consensus négatif car la société était catégoriquement opposée aux idées de Galilée. Bien que fidèle à l'opinion du groupe, la validité *sociale* (relative au groupe qui s'est exprimé) d'une annotation ne permet pas de conclure sur sa validité *universelle*.

Une seconde limite de la validation sociale concerne sa sensibilité aux réponses hors-sujet (*discussion drift*). Par exemple, une discussion à propos de l'équation « $E = mc^2$ » peut contenir des interventions hors-sujet concernant la vie privée d'Albert Einstein, alors que ce n'est pas le sujet original de la discussion. Par conséquent, la validité sociale de l'annotation initiale est biaisée car elle prend en compte les arguments hors-sujet. Or, Radev (1999) détecte le changement de thématique entre différents articles de presse en évaluant la distance entre leurs contenus préalablement indexés. C'est une approche exploitable pour identifier les sous-arbres du fil de discussion qui sont hors-sujet, dans le but de les ignorer lors du calcul de la validation sociale.

La dernière limite que nous avons identifiée concerne le caractère impersonnel de la validation sociale. Quel que soit le lecteur, la validité sociale d'une annotation est inchangée. Pourtant, chacun possède ses critères d'évaluation de la validité d'un écrit. Par exemple, Marshall (1998) a observé que l'origine d'une annotation est un critère important pour le lecteur. De ce fait, le lecteur pourrait personnaliser le calcul de la validation sociale en lui permettant d'affecter des valeurs de « confiance » aux individus présents dans son réseau social.

4 Définition d'une mesure de similarité basée sur l'usage des documents

"... user's grouping behavior (such as the placement of subjects in folders) as an indication of semantic coherency or relevant groupings between subjects."

James Rucker and Markos J. Polanco (1997)

COMPARER des documents est une tâche fondamentale car récurrente en Recherche d'Information. Par exemple, certains systèmes de recommandation d'information (Resnick et Varian, 1997) comparent des documents candidats avec ceux de l'usager, afin de ne lui recommander que les plus similaires. Dans ce contexte, la comparaison est communément effectuée sur la base du contenu des documents. Ainsi, l'usager se voit recommander des documents qui contiennent majoritairement les mêmes termes que les documents qu'il possède.

En fait, calculer la similarité sur le contenu permet d'identifier les documents semblables à un document donné. En revanche, cette approche ne permet en aucun cas de trouver les documents *utilisés conjointement* avec un document donné : deux documents utilisés ensemble n'ont pas forcément le même contenu, et vice versa. Pourtant, un tel indicateur pourrait améliorer l'efficacité des individus en leur suggérant des groupes de documents qui ont fait sens par le passé, dans le cadre de la réalisation des projets de l'organisation.

Ce chapitre poursuit les travaux initiés par Chevalier (2002) en introduisant la notion d'usage de document (Cabanac *et al.*, 2007a). Nous définissons ensuite une mesure de similarité d'usage entre les documents. Elle se veut complémentaire aux mesures de similarité sur le contenu. Par extension, nous définissons également une mesure d'usage entre individus. Enfin, nous discutons les limites de cette proposition qui fait l'objet d'une validation expérimentale dans le chapitre III.3 (p. 117).

4.1 Définition de la notion d'usage d'un document

Les membres organisationnels ont recours à des documents pour mener à bien leurs activités, notamment leurs projets. L'observation des documents employés pour chaque activité peut révéler certains regroupements récurrents. En effet, des groupes de documents fréquemment utilisés ensemble peuvent se détacher. Chaque groupe forme alors un *usage* déterminé. Cette relation d'usage est d'autant plus forte que de nombreux individus associent les mêmes documents ensemble. Par exemple, des biologistes réalisant des essais cliniques peuvent regrouper des documents concernant des médicaments pour lesquels ils ont observé des interactions. Par la suite, un autre biologiste travaillant dans une équipe différente peut bénéficier de cette information si elle lui est présentée.

Afin d'expliciter les usages des documents organisationnels, nous exploitons les EPA qui sont structurés en projets et découpés en sous projets (section I.1.5). Dans le but d'identifier les regroupements récurrents de documents, nous tirons parti de la structure de données « multi-arbres » introduite par Furnas et Zacks (1994).

4.2 Modélisation des EPA organisationnels dans un multi-arbres

La structure du multi-arbres factorise au sein d'une structure de données unique les documents éparpillés dans tous les EPA de l'organisation. Les hiérarchies des EPA y sont également représentées. La figure II.4.1 propose un exemple de multi-arbres construit à partir des EPA appartenant à deux usagers (u_1 et u_2) d'une organisation minimale. Ce multi-arbres contient deux racines, une par usager.

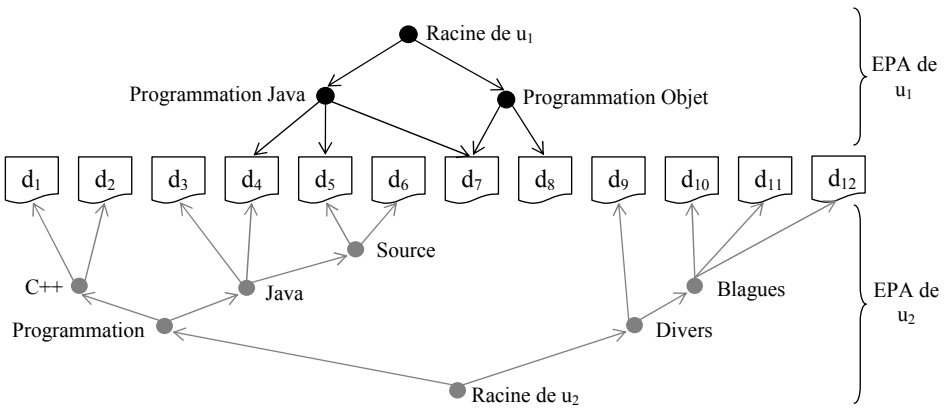


Figure II.4.1 – Exemple d'un multi-arbres construit à partir des EPA de deux individus.

Définition 5. Un multi-arbres $\mathcal{M} = \langle D, R, U, R_D, R_R, R_U \rangle$ est représenté par un sextuplet pour lequel $D = \{d_1, \ldots, d_n\}$ est un ensemble de documents, $R = \{r_1, \ldots, r_m\}$ est un ensemble de répertoires et $U = \{u_1, \ldots, u_l\}$ est un ensemble d'usagers. Par ailleurs, la structure du multi-arbres est spécifiée à partir des relations suivantes :

– R_D est une relation binaire sur $D \times R$ traduisant la présence de documents dans les répertoires. Ainsi, $(d_i, r_j) \in R_D$ signifie que le document d_i est contenu dans le répertoire r_j.

- R_R est une relation binaire sur $R \times R$ traduisant l'inclusion de répertoires, c'est-à-dire la relation père-fils entre répertoires. Ainsi, $(r_i, r_j) \in R_R$ signifie que le répertoire r_i est un fils direct du répertoire r_j.
- R_U est une relation binaire sur $U \times R$ traduisant l'appartenance d'un EPA à un membre organisationnel, l'EPA étant accessible à partir de son répertoire racine. Ainsi, $(u_i, r_j) \in R_U$ signifie que l'usager u_i possède comme racine de sa hiérarchie le répertoire r_j.

De plus, $R_R^+ : R \rightarrow R$ est une fonction (4.1) qui retourne le parent direct p d'un répertoire r donné. Dans le cas où r est une des racines du multi-arbres $R_R^+(r) = \lambda$, où λ représente la valeur nulle.

$$R_R^+(r) = p \mid \exists (r, p) \in R_R \tag{4.1}$$

Définition 6. Soit $\mathcal{G}$ le graphe associé au multi-arbres $\mathcal{M}$. Un sommet de $\mathcal{G}$ est soit un nœud d'un EPA (un répertoire), soit une feuille d'un EPA (un document). Un arc de $\mathcal{G}$ provient de $R_D \cup R_R$. Un *chemin* à partir d'une racine r jusqu'à un document d forme une séquence notée « $/r/r_1/r_2/\ldots/r_k/d$ » où $r_1 R_R r, r_2 R_R r_1, \ldots, d R_D r_k$. Nous appelons *branche* le répertoire fils direct $r_1 \in R$ de la racine r. La fonction (4.2) de signature $b : R \rightarrow R$ fournit la branche correspondant à un répertoire r donné.

$$b(r) = \begin{cases} \lambda & \text{si } R_R^+(r) = \lambda \\ r & \text{si } b(R_R^+(r)) = \lambda \\ b(R_R^+(r)) & \text{sinon} \end{cases} \tag{4.2}$$

La représentation de la structure et du contenu des EPA sous la forme d'un multi-arbres permet de calculer la similarité d'usage entre deux documents donnés, comme détaillé dans la section suivante.

4.3 Calculs de similarités basés sur l'usage

La similarité d'usage inter-documents (définition 8) reflète le caractère récurrent de certains regroupements de documents au sein des EPA. Son calcul est basé sur la similarité inter-répertoires (définition 7) présentée dans la section suivante.

4.3.1 Similarité d'usage entre répertoires

Définition 7. Sur la base des travaux de Jaczynski et Trousse (1998) relatifs à la similarité entre URL, la fonction (4.3) de signature $\sigma_R : R^3 \rightarrow [0;1]$ évalue la similarité entre deux répertoires au travers du multi-arbres. Concrètement, son calcul repose principalement sur les deux facteurs suivants : leur profondeur et le nombre d'ancêtres que les deux répertoires donnés ont en commun.

$$\sigma_R(b, r_1, r_2) = 1 - \frac{s(r_1, m(r_1, r_2)) + s(r_2, m(r_1, r_2))}{s(r_1, b) + s(r_2, b) + 2} \tag{4.3}$$

Par la suite, l'opérateur ensembliste « $\ominus$ » figure la différence symétrique, correspondant à l'opérateur ou-exclusif (xor) de la logique Booléenne, tel que $A \ominus B = (A \cup B) \setminus (A \cap B) = (A \setminus B) \cup (B \setminus A)$.

De plus, $|A|$ représente le cardinal de l'ensemble A. La fonction (4.4) de signature $s : R^2 \to \mathbb{N}_+$ calcule le nombre de « sauts » (c.-à-d. le nombre d'arcs) dans le chemin menant du répertoire r_1 au répertoire r_2, sachant que ces deux répertoires appartiennent à la même branche.

$$\forall r_1, r_2 \in R \mid b(r_1) = b(r_2) \quad s(r_1, r_2) = |a(r_1) \ominus a(r_2)| \tag{4.4}$$

Pour ce faire, la fonction (4.5) de signature $a : R \to R$ restitue l'ensemble des répertoires ancêtres d'un répertoire r donné, celui-ci y compris.

$$a(r) = \begin{cases} \varnothing & \text{si } r = \lambda \\ \{r\} \cup a\left(R_R^+(r)\right) & \text{sinon} \end{cases} \tag{4.5}$$

Enfin, la fonction (4.3) dépend de la fonction (4.6) de signature $m : R^2 \to R$. Elle retourne le plus proche ancêtre commun aux deux répertoires r_1 et r_2. C'est le répertoire qui a la profondeur maximale parmi les répertoires ancêtres que r_1 et r_2 ont en commun. Dans l'expression de (4.6) la fonction $d : R \to \mathbb{N}_+$ retourne la profondeur d'un répertoire.

$$m(r_1, r_2) = f \mid \forall r, r' \in a(r_1) \cap a(r_2) \quad \left(r \neq r'\right) \wedge \left(d(r) > d(r')\right) \tag{4.6}$$

4.3.2 Similarité d'usage entre documents

La similarité d'usage inter-documents est fonction de deux facteurs : leur proximité moyenne au travers du multi-arbres, ainsi que la récurrence du regroupement observée dans plusieurs EPA. Par exemple, si plusieurs personnes regroupent deux documents donnés dans un même répertoire ou dans deux répertoires proches — au sens de la similarité inter-répertoires — cela signifie qu'il existe un lien entre ces documents. Ce lien résulte potentiellement de multiples causes à la discrétion des individus : proximité thématique, sémantique, temporelle, intentionnelle, etc. Notre proposition vise à expliciter de tels liens implicites afin de caractériser l'usage des documents.

Définition 8. La fonction symétrique (4.7) de signature $\sigma_D : D^2 \to [0; e]$ calcule la similarité d'usage inter-documents. Deux cas particuliers pour lesquels $\sigma_D(d_1, d_2) = 0$ sont à considérer, lorsqu'au moins l'un des deux documents est stocké :

1. à la racine d'un EPA, car il n'a pas fait l'objet d'un classement,
2. dans un répertoire « divers » ou « fourre-tout », pour la même raison.

Par contre, lorsque les deux documents sont dans une même branche b, ils ont fait l'objet d'un effort cognitif de classement. Par conséquent, il existe un lien d'usage entre eux fourni par leur propriétaire. La fonction symétrique σ_D a recours à la fonction (4.8) de signature $R_D^+ : D \times R \to R$ qui retourne le répertoire r contenant un document d donné (père direct), à condition que r soit dans la branche b.

$$\sigma_D(d_1, d_2) = \frac{e^{\frac{u}{|U|}}}{|B|} \sum_{b \in B} \sigma_R\left(b, R_D^+(d_1, b), R_D^+(d_2, b)\right) \tag{4.7}$$

$$R_D^+(d,b) = r \mid (\exists (d,r) \in R_D) \wedge (b \in a(r)) \tag{4.8}$$

Dans la fonction (4.7) l'ensemble $B = b(d_1) \cap b(d_2)$ désigne les branches contenant à la fois d_1 et d_2. De plus, u est le nombre d'usagers dont une branche contient à la fois d_1 et d_2. Enfin, l'expression $e^{\frac{u}{|U|}}$ traduit le fait que plus on observe un regroupement dans une branche donnée chez des usagers distincts, plus les documents regroupés sont similaires par l'usage. La seconde partie de (4.7) calcule la distance moyenne des répertoires contenant les deux documents considérés.

Le chapitre III.3 (p. 117) présente l'expérimentation de la mesure d'usage entre documents exposée dans cette section. Elle suggère la pertinence et la complémentarité de cette mesure eu égard à la mesure classique de similarité entre documents calculée sur leur contenu.

4.3.3 Similarité d'usage entre usagers

Définition 9. Soit la fonction $d : U \rightarrow R^*$ définie à partir des relations binaires R_D, R_R et R_U du multi-arbres (définition 5) qui retourne l'ensemble des documents d'un usager. Pour calculer la similarité d'usage entre deux usagers u_1 et u_2, nous considérons les deux ensembles suivants :

- $D^\cap = d(u_1) \cap d(u_2) = \left\{d_1^\cap, \ldots, d_k^\cap\right\}$ contient les documents $d_i^\cap$ que u_1 et u_2 ont en commun ;
- $D^\ominus = d(u_1) \ominus d(u_2) = \left\{d_1^\ominus, \ldots, d_l^\ominus\right\}$ contient les documents $d_i^\ominus$ possédés par u_1 ou (exclusif) par u_2, mais pas par u_1 et u_2 simultanément (ces derniers formant l'ensemble $D^\cap$).

À partir de la similarité d'usage inter-documents σ_D, nous définissons la similarité d'usage inter-usagers par la fonction symétrique (4.9) de signature $\sigma_U : U^2 \rightarrow \mathbb{R}_+$. Notons que l'initialisation des sommes (indices i et j) prend en compte le caractère symétrique de la fonction σ_D, en évitant de calculer à la fois $\sigma_D(x,y)$ et $\sigma_D(y,x)$ qui sont identiques.

$$\sigma_U(u_1,u_2) = f\left(\sum_{i=1}^{k}\sum_{j=i+1}^{k}\sigma_D\left(d_i^\cap, d_j^\cap\right), \sum_{i=1}^{l}\sum_{j=i+1}^{l}\sigma_D\left(d_i^\ominus, d_j^\ominus\right)\right) \tag{4.9}$$

Dans (4.9), la fonction $f(x,y)$ de signature $f : \mathbb{R}_+^2 \rightarrow \mathbb{R}_+$ est croissante en x et en y. Cette caractéristique vise à accroître la valeur de σ_U d'autant plus que les liens d'usage entre les documents possédés par u_1 et u_2 sont forts. Autrement dit, cette fonction traduit le fait que deux personnes sont d'autant plus proches par l'usage qu'elles possèdent les mêmes documents et qu'elles les organisent de façon similaire. Une instanciation possible est $f(x,y) = (y+1)\cdot e^{(x+1)}$ pour favoriser les liens d'usage des documents que les deux individus considérés ont en commun (x) par rapport à ceux qui ne sont possédés que par l'un d'entre eux (y).

4.4 Apports et discussion de la similarité d'usage

Le concept de similarité d'usage présente les trois principaux avantages suivants.

1. La similarité d'usage tire parti de la structuration des documents dans les EPA, résultant des efforts cognitifs fournis par les membres organisationnels. À notre connaissance, une telle approche n'a pas été explorée dans la littérature.

2. Son calcul ne nécessite pas de connaître le contenu des documents mis en jeu, ni de les indexer *a fortiori*. C'est un avantage clé étant donné que les contenus issus du Web sont très éphémères : de nombreux signets sont cassés lorsque les sites référencés évoluent, par exemple. Malgré l'absence de contenu, on peut calculer la similarité d'usage inter-usagers alors que cela devient impossible avec des approches par contenu. Dans le même contexte, l'approche basée sur l'usage est plus dynamique que celle basée sur le contenu car elle prend en compte l'aspect évolutif des hiérarchies, alors que le contenu des documents ne change pas.
3. Son calcul est indépendant de la langue des documents. En effet, même si des documents sont rédigés dans des langues différentes, ils sont similaires par l'usage dès lors qu'ils appartiennent à une même branche d'un EPA. Dans la même situation, ces documents seraient certainement très éloignés par leur contenu : deux documents, l'un en arabe et l'autre en chinois, seraient peu semblables sur le contenu car ils n'ont que très peu de termes en communs, par exemple. Ainsi, la similarité d'usage peut être exploitée pour compléter des résultats de recherche : une requête en anglais retournerait alors les documents originaux en anglais en fonction d'une similarité de contenu, ainsi que les documents en français (ou toute langue comprise par l'usager) qui leur sont liés par l'usage.

Concernant la mesure de similarité d'usage proposée, nous avons identifié les deux points de discussion suivants.

1. L'usage traduit la récurrence d'un regroupement de documents. Par contre, cette notion n'évalue pas leur *utilisation* en termes d'utilité effective pour les activités de l'organisation. Par exemple, des fichiers peuvent être conservés dans un EPA alors que leur propriétaire n'en a pas (ou plus) l'utilité. Toutefois, un tel regroupement devenu inutile doit être marginal au sein de l'organisation. De ce fait, la pondération relative au nombre d'usagers chez qui l'on observe le regroupement — fonction (4.7) — sera faible. Par ailleurs, l'utilité réelle d'un document pourrait être évaluée selon de sa fréquence d'utilisation (ouvertures, impressions et envois du fichier, par exemple). Associée à un intervalle temporel, cette évaluation permettrait d'identifier la perte et le regain d'intérêt pour un document donné.
2. La mesure de similarité sur l'usage est fortement dépendante des critères de regroupement retenus par les usagers. Bien que différentes études identifient que le classement par projet est privilégié (Jones *et al.*, 2005; Jones, 2007; Khoo *et al.*, 2007), toute autre stratégie est envisageable : thématiquement, chronologiquement, par auteur… Toutefois, ce sont les stratégies les plus adoptées qui seront restituées par la mesure d'usage définie dans ce chapitre. À l'opposé, l'identification de stratégies de regroupement non triviales ou marginales peut être pertinente, notamment pour identifier l'émergence de nouvelles tendances. Cette tâche peut être mise en œuvre en adaptant les formules de calcul présentées dans ce chapitre, de façon à davantage pondérer les regroupements rares que les regroupements fréquents.

En se basant sur le modèle unifié des activités documentaires (chapitre II.2), sur la validation sociale (chapitre II.3) et sur la mesure de similarité d'usage (présent chapitre), nous détaillons dans le chapitre suivant les processus destinés à améliorer les six activités documentaires.

5 Amélioration des six activités documentaires : détail des processus intégrés

> "Knowledge management heavily depends on the willingness of knowledge workers to take part in it. We encountered various reasons for knowledge workers to actually engage in knowledge management initiatives, such as increase of job efficiency, status, and fun. If the condition of a win-win situation is not established, managers will be confronted with major rejections from the side of the knowledge workers."
>
> *Marleen Huysman et Dirk de Wit (2003, p. 45)*

À LA BASE DE NOTRE CONTRIBUTION figure le modèle unifié pour les activités documentaires introduit dans le chapitre II.2. En modélisant les usagers, les ressources et les annotations organisables dans les EPA, nous fédérons les activités de rédaction ② et finalisation ③, d'exploitation ⑤ mais aussi de classement ⑥ de documents électroniques. En particulier, cette modélisation prend en compte le rôle dual de la pratique d'annotation. Pour un usage *individuel*, c'est un outil mis en œuvre lors de la réflexion critique, en permettant la prise de notes en contexte (AnnotationRemarque) contrairement au recours à des services tiers (forum, par exemple) où les individus doivent préciser le contexte de leurs interventions. À des fins d'accès ultérieur, ces notes peuvent être conservées et organisées au sein de l'EPA de l'usager (AnnotationStockage). D'autre part, pour un usage *collectif*, les annotations peuvent être partagées à des fins de travail collaboratif : les lecteurs prennent connaissance des commentaires et opinions de leurs prédécesseurs, ils peuvent également participer à des débats en contexte, ancrés dans le document annoté (AnnotationArgumentative).

Ensuite, nous avons exposé les concepts de validation sociale (chapitre II.3) et de mesure de similarité d'usage (chapitre II.4). Nous tirons profit de ces deux propositions dans le présent chapitre qui expose les six processus intégrés au modèle unifié proposé (figure II.5.1). Ces processus

visent à enrichir la relation usagers-documents en exploitant notamment les EPA qui sont considérés en tant que capital documentaire de l'organisation. Les processus privilégient une approche donnant-donnant en exploitant ce capital, tout en favorisant l'enrichissement d'une activité par les autres activités. Par exemple, le processus NAVI recommande des documents extraits des EPA des membres organisationnels à tout usager en fonction de sa navigation dans une source d'information telle que le Web. L'individu contribue alors au capital du groupe (aspect donnant-donnant) en stockant les documents qu'il juge utile de conserver, ils seront par la suite valorisés rétroactivement et de façon non-intrusive pour aider les autres usagers lors de leurs navigations.

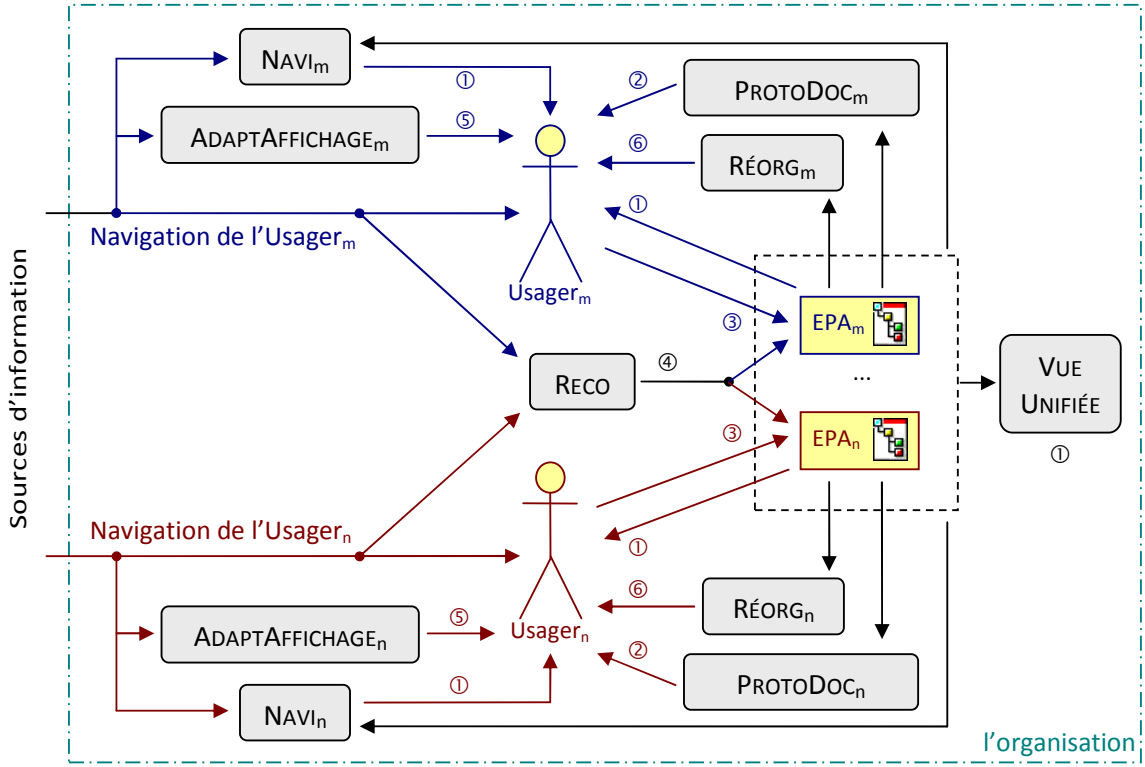


Figure II.5.1 – Schéma synoptique de l'architecture proposée représentant une organisation minimale composée de deux individus. Les flèches représentent les flux de données entre les usagers, leurs EPA et les six processus interdépendants. Les symboles ① à ⑥ font référence au cycle de vie du document (figure I.1.1).

Nous présentons dans les sections suivantes le modèle unifié ainsi que les processus intégrés couvrant les six activités du cycle de vie du document (figure I.1.1).

5.1 ADAPTAFFICHAGE : améliorer l'exploitation des documents

L'annotation collective a été retenue comme élément fédérateur des activités documentaires. Les individus peuvent en créer pour commenter, mémoriser, ou discuter de tout ou partie d'un document. Par la suite, chaque individu visualise en contexte, au sein des documents, les annotations pour lesquelles il dispose de droits d'accès adéquat. Pour une organisation comprenant de nombreux membres qui annotent massivement les documents, des problèmes liés au passage à l'échelle du dispositif de visualisation surviennent (Bouvin *et al.*, 2002). En effet, l'affichage des

annotations au sein du document, replacées sur leur point d'ancrage, peut déranger le lecteur en altérant significativement leur contenu, au détriment de leur lisibilité (figure I.3.6, p. 38).

Afin de limiter la profusion des annotations ②, notamment celles qui présentent des contenus invalides ou *a fortiori* incorrects (publicité, diffamation, pornographie, etc.), le système d'annotation JotBot associe aux annotations une « durée de vie » qui est prolongée par les lecteurs, lorsqu'ils les jugent utiles (Vasudevan et Palmer, 1999). Ce fonctionnement a le mérite d'éliminer incrémentalement les annotations « graffitis ». Toutefois, des annotations pertinentes seront effacées lorsqu'une ressource est peu visitée, ou bien à cause du manque de motivation ou d'implication des lecteurs qui ne cliquent pas sur le bouton de vote approprié. Plutôt que de supprimer quasi arbitrairement des annotations pour en limiter le nombre, nous spécifions dans cette section des indicateurs exploités au travers du processus ADAPTAFFICHAGE. Ils visent à réduire les efforts cognitifs que le lecteur met en œuvre pour filtrer le bon grain de l'ivraie lors de l'activité d'exploitation des documents ⑤ ; ce dernier pouvant alors juger sur pièce de la pertinence des annotations.

1. Une première amélioration de l'affichage consiste à l'adapter aux préférences de l'usager en ne restituant que les annotations exprimées dans une des Langues qu'il a spécifiées, cf. le modèle UML de figure II.2.3. Bien que très basique, nous n'avons pas identifié dans l'étude des systèmes d'annotation (section I.3.2.3) cette stratégie qui évite d'afficher des éléments incompréhensibles à l'usager.

2. Une seconde adaptation consiste à restituer les annotations de façon plus informative qu'avec une icône telle que «  » d'Amaya (Kahan *et al.*, 2002). Pour ce faire, nous associons au point d'ancrage de toute annotation sa sémantique et l'opinion de son créateur au travers de pictogrammes visuels (icônes) correspondant aux types spécifiées dans le tableau II.2.1. De plus, si un fil de discussion est associé à une annotation, le nombre de ses réponses est alors indiqué. Par ailleurs, la présence de commentaires et de références est également identifiable au travers de pictogrammes visuels. Enfin, les nouvelles annotations ancrées sur un document depuis sa dernière Visite (figure II.2.3) sont mises en exergue grâce à une icône de notification spécifique. Grâce à ces éléments d'interface, le lecteur peut se focaliser sur certaines annotations, selon des critères de choix qui lui sont propres.

3. Une troisième adaptation concerne les AnnotationsArgumentatives et les débats qu'elles suscitent. Lorsque de nombreux débats sont ancrés sur un document, le lecteur est tenté de tous les consulter tour à tour. Pour chaque débat, il peut alors déduire mentalement l'opinion globale de ses participants pour évaluer la validité sociale de l'AnnotationArgumentative. Bien que nécessaire à la lecture critique et approfondie d'un document, consulter chaque débat et synthétiser leurs opinions demande un temps non négligeable, ainsi qu'une charge cognitive élevée. En effet, il faut tout d'abord identifier les opinions exprimées dans chaque argument, puis les synthétiser de façon récursive, des feuilles vers la racine de l'arbre représentant le débat (figure II.2.1). Afin de soulager l'usager de ce fardeau, nous mettons à contribution les algorithmes de validation sociale définis dans le chapitre II.3. Ainsi, les usagers sont informés du degré de consensus (resp. controverse) atteint dans chaque débat. Ils peuvent alors choisir de se focaliser sur des informations stabilisées, ou bien sur des discussions pour lesquelles les divers participants n'ont pas encore trouvé de consensus.

5.2 PROTODOC : améliorer la création et la finalisation de documents

Une grande partie de la connaissance d'une organisation est contenue dans les documents qu'elle produit, étant donné que les travailleurs du savoir consacrent beaucoup de temps à la rédaction de documents. Pour ce faire, ils extraient des informations à partir de divers documents avant de les synthétiser en un nouveau document, tel qu'un rapport. Eu égard à l'architecture proposée, l'*AnnotationAncrée* fournit au lecteur un moyen efficace pour collecter des pépites informationnelles et y associer un Contenu pour garder une trace de son interprétation, par exemple. Pour un projet spécifique, tel que l'analyse quotidienne des cours de la Bourse par exemple, l'usager peut créer un Répertoire idoine dans son EPA afin de regrouper l'ensemble des bribes de documents qu'il désire conserver à cet effet. Par ailleurs, la notion de Droit associée aux Répertoires restreint l'accès aux répertoires partagés entre différentes *Entités* collaborant à des tâches telles que la recherche et l'analyse collectives.

Afin d'assister l'activité de création de document ②, le processus PROTODOC ébauche un proto-document à partir des Répertoires de l'EPA désignés par l'usager. Ce résultat permet notamment de constituer le squelette d'un dossier d'analyse. Pour chaque annotation des répertoires concernés, le processus intègre au proto-document les *Ancrages* des Ressources annotées, leur validité sociale, ainsi que les diverses informations fournies par l'annotateur : Contenu, Tags, références, etc. Par la suite, l'usager peut compléter et retravailler cette ébauche à l'aide du traitement de texte de son choix.

5.3 RECO : améliorer la diffusion des documents

L'étude de Feldman (2004) montre à quel point une diffusion d'information limitée au sein de l'organisation mène à des situations contre-productives et coûteuses : méconnaissance des compétences au sein de l'organisation, perte de temps en recherches infructueuses, recréation d'information, etc.

C'est pourquoi nous proposons de tirer parti des informations recueillies par les individus, qui demeurent manifestement en sommeil dans le cas contraire. Pour ce faire, nous offrons aux Usagers la possibilité d'émettre des Recommandations *manuelles* à destination d'autres *Entités* de sa connaissance (figure II.2.3). Une fonctionnalité duale permet à chaque Usager de s'enregistrer pour recevoir des notifications concernant les événements sur les *Entités* de son choix. De ce fait, on peut notamment spécifier les répertoires des EPA pour lesquels on désire connaître tout ajout d'annotation. Cette demande proactive de notification est similaire à la syndication de flux RSS sur une page Web (Hammond *et al.*, 2004).

Comme le souligne la section I.2, la diffusion manuelle est limitée par divers facteurs liés à l'usager : son réseau social, sa volonté de partager alors que l'information est fréquemment assimilée au pouvoir, les efforts cognitifs en jeu, etc. Pour dépasser ces limites, nous proposons de compléter la diffusion manuelle par une diffusion *automatique* ④ grâce au processus RECO. Ce dernier vise à exploiter la capacité des individus à rechercher de l'information, de façon non-intrusive. En effet, il considère tout document introduit dans l'organisation par un usager comme candidat à la recommandation car d'autres membres possédant des centres d'intérêt similaires peuvent en bénéficier. Ce processus formalisé dans (Chevalier, 2002; Chevalier et Julien, 2003) opère de la fa-

çon suivante : chaque document introduit dans l'organisation (objet de la navigation d'un usager, par exemple) est indexé (classes Terme et Index). Le système calcule ensuite sa similarité thématique avec les répertoires des EPA des autres Usagers. Pour ce faire, chaque répertoire est doté d'un classifieur créé sur le principe du « méga-document » (Klas et Fuhr, 2000), en considérant un répertoire comme un document dont le contenu résulte de la concaténation des documents qu'il contient. Enfin, le document indexé est recommandé dans le répertoire le plus spécifique de l'EPA considéré, à condition que la similarité calculée excède un seuil calculé dynamiquement. Ce seuil permet de limiter le nombre des recommandations de façon à ne pas surcharger l'usager.

5.4 RÉORG : aider à l'organisation thématique des documents

Concernant l'activité de gestion de l'EPA ⑥, un Usager crée donc une *AnnotationAncrée* pour conserver de l'information dont il prend connaissance au cours de ses activités. Il peut ensuite organiser cette annotation selon ses besoins (par thème, par projet, chronologiquement, etc.) au sein de son EPA composé des Répertoires qu'il crée et gère. Le processus RÉORG adjoint à l'activité ⑥ assiste l'usager lors de la réorganisation de son EPA. Cette tâche est reconnue hautement cognitive et fréquemment reportée à une date ultérieure (par ex. : quand j'aurai plus de temps, une fois ce projet terminé), comme l'indiquent Abrams *et al.* (1998) et plus récemment Jones (2007). Or, sur une longue période d'utilisation, réorganiser son EPA est indispensable afin de pouvoir accéder à l'information efficacement. Par exemple, certains répertoires « grossissent » quotidiennement lorsque l'usager complète sa connaissance du domaine associé, en y insérant de plus en plus d'annotations pointant vers les passages d'intérêt des documents.

C'est dans un pareil cas que le raffinage (éventuellement thématique) du répertoire en sous-répertoires est requis, afin de structurer sa perception de son EPA tout en conservant la faculté d'accéder à ces informations en un temps minimal. C'est pourquoi le processus RÉORG se base sur l'EPA d'un usager pour lui proposer une réorganisation thématique suite à sa demande explicite, qu'il peut accepter partiellement ou en totalité. Ce processus initialement proposé dans (Cabanac, 2002; Chevalier, 2002) met en œuvre une succession d'algorithmes classiques en Recherche d'Information (Manning *et al.*, 2008, ch. 17) : l'Indexation du contenu des Ressources annotées (Baeza-Yates et Ribeiro-Neto, 1999, ch. 2), une classification ascendante hiérarchique (Jardine et van Rijsbergen, 1971) fournissant une arborescence thématique binaire de l'EPA, un seuillage de cette dernière comme proposé par Maarek et Ben-Shaul (1996), ainsi qu'un étiquetage des répertoires obtenus à l'aide du coefficient du χ^2 pour déterminer les termes les plus discriminants de chaque répertoire.

5.5 NAVI : améliorer l'accès à l'information

L'architecture proposée améliore les deux modalités qu'Agosti (1996) situe au cœur de l'activité de recherche d'information ① : l'interrogation et la navigation.

Premièrement, nous proposons de compléter la modalité d'interrogation par la prise en compte des annotations collectives. L'idée est d'exploiter les annotations des lecteurs et leurs fils de discussion en tant que contributions reflétant un *feedback* social, dans le but d'améliorer le rappel (en retrouvant davantage de documents pertinents par rapport à la requête) ainsi

que la précision des résultats de recherche (en ne retrouvant que les documents pertinents). Au sujet du rappel, le « problème du vocabulaire » formulé par Furnas *et al.* (1987) indique que la requête d'un usager contient rarement (< 20 %) les mêmes termes que ceux constituant les documents pertinents pour le besoin en information de l'individu. Or, Fraenkel et Klein (1999) rapportent de nombreux exemples où le contenu des annotations permet de trouver davantage de documents pertinents, car les termes employés par les lecteurs dans les annotations sont parfois complémentaires de ceux présents dans le passage rédigé par l'auteur du document. De ce fait, la prise en compte des annotations permet d'améliorer le rappel, en retrouvant des documents de façon indirecte, ainsi que la recherche contextuelle en retrouvant des passages annotés plutôt que des documents entiers. Par ailleurs, concernant la précision des recherches, la considération des Contenus des annotateurs permet la désambiguïsation des documents annotés, ainsi que l'intégration de leurs termes dans le processus d'indexation. De plus, la validation sociale des débats (AnnotationArgumentative) offre un moyen de caractériser une Ressource en tant que fiable, controversée, populaire, abandonnée, etc. La prise en compte de tels indicateurs permet alors d'adapter le moteur de recherche aux préférences des usagers.

Deuxièmement, le processus NAVI introduit par Chevalier (2002) assiste l'usager de façon non-intrusive, alors qu'il navigue sur le Web. Pour ce faire, ce processus exploite les EPA de l'organisation — résultant des efforts cognitifs de classement de chacun — afin d'en extraire des documents pertinents eu égard à la navigation courante de l'usager. Ces documents identifiés dans le capital documentaire collectif (les EPA) sont alors recommandés à l'usager de façon synchrone lors de sa navigation. La mise en œuvre de ce processus est originale à deux égards :

1. Le processus NAVI exploite la capacité des membres organisationnels à trouver, filtrer et classer des documents pertinents pour l'organisation, étant donné que les recommandations calculées proviennent de leurs EPA. De ce fait, des documents recherchés il y a longtemps puis oubliés dans les répertoires des individus sont automatiquement rentabilisés en les proposant à leurs collègues. Ici nous posons deux hypothèses de travail : *i*) un document contenu dans un répertoire a été jugé intéressant par le propriétaire de l'EPA car il a mis en œuvre un effort cognitif en sélectionnant le répertoire le plus approprié dans son EPA (Rucker et Polanco, 1997). De plus, *ii*) d'autres membres du groupe ayant des intérêts communs peuvent être intéressés par un tel document pour réaliser leurs propres activités.

2. Le second aspect original concerne l'algorithme d'identification des documents à recommander en fonction de celui qui est visualisé par l'usager. Les approches classiques se basent uniquement sur le contenu des documents, après une phase d'indexation. Prenant le contrepied de ces approches, nous avons défini une mesure de similarité ne nécessitant pas de les indexer : elle est basée sur l'usage des documents (chapitre II.4) qui évalue à quel point les individus classent les documents ensemble pour réaliser leurs activités. Il est à noter que les documents sont recommandés à l'usager accompagnés de leurs chemins dans les EPA, de façon à ce que l'individu puisse appréhender le contexte et l'intention de classement du propriétaire du document recommandé. L'indication de sa provenance évite que l'usager n'ait l'impression d'accéder au document par « téléportation », ce qui est souvent reproché aux moteurs de recherche (comparés aux systèmes permettant d'accéder aux documents par navigation) selon Teevan *et al.* (2007). Cette indication lui permet également de parcourir directement les EPA des autres usagers.

Le processus NAVI recommande de façon synchrone (durant la navigation) des documents introduits dans l'organisation en fonction d'un besoin précis (selon le document visualisé). Afin d'améliorer l'accès au capital organisationnel en dehors de la tâche de navigation, le processus VUE UNIFIÉE en offre une vision globale, permettant ainsi aux membres organisationnels de prendre connaissance de leur environnement de travail et des ressources capitalisées par le groupe. Ce processus original est exposé dans le chapitre suivant.

6 Visualisation multi-facettes et exploration du capital organisationnel

"If HP knew what HP knows, we would be three times as profitable."
Lewis E. Platt (1941 — 2005), cité par Hinds et Pfeffer (2003).

LES PROCESSUS INTÉGRÉS présentés dans le chapitre précédent aident chaque individu en lui proposant des documents intéressants de façon synchrone (NAVI) ou asynchrone (RECO). De plus, la restitution des documents est adaptée afin que le lecteur puisse apprécier des indicateurs quant aux annotations associées (ADAPTAFFICHAGE). Enfin des fonctionnalités rassemblent les annotations créées dans un nouveau document (PROTODOC) et assistent l'usager lors de la maintenance de leur EPA (RÉORG). Ces processus apportent une assistance au niveau des usagers, que nous qualifions de « microscopique ».

Par ailleurs, ces mêmes individus ont tout intérêt à compléter leur connaissance de l'environnement informationnel dans lequel ils évoluent, correspondant à l'organisation dans son ensemble. Pour ce faire, une assistance au niveau « macroscopique » leur permettra d'être plus efficaces en évitant les pièges du retravail inutile, résultant souvent d'une méconnaissance des compétences et connaissances voisines. Or, la section I.2.3 a souligné le fait que les EPI tels que les arborescences personnelles de fichiers ne sont que trop peu valorisées car difficiles à partager. Par conséquent, ce capital documentaire interne sommeille dans les EPI, alors qu'il serait utile à chaque membre organisationnel s'il était mis en valeur par un système de type donnant-donnant.

Dans notre approche basée sur les EPA, nous souhaitons valoriser ces espaces d'information au maximum car ce sont de véritables mines d'informations en relation directe avec les activités de leurs propriétaires. Dans cette optique, le présent chapitre définit le processus VUE UNIFIÉE basé sur une interface multi-facettes d'accès au capital documentaire de l'organisation (Cabanac, 2008b). Cette proposition ne vise pas à constituer une nouvelle base documentaire ou à enrichir une mémoire d'entreprise ; elle se veut plutôt complémentaire à ces approches. Afin de rendre

possible la visualisation et l'exploration du capital documentaire, notre proposition exploite les EPA des usagers, sans nécessiter de modification structurelle ou organisationnelle spécifique. Ainsi, un avantage de l'interface multi-facettes proposée réside dans son aspect non-intrusif : la façon de travailler des membres organisationnels — en termes de pratiques documentaires — n'est pas remise en cause. Concrètement, l'interface proposée (Cabanac *et al.*, 2009a) permet la visualisation des documents et des personnes selon deux axes : thématiquement (en fonction du contenu des documents possédés) ou selon leur usage (mesurant la régularité d'organisation des documents les uns par rapport aux autres dans les EPA).

6.1 Interface multi-facettes d'accès au capital organisationnel

L'interface proposée à destination des membres organisationnels vise à répondre aussi bien à des besoins opérationnels que stratégiques :

– **Besoins opérationnels.** L'adjectif « opérationnel » fait référence à la réalisation des tâches affectées aux membres organisationnels. Dans ce cadre, l'interface offre une vue globale des documents de l'organisation et en permet l'exploration, par thématique ou par usage (chapitre II.4). Ces deux mesures de similarité sont complémentaires comme l'illustre la figure II.6.1 résultant de l'expérimentation rapportée dans (Cabanac *et al.*, 2007a). Les deux graphes de cette figure montrent les deux types de similarités calculées à partir des douze documents provenant des EPA du multi-arbres de la figure II.4.1 (p. 66).

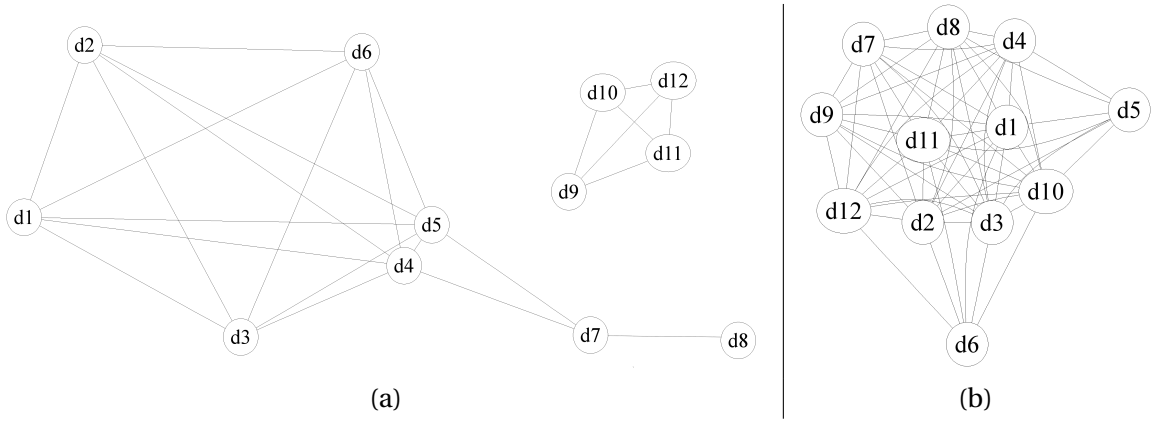


Figure II.6.1 – Comparaison de la similarité des documents sur l'usage (a) par rapport à leur similarité sur le contenu (b). La longueur des arcs est inversement proportionnelle à la similarité entre les nœuds associés qui représentent les documents d_1 à d_{12}.

Deux groupes de documents sont distinguables dans (II.6.1.a) contre un seul groupe dans (II.6.1.b). De plus, d_1 et d_{11} sont proches thématiquement alors qu'ils ne sont pas utilisés ensemble. Enfin, d_4 et d_5 sont les documents les plus proches par l'usage alors que leurs thématiques sont relativement éloignées. La complémentarité des mesures de similarité de thématique *versus* d'usage est confirmée dans le chapitre III.3 (p. 117). Grâce à l'interface basée sur ces deux mesures, chaque individu peut identifier, à partir de l'ensemble des EPA, les documents connexes ou complémentaires à ses propres documents. Hertzum et Pejtersen (2000) soulignent à quel point les travailleurs du savoir cherchent des documents pour trouver les individus associés, et vice versa. Partant de cette nécessité, l'interface permet de

basculer de l'une à l'autre de ces deux dimensions.

– **Besoins stratégiques.** L'adjectif « stratégique » concerne les activités propres au pilotage de l'organisation, notamment au service des ressources humaines. Dans ce contexte, notre proposition permet de visualiser, au travers des EPA, les activités documentaires de tout ou partie des membres organisationnels. Une application directe consiste à identifier les documents utilisés pour réaliser les activités associées à un poste donné. Prendre en compte cette information peut aider à trouver des personnes-ressources dans un domaine donné, à composer un groupe de travail adapté aux besoins d'un projet, à identifier les centres d'intérêts émergeants, à lutter contre le *turnover* en anticipant les compétences à renouveler (Boyer *et al.*, 2007)… Les auteurs de ces travaux proposent de cartographier une organisation en fonction du contenu des documents et des relations établies entre les différents acteurs. Par rapport à cette approche, notre proposition introduit la notion d'usage, permettant d'identifier des liens complémentaires entre les documents en selon leur organisation dans les EPA.

Nous exposons dans cette section l'interface multi-facettes proposée, en décrivant en premier lieu ses différentes composantes complémentaires qui donnent accès au capital organisationnel (aspect statique). Dans un deuxième temps, nous formalisons les diverses actions que l'usager peut réaliser sur l'interface (aspect dynamique) afin de pouvoir explorer le capital organisationnel au travers des facettes. Enfin, nous présentons dans un troisième temps les aspects relatifs à la mise en œuvre de cette interface : concepts de similarité de contenu et d'usage, ainsi que techniques de visualisation.

6.1.1 Aspect statique de l'interface : représentation du capital organisationnel

L'interface proposée permet de visualiser les deux dimensions du capital organisationnel : les documents et les personnes. Pour une dimension donnée, l'usager peut explorer un ensemble d'éléments (un groupe de documents ou de personnes) ou un seul élément (un document ou une personne). La combinaison de ces deux paramètres représente quatre cas, matérialisés par des « vues » dans la figure II.6.2 qui schématise l'architecture globale de l'interface proposée.

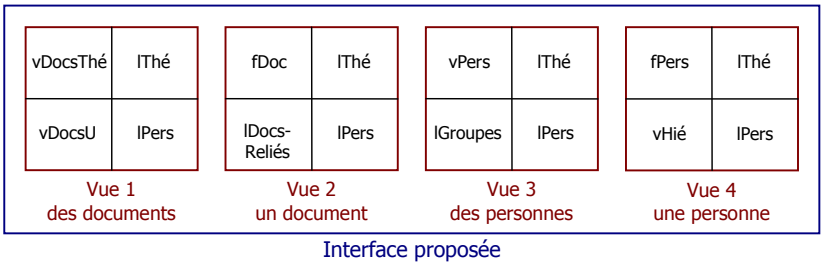


Figure II.6.2 – Architecture générale de l'interface comprenant des vues et des facettes.

Une vue peut être assimilée à une fenêtre dans une interface graphique. De plus, chaque vue comprend quatre « facettes » où figurent des informations relatives à la vue choisie. Elles permettent également l'exploration du capital organisationnel car c'est par leur intermédiaire que l'usager passe d'une vue à l'autre (aspect dynamique). Pour chacune des quatre vues, le tableau II.6.1 recense les facettes disponibles. On distingue trois types de facettes : visualisation, liste et fiche. Chaque nom de facette est préfixé par l'initiale de son type, soit « v », « l » ou « f ».

Facettes		Dimensions			
		Document		Personne	
Nom	Description	Groupe *Vue 1*	Unité *Vue 2*	Groupe *Vue 3*	Unité *Vue 4*
fDoc	fiche d'un document		✓		
fPers	fiche d'une personne				✓
lDocsReliés	liste des documents liés		✓		
lThé	liste de thématiques	✓	✓	✓	✓
lGroupes	liste de groupes			✓	
lPers	liste de personnes	✓	✓	✓	✓
vDocsThé	vue thématique des documents	✓			
vDocsU	vue de l'usage des documents	✓			
vHié	hiérarchie d'une personne				✓
vPers	représentation de personnes			✓	

Tableau II.6.1 – Description des facettes associées aux quatre vues composant l'interface.

Nous détaillons dans les sections suivantes chacune des quatre vues, en décrivant les informations accessibles par l'intermédiaire de chaque facette et en donnant un scénario d'utilisation.

6.1.1.1 Vue 1 : représentation d'un groupe de documents

Les facettes de cette vue permettent de visualiser les documents de l'organisation regroupés par thématique (vDocsThé) et par usage (vDocsU). Ces deux modalités sont respectivement basées sur le contenu des documents et sur leur organisation au sein des EPA. De façon intuitive, deux documents sont d'autant plus proches par le *contenu* qu'ils partagent un nombre important de termes. D'autre part, deux documents sont d'autant plus proches par l'*usage* qu'ils sont organisés ensemble dans les EPA. Le détail des mesures de similarité sur le contenu et sur l'usage est présenté en section II.6.1.3.2. L'usager interagit avec ces facettes en sélectionnant un ou plusieurs documents, il peut alors se focaliser sur ce(s) dernier(s). La facette lPers contient la liste des propriétaires des documents sélectionnés, triable par nombre de documents. Enfin, la facette lThé liste les thématiques concernant les documents sélectionnés, par ordre d'importance.

Grâce à cette vue, un individu obtient une représentation des thématiques du fonds documentaire constitué à partir des documents extraits des EPA. En se focalisant sur une thématique particulière, il visualise les personnes qui possèdent ces documents. Il voit également quels documents sont classés avec les documents sélectionnés. Ces documents connexes, issus des EPA de l'organisation, apportent des informations complémentaires par rapport à la sélection originale de l'usager. En reflétant les associations d'idées des membres organisationnels, la facette vDocsU offre un véritable retour sur investissement qui rentabilise l'effort de chaque membre.

6.1.1.2 Vue 2 : représentation d'un seul document

Cette vue présente la fiche d'un document (fDoc) qui donne accès à son titre, à son contenu et aux chemins absolus des EPA qui le contiennent (par exemple, « `/home/userX/informa-`

tique/bdr/indexation/arbreBalancé/cours.pdf » et « /home/userY/inventeurs/science/info/Rudolf_Bayer/bio.html »). La date de création du document dans chacun de ces chemins est précisée. Les thématiques du document sont listées dans la facette lThé. Les individus qui le possèdent sont recensés dans la facette lPers. Enfin, les documents connexes (utilisés avec le document visualisé) sont listés dans la facette lDocsReliés, ordonnés par similarité d'usage.

Au travers des facettes de cette vue, l'usager identifie les thématiques traitées dans un document. Il connaît également les autres individus qui l'ont rangé dans leur EPA ; les noms des chemins absolus associés fournissent des indications complémentaires sur l'utilisation qui en est faite. Comme l'usager identifie les personnes intéressées par le document, il peut par la suite explorer leurs EPA pour trouver d'autres documents intéressants et éventuellement prendre contact avec eux. Cette fonctionnalité répond aux besoins identifiés dans (Hertzum et Pejtersen, 2000).

6.1.1.3 Vue 3 : représentation d'un groupe de personnes

Cette troisième vue représente dans la facette vPers un ensemble de personnes et les liens qui les unissent, qu'ils soient d'usage ou de thématique. Cette facette privilégie la visualisation des liens, elle est complétée par la facette lPers qui liste les personnes visualisées. Au sein de l'organisation, chaque individu fait partie de groupes explicites (équipes, groupes de travail, commissions...). Ces derniers sont représentés dans la facette lGroupes : elle contient les noms et le nombre de représentants des groupes distincts correspondant aux personnes visualisées dans vPers, par exemple « Service des ventes (12) ». Enfin, la facette lThé recense les thématiques associées aux EPA des personnes visualisées, triées par nombre de documents associés.

Cette vue permet à un usager d'identifier les intérêts thématiques caractérisant tout groupe de personnes, qu'il soit explicite (une équipe mentionnée dans l'organigramme) ou tacite (des personnes qui ont des affinités, qui déjeunent ensemble, etc.). De ce fait, un membre organisationnel peut identifier et explorer par la suite les thématiques de son équipe. Cette fonctionnalité est très utile en phase d'intégration d'un nouveau collaborateur, lorsque ce dernier doit s'adapter et se former en assimilant les thématiques manipulées par son équipe d'accueil (Boyer *et al.*, 2007). De la même façon, l'identification des thématiques principales d'une équipe, à partir des EPA, peut aider le service des ressources humaines à établir des fiches de poste. Ces dernières pourront notamment être utilisées pour la création ou le renouvellement d'un poste.

6.1.1.4 Vue 4 : représentation d'une seule personne

Une personne est représentée par sa fiche (fPers) qui contient les informations suivantes : identité (nom, prénom) et groupes d'appartenance. Une représentation hiérarchique des documents structurés dans son EPA est accessible au travers de la facette vHié. La liste des thématiques relatives à son EPA est présentée dans la facette lThé, elles sont classées par ordre alphabétique ou par importance décroissante. Enfin, la facette lPers recense les personnes qui partagent les mêmes thématiques ou qui utilisent les documents de la même façon que celle représentée dans cette vue.

Un scénario concret d'utilisation consiste, pour un usager donné, à visualiser sa propre fiche pour identifier les personnes proches de lui (par thématique ou par usage). Par la suite, la visualisation de leur fiche lui permet de connaître les thématiques qui les caractérisent. Il peut alors

explorer le contenu de leurs EPA en fonction des thématiques qui l'intéressent et de leur structure.

6.1.2 Aspect dynamique de l'interface : exploration du capital organisationnel

L'interface proposée permet de visualiser le capital organisationnel selon quatre vues spécifiques. Afin de permettre l'exploration et la navigation dans ce capital, nous définissons dans cette section deux types d'interaction entre l'usager et l'interface : l'interaction « intra-vue » et l'interaction « inter-vues ».

Au sein d'une vue quelconque, l'interaction intra-vue consiste à automatiquement répercuter la sélection d'un ou de plusieurs éléments d'une facette sur les trois autres facettes de la vue. Par exemple, la sélection d'un ensemble de thématiques associées à une personne (dans la facette lThé de la vue 4) permet d'identifier, au même moment, ces thématiques dans l'EPA de la personne (facette vHié) et de voir les personnes qui partagent ces mêmes thématiques (facette lPers). Alternativement, l'usager peut formuler une requête composée de mots-clés et de connecteurs booléens afin de sélectionner les éléments correspondants. Concrètement, chaque facette met en évidence les éléments associés à la sélection grâce à une mise en forme adaptée (couleur différente, graisse de la police, etc.). En fait, l'interaction intra-vue permet de localiser un même élément dans toutes les facettes qui constituent une vue, ces facettes proposant des représentations complémentaires de l'information extraite des EPA.

Le second type d'interaction introduit, appelé inter-vues, permet la navigation d'une vue à l'autre. Concrètement, en fonction d'une action réalisée sur une facette, l'interface remplace la vue actuelle par une autre vue répondant plus précisément au besoin exprimé. La sélection d'une personne au sein de la facette vPers (vue 3) permet par exemple de basculer sur la vue 4, car elle apporte davantage d'informations sur cette personne. De cette façon, les différentes actions réalisées sur les facettes permettent d'explorer le capital organisationnel. Nous avons modélisé la dynamique de l'interface à l'aide du diagramme états-transitions de la figure II.6.3.

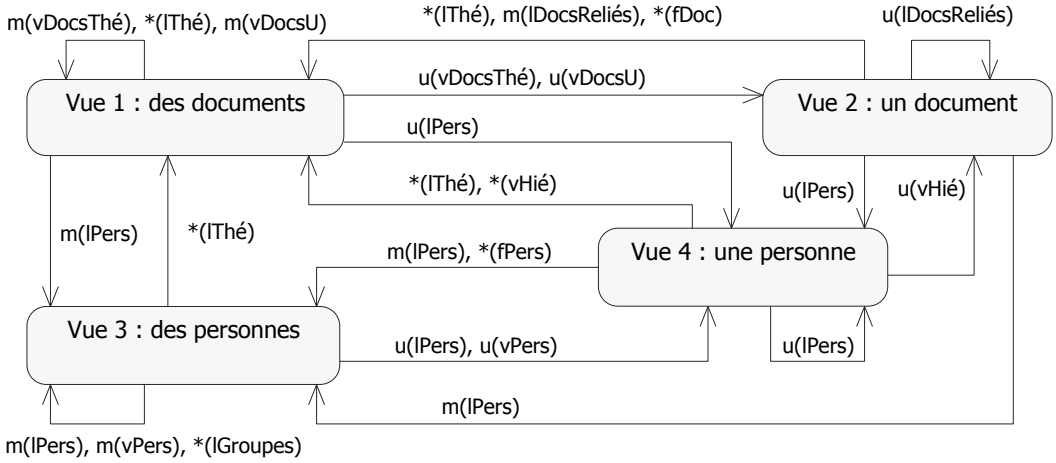


Figure II.6.3 – Diagramme états-transitions décrivant la dynamique de l'interface.

Les quatre états représentent les vues explicitées dans le tableau II.6.1. Une transition d'un état e_1 vers un état e_2 est déclenchée par des actions sur une facette de la vue correspondant à e_1. Le détail de ces actions figure sur l'étiquette de la flèche reliant les deux états. Plusieurs actions

possibles sont séparées par une virgule. La notation d'une action est du type $s(f)$ où s représente une sélection et f une facette. Plus précisément la sélection multiple est notée « m », la sélection d'un seul élément est notée « u », et « * » désigne une sélection multiple ou unique. Par exemple, l'étiquette « m(lPers), *(fPers) » entre la vue 4 et la vue 3 signifie qu'au travers de la vue 4, une sélection multiple dans la liste des personnes lPers ou une sélection quelconque dans la fiche de la personne fPers mènent à la vue 3.

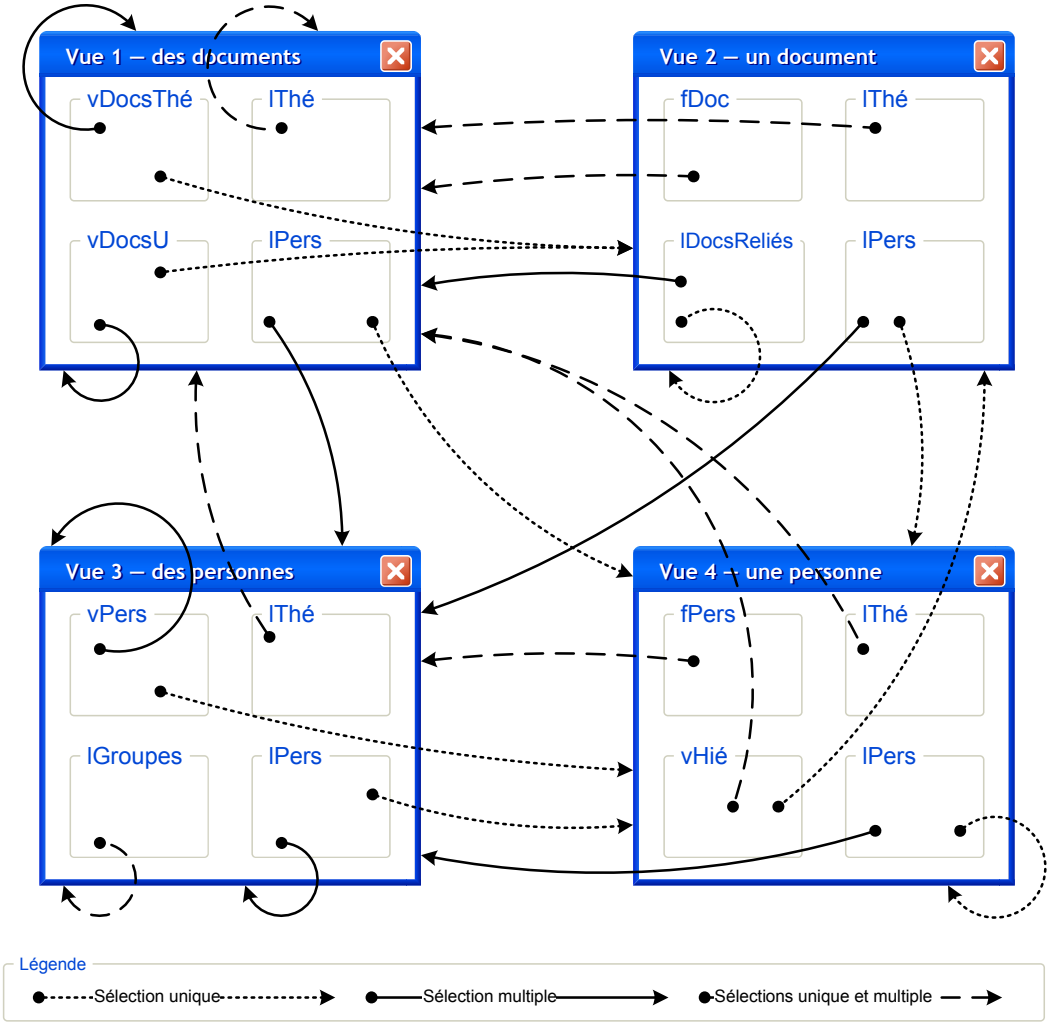


Figure II.6.4 – Synthèse des aspects statique et dynamique de l'interface proposée.

Afin de donner une vision d'ensemble de l'interface proposée ainsi que des interactions possibles entre les vues, la figure II.6.4 synthétise les aspects statique (tableau II.6.1) et dynamique (figure II.6.3). Les nombreux liens entre les vues montrent le caractère interactif de l'interface, qui facilite la navigation dans le capital organisationnel. Le calcul des facettes incorporées dans les quatre vues fait l'objet de la section suivante.

6.1.3 Mise en œuvre de l'interface multi-facettes proposée

La mise en œuvre de l'interface multi-facettes d'accès au capital organisationnel nécessite de modéliser les données sources à partir desquelles les similarités sur le contenu et sur l'usage sont

calculées. Nous considérons les données issues du SI de l'organisation, incluant de ce fait les EPA. Ces données peuvent alors être représentées au sein des facettes composant les quatre vues.

6.1.3.1 Modélisation des composants du SI nécessaires à notre approche

L'interface proposée ne vise pas à constituer une nouvelle source d'information, mais plutôt à explorer les EPA gérés par les membres organisationnels. Notre approche est de ce fait basée sur l'exploitation du SI pour extraire des données relatives aux personnes et aux documents de l'organisation. Comme recommandé par la CNIL, nous ne tenons pas compte des répertoires et fichiers personnels afin de respecter la vie privée des membres organisationnels [1]. Ces éléments sont identifiables grâce à leur nom, qui contient une chaîne de caractères spécifique telle que « perso ». De telles chaînes peuvent être définies au niveau de l'organisation. Ainsi, seuls les répertoires et fichiers non personnels sont exploités. Nous reprenons dans la figure II.6.5 les éléments du modèle unifié que nous considérons dans ce chapitre, en introduisant les attributs nécessaires à la compréhension.

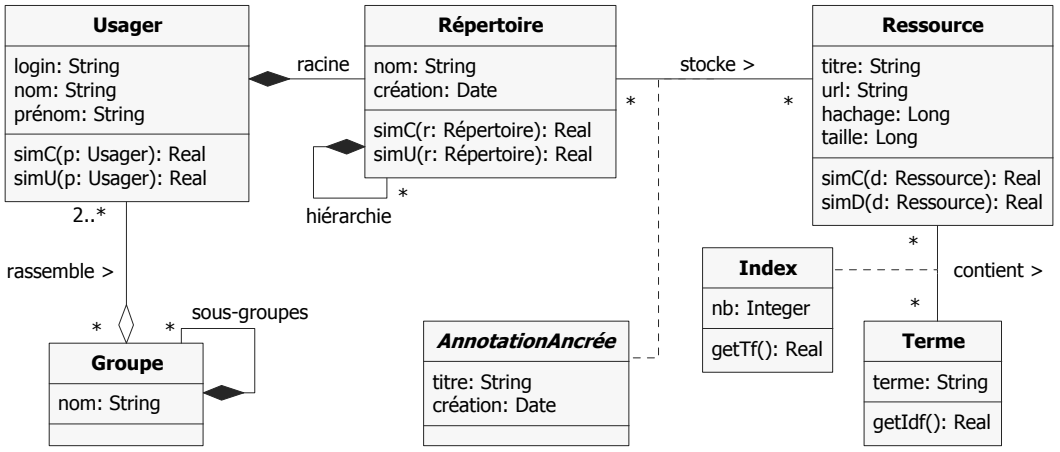


Figure II.6.5 – Diagramme des classes UML représentant les données exploitées par l'interface.

Chaque membre est modélisé par la classe Usager, il est caractérisé par son login et son identité (nom et prénom). Il fait éventuellement partie de Groupes, généralement explicités dans l'organigramme de l'organisation ou bien constitués pour des activités spécifiques telles que des projets. Une personne possède et gère son EPA (hiérarchie de Répertoires). Chaque répertoire peut contenir des sous-répertoires et des *AnnotationsAncrées* sur des Documents. Nous avons représenté cette classe par une classe d'association pour ne pas faire figurer la classe *Ancrage* intermédiaire. Une *AnnotationAncrée* est caractérisée par sa date de création et le titre attribué par son propriétaire. L'attribut titre d'un Document accessible à une url donnée correspond au titre extrait de ses méta-données. Un même document n'est indexé qu'une seule fois même s'il est annotaté plusieurs fois. Le résultat de son indexation (classes Terme et Index) est mis à jour lorsqu'une modification de son contenu est détectée. Pour ce faire, nous comparons une valeur de hachage calculée

1. « un message envoyé ou reçu depuis le poste du travail [...] revêt un caractère professionnel, sauf indication manifeste dans l'objet du message ou dans le nom du répertoire où il pourrait avoir été archivé par son destinataire qui lui conférerait alors le caractère et la nature d'une correspondance privée protégée par le secret des correspondances. » (Bouchet, 2004, p. 25)

suite à la visite du document par un usager avec l'attribut hachage calculé lors de l'insertion du document ou de sa dernière mise à jour. En complément de cet attribut hachage, la donnée de la taille des documents permet de limiter le problème des collisions de hachage (deux contenus différents possédant une valeur de hachage identique).

La section suivante décrit l'exploitation du contenu des documents représenté par les classes Terme et Index (resp. de l'organisation des documents représentée par la classe Répertoire) et le calcul d'une mesure de similarité thématique (resp. liée à l'usage des documents) correspondant à la méthode simC (resp. simU).

6.1.3.2 Mesures de similarité sur le contenu et sur l'usage des documents

Les informations présentées dans diverses facettes de l'interface sont basées sur le calcul de similarités thématique et d'usage. C'est pourquoi nous détaillons ces similarités avant d'en montrer l'exploitation par des techniques de visualisation adaptées.

Similarité basée sur le contenu des documents indexés Évaluer la similarité entre deux documents est une opération fondamentale dans le domaine de la RI (Baeza-Yates et Ribeiro-Neto, 1999, ch. 2). Une telle similarité est classiquement fonction du contenu textuel des documents indexés (section II.2.2.2). Plusieurs modèles mathématiques ont été proposés, le plus répandu étant le modèle vectoriel (Salton *et al.*, 1975) où chaque document est représenté par un vecteur dans l'espace vectoriel des termes distincts du corpus. Ainsi, un document d_i aura pour représentation $\vec{d}_i = (w_i^1, \ldots, w_i^n)$ où chaque $w_i^j \in \mathbb{R}_+$ correspond au poids du j^{e} terme dans le document d_i, sachant que le corpus comprend n termes. Classiquement, son poids dépend de deux facteurs : sa fréquence relative dans le document tf_i^j et l'inverse de sa fréquence dans le corpus idf^j. Le premier facteur, correspondant à la fonction getTf, est d'autant plus élevé que le terme est fréquent dans le document. Le second facteur, correspondant à la fonction getIdf, est d'autant plus élevé que le terme est rare dans le corpus car, dans ce cas, il a un fort pouvoir discriminant pour les documents qui le contiennent. Baeza-Yates et Ribeiro-Neto (1999, ch. 2) synthétisent les variantes proposées dans la littérature pour calculer ces deux facteurs, que nous ne détaillons pas ici. La combinaison des deux facteurs selon $w_i^j = tf_i^j \cdot idf^j$ fournit alors une valeur d'autant plus élevée que le terme est fréquent dans le document et globalement rare dans le corpus. Par la suite, le calcul de la similarité entre deux documents d_1 et d_2 repose sur une fonction appliquée aux deux vecteurs qui les représentent, par exemple $\cos(\vec{d}_1, \vec{d}_2)$.

Pour évaluer la similarité entre deux répertoires, nous exploitons l'approche du « méga-document » proposée par Klas et Fuhr (2000). Elle consiste à représenter un répertoire comme un document unique, créé en concaténant le contenu textuel des documents qu'il contient. Nous utilisons le même principe pour évaluer la similarité entre personnes, où une personne est représentée par un document unique créé en concaténant tous les documents de son EPA.

Dans la vue 1, la facette vDocsThé est construite à partir des valeurs de similarité calculées pour les documents pris deux à deux. Quant aux thématiques listées dans la facette lThé, elles correspondent aux termes issus de l'indexation, classés par fréquence décroissante. Enfin, la facette vPers de la vue 3 repose sur le calcul des similarités entre personnes prises deux à deux.

Similarité basée sur l'usage des documents classés dans les EPA Contrairement à la similarité de contenu basée sur le résultat de l'indexation, la similarité d'usage définie dans le chapitre II.4 repose uniquement sur la structure des EPA. Cette mesure n'évalue pas à quel point deux documents contiennent des termes identiques, mais plutôt à quel point ils sont utilisés ensemble par les individus.

Le calcul de similarités inter-documents sur l'usage est restitué dans les facettes vDocsU de la vue 1 et lDocsReliés de la vue 2. Concernant la similarité inter-personnes, elle figure dans la facette lPers.

6.1.3.3 Techniques de visualisation utilisées pour représenter documents et personnes

Comme le souligne la section I.2.4.2 (p. 19), pléthore de techniques et outils de visualisation ont été proposés dans la littérature (Herman *et al.*, 2000; Chen, 2006; Yang *et al.*, 2008). Or, nous devons sélectionner des visualisations adaptées à notre objectif : offrir une vue globale du capital organisationnel. Deux critères de choix primordiaux sont à considérer. Premièrement, la visualisation doit permettre de représenter des éléments en fonction de leur similarité (de thématique ou d'usage). Deuxièmement, elle doit permettre l'affichage d'un nombre d'éléments d'autant plus important que l'organisation considérée comprend de nombreux membres, ce qui a trait au problème du passage à l'échelle. En tout état de cause, la contribution de ce chapitre ne repose pas sur les choix effectués en matière de techniques de visualisation, mais plutôt sur l'exploitation conjointe des similarités de thématique et d'usage dans l'interface multi-facettes proposée. De ce fait, les choix que nous présentons dans cette section peuvent être remis en question en fonction de critères propres à l'organisation.

Pour représenter les liens d'usage entre les documents et entre les personnes, nous avons retenu une visualisation sous forme de graphe. Cette représentation favorise l'identification de groupes de documents utilisés ensemble, formant des sous-graphes connexes. Les nœuds représentent les documents ou personnes, ils sont reliés par des arcs dont la longueur est inversement proportionnelle à leur similarité. Les arcs entre les documents sont étiquetés avec les chemins absolus issus des EPA qui les contiennent. La construction du graphe tenant compte des similarités d'usage calculées est réalisée par l'application d'un algorithme de placement dirigé par les forces d'attraction-répulsion (Eades, 1984; Fruchterman et Reingold, 1991). Le graphe de la figure II.6.6 a été ainsi obtenu pour l'expérimentation rapportée dans (Cabanac *et al.*, 2007a).

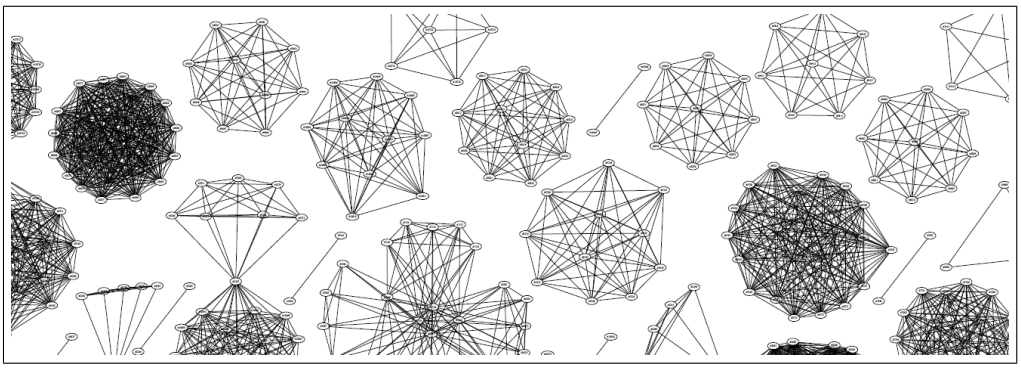


Figure II.6.6 – Graphe de l'usage des documents (Cabanac *et al.*, 2007a).

D'autre part, nous avons retenu deux visualisations pour restituer les thématiques des documents : les cartes auto-organisatrices de Kohonen (2001) et la représentation arborescente qui est davantage familière aux usagers. Comme le montre la figure I.2.2 (p. 20), les cartes auto-organisatrices mettent en évidence les différentes thématiques d'un corpus et leur importance relative en nombre de documents. Par ailleurs, nous proposons de représenter un ensemble de documents en construisant une arborescence de répertoires thématiques sur le même principe que le processus RÉORG (section II.5.4). Ce dernier met en œuvre l'algorithme de classification ascendante hiérarchique de Jardine et van Rijsbergen (1971) qui construit une arborescence binaire dont nous réduisons la profondeur afin de la rendre exploitable, selon le réglage désiré par l'usager. Celui-ci est pris en compte par la procédure de seuillage appliquée (Maarek et Ben-Shaul, 1996). Enfin, les répertoires sont étiquetés avec les termes les plus représentatifs (valeur $tf \cdot idf$ élevée) issus des documents qui les composent.

6.2 Discussion de la proposition

Nous avons pris le parti de concevoir une interface non-intrusive, ne requérant aucune adaptation de la part des membres organisationnels. Ce choix permet notamment de limiter la résistance au changement des usagers. Toutefois, une intervention de ces derniers peut être souhaitable lorsqu'ils désirent indiquer que certains documents doivent rester confidentiels, ou cantonnés à un périmètre d'usagers donné, par exemple. À l'opposé de l'intervention qui réduit la richesse du capital explorable, certains usagers pourraient avoir envie noter les documents qu'ils possèdent, sur une échelle de un à cinq par exemple. Un tel jugement, éventuellement accompagné d'une annotation, serait alors utilisé par l'interface pour mettre en valeur les documents jugés les plus intéressants.

Concernant l'analyse des EPA, il convient de souligner un point important : l'interface proposée ne permet pas l'identification d'expertise à proprement parler. En effet, le fait qu'un individu conserve de nombreux documents sur une thématique donnée ne fait pas de lui un expert. Cette observation permet tout au plus de savoir qu'il s'intéresse à cette thématique-là. À l'opposé, un réel expert peut ne pas avoir besoin de stocker dans son EPA des documents qu'il aurait assimilés ou bien qu'il pourrait facilement retrouver par un autre moyen.

Par ailleurs, une étude des motivations d'archivage de documents papier montre que la construction d'un « héritage documentaire » est la seconde motivation après le fait de pouvoir retrouver un document (Kaye *et al.*, 2006). Si les mêmes motivations s'appliquent aux EPA, l'interface proposée permet effectivement le partage des documents et donc la mise en commun de l'héritage documentaire de chacun, sans pour autant demander aux individus de modifier leur façon de travailler. Il est probable que cette faculté sur le principe du donnant-donnant motive les différents membres organisationnels, qui sauront par la suite que leurs efforts d'organisation de leurs EPA bénéficient également à l'organisation dans sa globalité.

7 Limites et synthèse de la contribution

> "Most people are fools, most authority is malignant, God does not exist, and everything is wrong."
>
> *Ted Nelson (1937 —)*

Ce chapitre clôt la partie « contribution » du mémoire en revenant sur les propositions des deux volets que nous avons exposés :

1. au niveau « microscopique » par la modélisation unifiée des activités documentaires dans une architecture multi-utilisateurs. Les processus associés visent à améliorer chaque activité à partir des autres activités, de façon à aider l'individu grâce aux résultats du groupe, et vice versa. L'objectif ciblé est d'enrichir la relation usagers-documents en favorisant une approche donnant-donnant qui respecte le principe de non-intrusion ;
2. au niveau « macroscopique » par la conception d'une interface d'accès au capital documentaire de l'organisationnel. Celle-ci répond à des besoins organisationnels comme stratégiques en offrant la possibilité de visualiser et d'explorer les EPA, considérés comme de véritables vecteurs d'information à forte valeur ajoutée.

La première section de ce chapitre expose les limites de notre contribution que nous avons identifiées. Puis, nous faisons le bilan de notre contribution dans la seconde section en proposant une synthèse des deux volets.

7.1 Limites de la contribution

Notre proposition repose sur la valorisation des EPA des membres organisationnels. En particulier, des documents annotés et organisés par leurs soins sont automatiquement proposés à leurs collègues. La dissémination de ces documents professionnels stockés dans des espaces personnels ne contrevient pas à la législation française, comme l'atteste une recommandation de la CNIL au

sujet des courriels (Bouchet, 2004, p. 25). Cette recommandation s'appliquerait *a fortiori* aux documents professionnels des EPA, moins sensibles par essence : ils n'incluent pas d'information sur l'expéditeur ou la date d'envoi, contrairement aux courriels, par exemple. Ainsi, tout document organisationnel peut légalement être exploité par notre approche, exceptés ceux contenant une mention personnelle dans le titre — l'organisation pouvant suggérer l'utilisation de mots clés spécifiques tel que « perso ». Toutefois, les travaux concernant la gestion d'information collective passés en revue par Lutters *et al.* (2007) montrent que les individus ont besoin de conserver des informations professionnelles privées (par ex. : choix stratégiques). Ces auteurs rappellent qu'il est essentiel de pouvoir garder le contrôle sur ce qui est échangé, le partage d'information étant flexible, nuancé et contextualisé. Afin de satisfaire ces contraintes et de limiter la résistance des usagers, nous avons proposé dans un premier temps un mécanisme de droits d'accès classique (figure II.2.3). Celui-ci n'est pas complètement adapté aux attentes des usagers qui préféreraient une approche adaptative, ne requérant qu'un minimum d'implication (Lutters *et al.*, 2007). Une piste d'amélioration pourrait s'appuyer sur les actions des usagers pour estimer des seuils de confiance en fonction des échanges observés : l'envoi d'une recommandation de lecture peut indiquer la confiance que l'expéditeur confère au destinataire, concernant ce document et éventuellement ceux qui sont utilisés avec (cf. mesure de similarité sur l'usage, section II.4). De tels indicateurs pourraient alors être utilisés pour adapter dynamiquement les politiques d'accès aux documents.

En relation avec la notion de droits d'accès, Lutters *et al.* (2007) rapportent à quel point le choix du droit d'accès par défaut (public ou privé) peut affecter le modèle de collaboration sous-jacent. En effet, « privé par défaut » peut conduire les usagers à faire de la rétention non intentionnelle d'information, nuisant alors à la performance des processus d'assistance proposés. D'autre part, « public par défaut » peut exacerber la résistance des usagers à adopter le système. Pour ces deux alternatives enfin, il faut être conscient que tout groupe est son propre pire ennemi (Shirky, 2003). Ainsi, il nous semble capital d'identifier et de modérer les passagers clandestins (*free riders*) : les usagers qui parasitent le système en consommant de l'information sans jamais en offrir.

Le dernier point de cette discussion concerne l'approche proposée : la fédération des activités documentaires, qui induit de nombreux avantages mais également des inconvénients critiques au regard de l'usager :

Avantages de la fédération. Elle offre une vision globale de l'ensemble des activités en les décloisonnant. La fédération permet d'utiliser les résultats d'une activité menée par un usager (par ex. : le stockage d'une annotation) pour enrichir l'activité d'un autre membre organisationnel (par ex. : lui recommander ce même document annoté dans son EPA avec le processus RECO). C'est aussi grâce à la fédération que les six processus donnant-donnant sont réalisables. De plus, elle a permis la définition d'un modèle de données unifié, formalisant l'annotation collective et le concept d'EPA, où les usagers stockent et organisent leurs annotations selon leurs besoins. L'interface multi-facettes exploite ainsi cette unification des données pour permettre leur visualisation. En résumé, c'est à travers la fédération que notre approche améliore les pratiques documentaires : annotation collective, recommandations synchrone et asynchrone, réorganisation des EPA, exploration du capital humain et documentaire dans une interface multi-facettes…

Inconvénients de la fédération. Les avantages de la fédération sont acquis au détriment de la liberté de l'usager. Jusqu'alors, il devait péniblement jongler avec une kyrielle d'applications,

en contrepartie il avait toujours le choix d'en utiliser une plutôt qu'une autre. Dans notre approche, c'est un système unique que l'usager manipule. Étant par conséquent limité dans sa liberté de choix, il faut s'attendre à ce qu'il rejette en bloc l'application proposée, même si le fait qu'elle améliore ses activités soit avéré.

La fédération représente le *moyen* retenu pour arriver à la *fin* consistant à améliorer les activités documentaires. Ce choix résulte de l'étude comparative des possibilités de communication entre les applications couvrant le cycle de vie du document. À l'heure actuelle, à notre connaissance, il n'existe aucune opportunité répandue qui constituerait une solution réalisable. Chaque application gère les données des usagers dans son format (propriétaire), certaines en donnent accès aux travers d'API qui sont parfois trop limitées pour implanter les processus présentés dans ce chapitre. Enfin, au niveau technique, les applications sont loin de partager un protocole universel et indépendant du langage d'implantation : RPC, CORBA, RMI-IIOP, DCOM, XPConnect, SOAP, etc. représentent uniquement la partie visible de l'iceberg. En faisant abstraction de ces limites techniques, la fédération dans un seul système pourrait être avantageusement (pour l'usager) remplacée par une approche fondée sur l'interopérabilité entre applications, comme le proposent Chevalier *et al.* (2008) au sujet de l'exploitation des profils d'usagers. Avec la maturité du Web 2.0 et les avancées quotidiennes du Web Sémantique, gageons que les applications intégreront à terme de véritables couches d'interopérabilité. Dès lors, nous pourrons mettre en œuvre les processus présentés dans cet article sans que l'usager n'ait à s'adapter à un système fédéré comme nous sommes obligés de l'imposer à cause des limites actuelles.

7.2 Synthèse de la contribution

Nous avons étudié dans la partie I les activités documentaires réalisées par les travailleurs du savoir au sein de l'organisation, pour lesquelles nous avons identifié trois problématiques. L'individu doit au quotidien maîtriser de nombreux systèmes qui ne communiquent pas entre eux. De ce fait, l'usager doit faire face à une surcharge cognitive ; de plus toute assistance proposée est sous-efficace car produite à partir d'une représentation parcellaire des usagers. Par ailleurs, les informations qu'ils trouvent et structurent dans leurs EPI ne sont pas valorisées au niveau de l'organisation, alors que ce sont des mines d'informations pertinentes pour l'organisation dans son ensemble.

Afin de répondre à ces trois problématiques, nous avons proposé de fédérer les activités documentaires grâce à l'activité transversale d'annotation collective. Ainsi, chaque usager dispose d'un EPA où il stocke et organise ses commentaires associés aux bribes de documents auxquels il désire pouvoir accéder ultérieurement. Le modèle unifié proposé supporte une architecture multi-utilisateurs mettant en œuvre des processus sur le principe du donnant-donnant. En capitalisant sur les activités du groupe qui s'auto-enrichissent, ils assistent chaque individu, notamment en émettant des recommandations synchrones et asynchrones et en adaptant l'affichage des documents annotés. Enfin, nous avons conçu une interface multi-facettes afin d'explorer le capital documentaire organisationnel qui demeurait en sommeil jusqu'à présent, les EPI n'étant souvent accessibles que par leurs propriétaires. Grâce à l'aspect fédéré de notre architecture, cette interface permet la visualisation des EPA des membres organisationnels en fonction de deux dimensions (documents et usagers) et selon deux types de similarités (thématique et usage).

La contribution présentée dans cette deuxième partie du mémoire repose sur différentes propositions originales. Afin de valider ces propositions, nous détaillons dans la partie III les expérimentations relatives à la validation sociale d'annotations collectives (chapitre II.3) et à la mesure de similarité d'usage (chapitre II.4). La mise en œuvre de l'architecture fédérée fait également l'objet de la partie suivante, où nous détaillons son implantation au travers du prototype TafAnnote. Ce dernier est une « preuve de concept » destinée à démontrer la faisabilité de nos propositions.

Troisième partie

Implantation et expérimentation des propositions

1 Introduction

« Ce qui est affirmé sans preuve peut être nié sans preuve. »
Euclide de Mégare (v. 450 av. J.-C. — v. 380 av. J.-C.)

CETTE TROISIÈME PARTIE du mémoire rend compte de la démarche de validation scientifique mise en place pour asseoir la contribution présentée en partie II. Cette contribution est constituée de plusieurs propositions : validation sociale, mesure de similarité d'usage… Afin de la valider incrémentalement, nous considérons en premier lieu les propositions individuellement, avant que d'élaborer une démarche de validation de l'approche dans sa globalité.

1.1 Aperçu des expérimentations réalisées

Notre démarche consiste à valider expérimentalement les six processus intégrés au modèle unifié, socle de l'architecture proposée.

D'une part, les deux processus NAVI et RECO ont été originellement proposés par Chevalier (2002). Ils ont d'ores et déjà fait l'objet d'expérimentations dont nous ne rapportons ici qu'une synthèse permettant d'apprécier leur pertinence :

- **NAVI.** Afin d'évaluer la qualité des recommandations fournies par NAVI durant la navigation, une expérimentation a été conduite à partir des EPI de 14 enseignants-chercheurs, totalisant 4 079 documents (resp. 486 répertoires) avec une moyenne de 291 documents (resp. 34 répertoires) par usager. Cinq individus sélectionnés ont réalisé une navigation prédéfinie par les expérimentateurs. Les résultats de cette expérimentation montrent que la pertinence des recommandations (par rapport au besoin de l'usager) augmente au cours de la navigation ;
- **RECO.** Différentes stratégies de recommandation d'un document dans les EPI ont été évaluées à partir de la collection TREC 2001 OHSUMED/MeSH. Les résultats analysés dans Chevalier (2002) montrent qu'à la fois performance et efficience des recommandations sont améliorées en pratiquant un parcours descendant des hiérarchies de l'arborescence cible.

D'autre part, nous exposons dans cette partie III deux expérimentations originales liées aux

processus ADAPTAFFICHAGE, PROTODOC et NAVI :

1. le chapitre III.2 traite de la validation sociale d'annotations collectives, introduite dans le chapitre II.3. Nous avons constitué un protocole expérimental, mis en œuvre avec la participation en ligne de 121 volontaires. Cette expérimentation vise à déterminer à quel point les algorithmes de validation sociale approximent la perception humaine du consensus dans des débats argumentatifs ;
2. le chapitre III.3 concerne l'expérimentation de la mesure de similarité sur l'usage. Nous montrons que similarités de contenu et d'usage sont complémentaires à partir d'un corpus hiérarchisé de documents médicaux. Leur utilisation conjointe, notamment dans l'interface multi-facettes (chapitre II.6) apporte ainsi un éclairage nouveau sur le capital documentaire organisationnel.

Ces expérimentations sont complétées par la mise en œuvre de la contribution au travers du développement du prototype « TafAnnote ».

1.2 Aperçu du développement réalisé : le prototype TafAnnote

Nous avons modélisé et conçu une architecture multi-utilisateurs dans le but d'améliorer les activités documentaires des individus au sein des organisations. Le chapitre III.4 démontre la faisabilité technique de cette architecture fédérée par le concept d'annotation électronique. En effet, nous avons implanté dans le prototype TafAnnote les différentes propositions au sein d'une application intégrée dans le navigateur Web Mozilla Firefox, le cas d'Internet Explorer étant actuellement à l'étude (Paramelle, 2008). Ce chapitre détaille également les différentes technologies auxquelles nous avons eu recours : XUL, Java, JavaScript, SQL, PL/SQL... Par ailleurs, nous offrons une vision de TafAnnote à l'œuvre, à l'aide d'un scénario réaliste d'utilisation.

2 Expérimentation de la validation sociale d'annotation collective

> "If [a group] has a means of aggregating all [its] different opinions, the group's collective solution may well be smarter than even the smartest person's solution."
>
> *James Surowiecki (2005, p. 75)*

ÉVALUER la proposition exposée dans le chapitre II.3 est une tâche qui s'inscrit dans une démarche globale de recherche scientifique. En effet, nous avons défini la « validation sociale » dans la partie II, nous l'expérimentons dans le présent chapitre et l'intégrons à l'architecture fédérée autour du concept d'annotation dans le chapitre III.4.

Cette expérimentation originale vise à évaluer à quel point les résultats des algorithmes de validation sociale sont proches de la perception humaine du consensus dans des débats argumentatifs. Pour ce faire, nous avons établi un protocole d'expérimentation conforme aux bonnes pratiques observées en psychologie expérimentale (Reips, 2002, 2007). Concrètement, la plate-forme d'expérimentation que nous avons développée permet à des individus de prendre part à cette expérimentation en ligne, à partir de la page « `http://www.irit.fr/~Guillaume.Cabanac/expe` ».

Contrairement à une expérimentation réalisée sous le contrôle des expérimentateurs dans un laboratoire, notre approche est « écologique » dans le sens où les sujets participent dans leur environnement habituel. Un autre avantage réside dans le fait que les participants recrutés sont diversifiés et volontaires, car sollicités à partir de listes de diffusion nationales et internationales.

Cette expérimentation débutée en avril 2007 a recueilli 180 inscriptions. Parmi ces participants, 121 ont commencé l'évaluation qui comprend 13 étapes et 53 l'ont terminée. Ce chapitre expose le protocole d'expérimentation que nous avons élaboré, il a fait l'objet d'une publication nationale (Cabanac, 2008a). Puis, nous analysons les données recueillies à partir de la participation des individus.

2.1 Méthodologie de l'expérimentation

Nous avons conçu une expérimentation originale afin d'évaluer les algorithmes de validation sociale proposés dans le chapitre II.3. Elle consiste en la comparaison de leurs résultats avec la perception humaine du consensus. Pour ce faire, nous avons dû recruter des participants pour évaluer et synthétiser les arguments des débats rassemblés à cet effet. Dans l'optique de conduire une expérimentation généralisable, nous avons opté pour la sollicitation d'un large spectre de participants volontaires, recrutés mondialement. Ils ont pris part à l'expérimentation en ligne, grâce à la plate-forme que nous avons développé à cet effet.

2.1.1 Constitution du corpus d'expérimentation : 13 débats argumentatifs

Notre expérimentation a nécessité de la part des individus qu'ils évaluent des débats argumentatifs. Or, des travaux récents sur l'annotation et les fils de discussion ont noté l'absence de tels débats (Agosti et Ferro, 2006, 2007; Frommholz et Fuhr, 2006). De fait, l'évaluation des propositions dans ce domaine est une tâche difficile. Nous avons alors pris le parti de constituer notre propre corpus. Dans le but de proposer des débats réalistes et variés, nous avons exploité les trois ressources présentées dans le tableau III.2.1. Ce sont des cartes d'argumentation (Twardy, 2004) publiées par des universitaires (groupes A et B), ainsi que des dialogues argumentatifs en français (groupe C) publiés par Cayrol et Lagasquie-Schiex (2004).

Groupe	Numéro de débat	Description	Auteurs
A	1, 3–5, 7, 10	Cartes d'argumentation [1]	Université de Melbourne (AU)
B	2, 9, 11, 12	Cartes d'argumentation [2]	Université d'Ohio (US)
C	6, 8, 13	Dialogues argumentatifs	Université de Toulouse (FR)

Tableau III.2.1 – Origines des 13 débats argumentatifs constituant le corpus expérimental.

Après avoir traduit les ressources françaises en Anglais, nous avons obtenu 13 débats en Anglais comportant 222 arguments, leur contenu peut être consulté à partir de la page Web « `http://www.irit.fr/~Guillaume.Cabanac/expe/corpus` ». Chaque argument est une courte phrase exprimant une seule opinion. Comme le montre le tableau III.2.2, leurs caractéristiques sont variées. Il en est de même de leurs thématiques qui couvrent des sujets tels que la consommation de tabac, le réchauffement climatique, la discussion d'alternatives à une opération chirurgicale, l'orientation d'étudiants, l'interprétation de la Bible…

Caractéristique	Minimum	Maximum	Moyenne	Écart type
Nombre d'arguments	5	34	17,1	8,1
Profondeur de l'arbre représentant un débat	3	7	4,2	1,3
Largeur de l'arbre représentant un débat	3	15	7,9	3,2

Tableau III.2.2 – Caractéristiques du corpus constitué comprenant 13 débats.

1. `http://austhink.com/reason/tutorials`
2. `http://jostwald.com/ArgumentMapping`

2.1.2 Tâches des participants : étiquetage et synthèse des opinions

Les participants à l'expérimentation ont réalisé deux tâches consécutives pour chacun des 13 débats comprenant 222 arguments en tout. Afin de limiter un éventuel abandon dès le commencement de l'expérimentation, nous avons ordonné les débats selon leur difficulté croissante, que nous avons subjectivement estimée à partir de leurs caractéristiques et thématiques. Le tableau III.2.1 montre l'ordre des débats en fonction de leur origine. Les deux tâches assignées aux participants, dénotées ❶ et ❷, sont définies ci-dessous.

– Pour la tâche ❶, les participants évaluent la sémantique des liens qui relient les arguments dans chaque débat. Ces liens sont représentés par $a_i \leftarrow a_j$ dans la figure III.2.1.

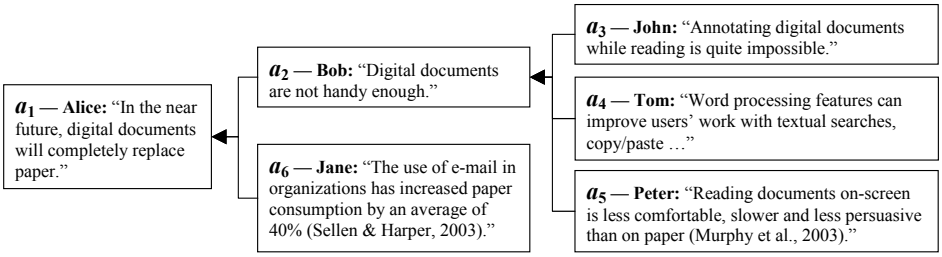


Figure III.2.1 – Exemple d'un débat argumentatif comprenant six arguments.

  Cet étiquetage consiste en l'identification du type opinion et des types commentaire existant entre l'argument a_j et son parent direct a_i dans le débat. En faisant référence au tableau II.2.1 (p. 49), trois types opinion (réfutation $\mathscr{R}$, neutre $\mathscr{N}$ et confirmation $\mathscr{C}$) et trois types commentaire (question $\mathscr{Q}$, modification $\mathscr{M}$, exemple $\mathscr{E}$) sont proposés. La combinaison des types opinion et commentaire permet aux participants de retranscrire la force de l'argument qu'ils évaluent : une confirmation appuyée par un exemple étant plus justifiée qu'une confirmation seule, par exemple. L'identification des opinions doit être objective : nous demandons aux participants de prendre en compte le contenu des arguments uniquement, car leurs jugements personnels sont indésirables. Dans la figure III.2.1, on peut identifier une confirmation pour $a_2 \leftarrow a_5$ ainsi qu'une réfutation avec un exemple pour $a_2 \leftarrow a_4$.
– Les participants passent à la tâche ❷ une fois que les liens reliant les arguments du débat courant sont tous étiquetés. Cette tâche consiste à synthétiser les opinions du débat mentalement. Pour un argument a_i ayant un ensemble de réponses $\{a_j\}$, les participants doivent associer à a_i une valeur de synthèse sur une échelle à 10 graduations. Elle varie entre « réfuté » et « confirmé » en passant par « neutre » (figure III.2.6) et permet aux participants de spécifier l'opinion globale des réponses a_j. En réalisant cette tâche récursivement, les participants arrivent à synthétiser l'opinion globale du groupe, qu'ils associent à l'argument racine du débat. Par exemple, sur la figure III.2.1 on peut estimer que a_2 est confirmée à 66 % en synthétisant les opinions de a_3, a_4 et a_5. Une telle évaluation serait rationnelle car elle correspond à la proportion $\frac{\text{nombre de confirmations}}{\text{nombre de refutations}}$. Notons que cette stratégie de synthèse spécifique n'est pas indiquée aux participants de façon à ne pas les influencer.

Les participants prennent part à l'expérimentation grâce à un logiciel spécifique permettant l'affichage de débats et le stockage des contributions des participants : les opinions identifiées ❶ et les valeurs de synthèse estimées ❷. La section suivante détaille la plate-forme logicielle que nous avons développée à cet effet.

2.1.3 Plate-forme pour l'expérimentation écologique en ligne

La méthodologie détaillée dans la section précédente a été mise en œuvre au sein d'une plate-forme d'expérimentation en ligne. Celle-ci est conforme aux standards de l'expérimentation en ligne établis par Reips (2002, 2007). Elle fédère trois composants :

1. la page Web « `http://www.irit.fr/~Guillaume.Cabanac/expe` » qui explique l'objectif de l'expérimentation et comment y prendre part ;
2. le logiciel associé pour afficher les débats à destination des participants ;
3. une base de données dédiée pour stocker les contributions recueillies.

La page Web permet à un participant de lancer l'application Java-Swing sur son ordinateur grâce à la technologie de déploiement Java Web Start. Suite à son inscription, un tutoriel décrit le travail à réaliser pour chacun des 13 débats, c'est-à-dire les tâches ❶ et ❷. Enfin, l'application affiche l'interface graphique présentant le premier débat à évaluer (figure III.2.2).

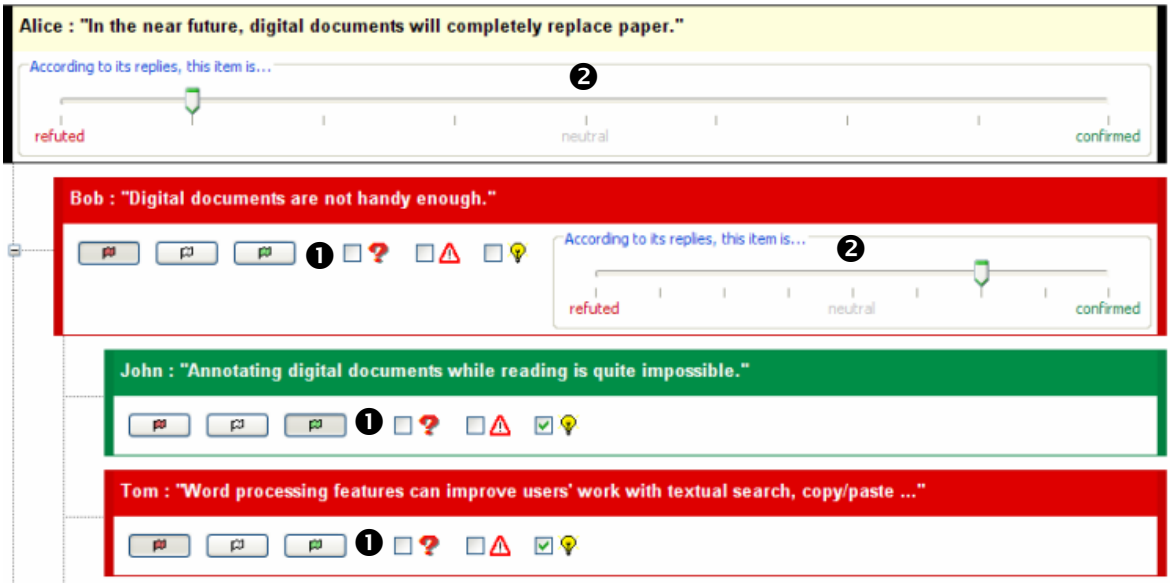


Figure III.2.2 – Capture d'écran de l'application affichant le premier débat à évaluer.

La figure III.2.2 illustre comment le débat de la figure III.2.1 est restitué au participant. Notons que cette figure ne montre que les arguments de a_1 à a_4 pour des raisons de concision. Afin de ne pas perturber les participants, nous avons opté pour une visualisation hiérarchique à laquelle ils sont certainement habitués car elle est couramment employée pour afficher les arborescences des fichiers.

Chaque argument est affiché dans une boîte horizontale, dont le fond est initialement coloré en jaune clair. Les participants réalisent la tâche ❶ à l'aide des boutons exclusifs représentant des drapeaux de couleur et des cases à cocher :

– chacun des types opinion est associé à une couleur métaphorique : réfutation ↔ rouge, neutre ↔ blanc et confirmation ↔ vert. Dans un souci d'améliorer la lisibilité de l'interface, la sélection d'un bouton modifie le fond de la boîte de l'argument en l'affichant dans la couleur associée au bouton, cf. figure III.2.2 ;

– les types commentaire sont sélectionnables au travers des cases à cocher associées à leurs icônes : question ↔ point d'interrogation, modification ↔ panneau de signalisation et exemple ↔ ampoule.

Lorsque le type opinion de chaque argument est sélectionné (obligatoire), les participants passent à la tâche ❷ en synthétisant les opinions des réponses (par ex : celles de John, Tom et Peter) au niveau de leur argument père (celle de Bob) sur la réglette associée. Ce composant graphique figure sur tous les arguments qui possèdent des réponses. C'est une échelle à 10 graduations qui varie de « réfuté » à « confirmé » en passant par « neutre » (figure III.2.6). La réglette associée à l'argument racine du débat reflète l'opinion globale du groupe, c'est-à-dire sa validité sociale estimée mentalement. Tant que les deux tâches ne sont pas terminées pour le débat en cours, l'application empêche le participant d'évaluer les débats suivants.

Une base de données dédiée stocke le contenu des 13 débats, ainsi que les données concernant les participants et leurs participations. La figure III.2.3 représente le modèle conceptuel de cette base de données implantée avec le SGBDR Oracle $10g2$. Le corpus est constitué à partir des Ressources dont ont été extraits les 13 Débats que nous avons ordonnés. Chaque débat est initié par un Argument racine qui est discuté par des arguments de type réponse. Pour conserver les informations relatives aux participants, nous avons modélisé la classe Participant qui contient son identité, son courriel, son sexe, son âge, son métier, la date de son inscription, son pays et ses langues maternelles (une seule étant requise). En plus de ces données, l'interface demande aux participants de spécifier leur niveau d'Anglais et à quel point ils sont familiarisés avec Internet et les forums. Enfin, ils peuvent demander à recevoir des informations liées à l'expérimentation (reqInfos) et écrire un commentaire. L'ensemble de ces informations sont demandées au travers du formulaire d'inscription (figure III.2.4), toutefois seules l'identité et l'adresse de courriel doivent être obligatoirement fournies.

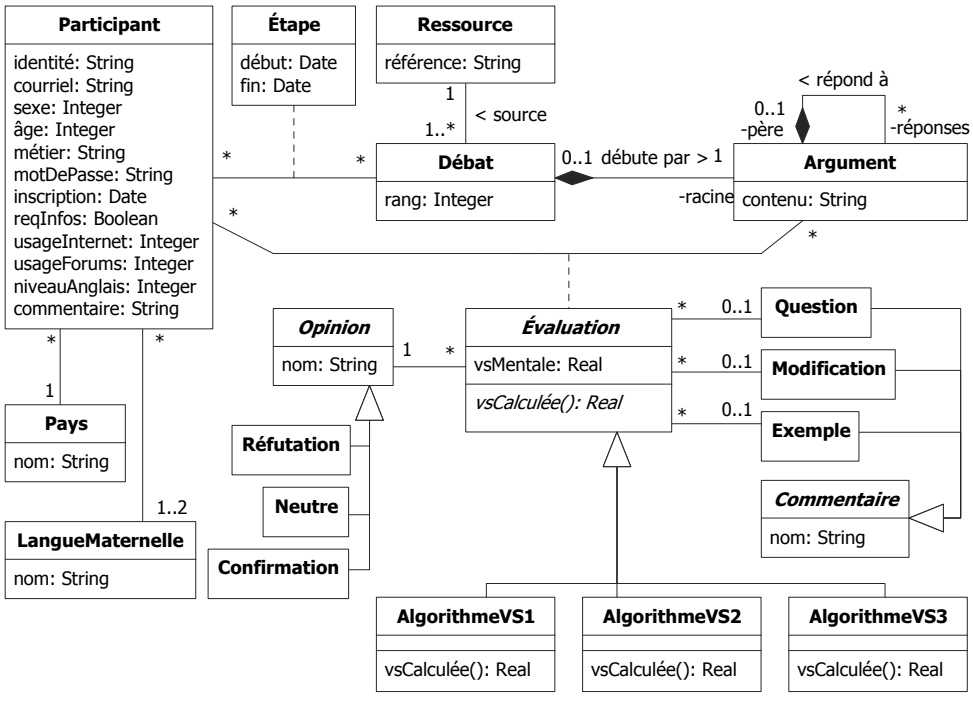


Figure III.2.3 – Diagramme des classes UML de la plate-forme d'expérimentation.

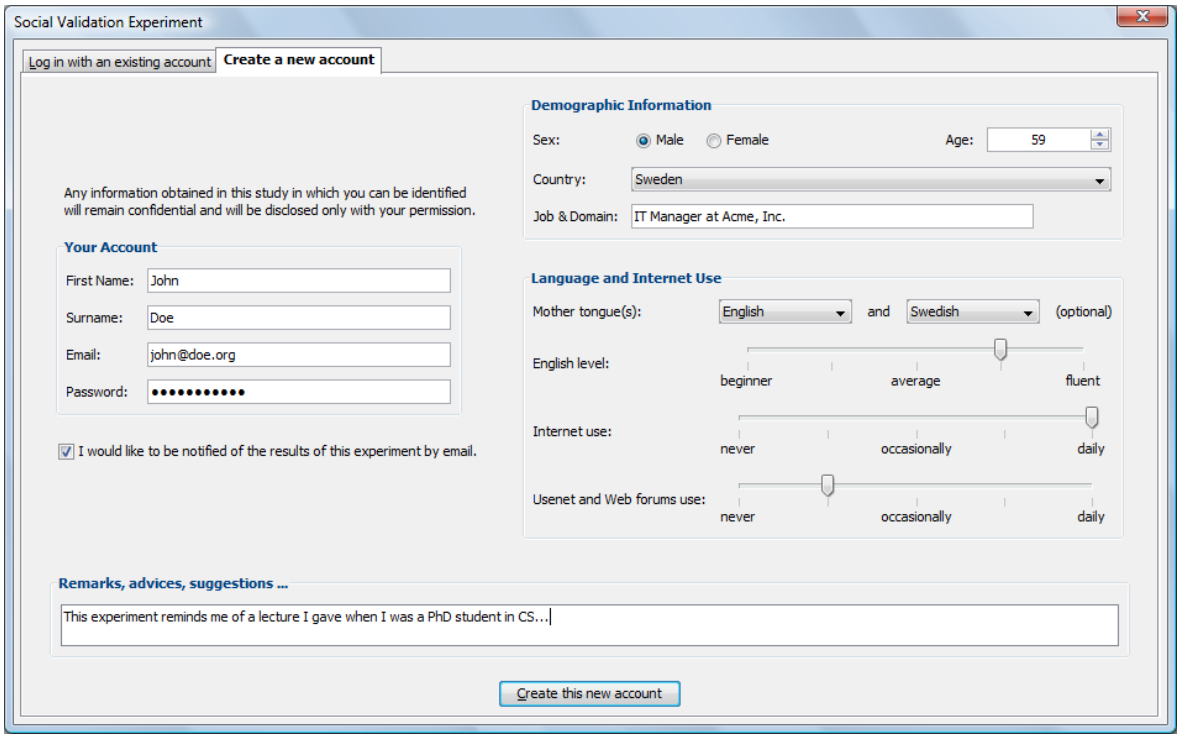


Figure III.2.4 – Capture d'écran du formulaire d'inscription à l'expérimentation.

L'expérimentation consiste à évaluer 13 débats, nous l'avons donc divisée en 13 Étapes indépendantes. Pour des raisons pratiques, les participants peuvent stopper l'expérimentation à tout moment, puis la reprendre plus tard. Une étape consiste à fournir une *Évaluation* pour chaque argument du débat en cours : réaliser la tâche ❶ (classes *Opinion* et *Commentaire*) et la tâche ❷ (attribut vsMentale, pour les arguments pères uniquement). Contrairement à vsMentale qui représente la perception humaine, la fonction *vsCalculée* calcule la validation sociale selon l'algorithme étudié (classes concrètes AlgorithmeVSi, détaillées dans la section III.2.2.3). Notons que ces algorithmes prennent en compte les types opinion et commentaire identifiés par le participant. Par conséquent, la perception humaine est comparée aux résultats des algorithmes, sur la base des mêmes types que le participant p. Par exemple, on compare la valeur de la réglette que le participant p a positionné pour l'argument de Bob (figure III.2.2) avec le résultat d'un algorithme de validation sociale exécuté à partir des types commentaire et opinion identifiés par le même participant p.

2.1.4 Recrutement des participants : appel à participation international

Le recrutement des participants est une étape clé de toute expérimentation. Parmi d'autres, Wolfe et Neuwirth (2001) dénoncent le recrutement de participants non indépendants ou *a priori* du même avis (les collègues ou les étudiants des expérimentateurs par exemple) car cela entraîne des biais de comportements. Désirant encourager l'indépendance et la diversité des participants, nous avons posté un appel à participation sur le Web. À partir d'avril 2007 nous l'avons progressivement envoyé sur des listes de diffusion françaises et internationales concernant l'informatique : bulle-i3, liste-egc, rtp-doc, chi-student et chi-web de l'ACM, www-annotation du W3C, semanticweb,

webir... En septembre 2007, nous avons ajouté notre expérimentation à des sites dédiés à la psychologie expérimentale : *Web Experiment List*[3] (Reips et Lengler, 2005) et *Psychological Research on the Net*[4]. Il est important de noter que les participants ont pris part à l'expérimentation sur la base du volontariat car aucune rétribution n'est proposée.

2.2 Résultats : analyse des évaluations des 121 participants

Cette section analyse les données acquises pendant la durée de l'expérimentation (16 mois). Pour des raisons de reproductibilité, ces données sont anonymisées et accessibles au format XML à partir de l'adresse « `http://www.irit.fr/~Guillaume.Cabanac/expe/data.xml` ». Nous rapportons dans cette section les analyses statistiques que nous avons réalisées pour mesurer à quel point les algorithmes de validation sociale sont proches de la perception humaine du consensus.

2.2.1 Analyse quantitative des 121 participations

Au mois d'août 2008, 180 personnes ont lancé le logiciel d'expérimentation et se sont inscrites. Par la suite, nous appelons « participants » les 121 personnes qui l'ont réellement commencée, en évaluant au moins un débat. Pour des raisons de respect de la vie privée, le remplissage du formulaire présenté en figure III.2.4 n'est pas obligatoire. Par conséquent, seulement 55 participants parmi les 121 au total ont fourni des données personnelles. Sur la base de ces 55 formulaires complétés, nous constatons que les participants résident dans 13 pays différents. Le tableau III.2.3 montre la proportion de leurs origines géographiques. Les participants résidant en France sont majoritaires, en partie parce que l'appel à participation a tout d'abord été envoyé sur des listes de diffusion françaises.

Origine	France	Amérique (Nord et Sud)	Europe	Autres
Proportion	67 %	13 %	10 %	10 %

Tableau III.2.3 – Origine des 55 participants qui ont rempli le formulaire d'inscription.

Ces données relatives aux origines des participants sont consistantes avec celles présentées dans le tableau III.2.4, montrant que la langue française est la langue maternelle de la majorité des participants.

Langue maternelle (code ISO)	fr	en	ar	de	pt	el	oc	ro	tr
Proportion (%)	62,1	19,7	6,1	3,0	3,0	1,5	1,5	1,5	1,5

Tableau III.2.4 – Langues maternelles des 55 participants ayant rempli le formulaire.

Le tableau III.2.5 illustre les caractéristiques des participants recueillies grâce au formulaire d'inscription (lignes 1 à 4) ou calculées à partir des données obtenues (dernière ligne). Les échelles ordinales à 5 graduations ont été encodées par un entier dans l'intervalle $[\![1;5]\!]$. Le participant typique est un homme (60 %) trentenaire déclarant un bon niveau d'Anglais. Les participants sont

3. `http://genpsylab-wexlist.unizh.ch`
4. `http://psych.hanover.edu/research/exponnet.html`

très familiers avec Internet car ils l'utilisent quotidiennement, bien que moins à l'aise avec les forums Usenet et forums car ils déclarent ne les utiliser qu'occasionnellement.

Variable	Minimum	Maximum	Moyenne	Écart type
Âge	20	61	32,1	9,4
Niveau d'Anglais	1	5	3,8	1,1
Utilisation d'Internet	3	5	4,9	0,4
Utilisation de Usenet et forums Web	1	5	3,2	1,1
Nombre d'interruptions [5]	0	12	4,5	3,9

Tableau III.2.5 – Données quantitatives des 55 participants ayant pris part à l'expérimentation.

Nous avons conçu l'expérimentation de façon à ce qu'elle puisse être stoppée et reprise à tout instant ; en moyenne un participant la termine avec 4,5 interruptions (tableau III.2.5). Comme seuls 53 participants sur 121 l'ont terminée, le nombre de participants par étape décroît en fonction de son rang (figure III.2.5). Ceci correspond à un taux d'abandon de 56 % qui est sensiblement supérieur aux 45 % observés par Reips (2007, p. 387) dans le cadre des expérimentations non rémunérées. Notons que les participations des personnes qui ne terminent pas l'expérimentation sont tout de même exploitables car elles ont au moins achevé de une à douze étapes.

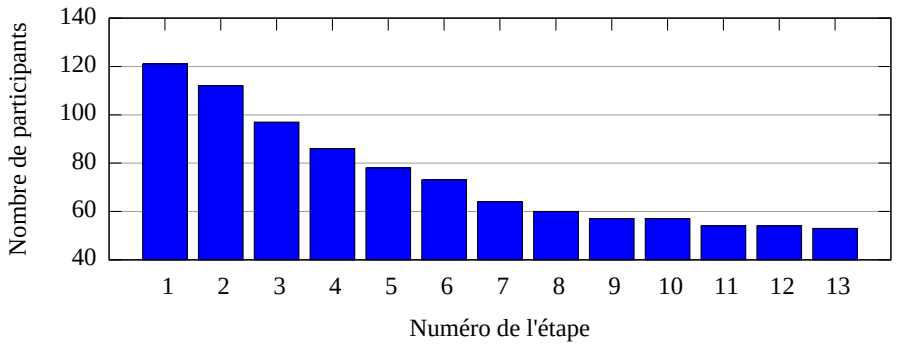


Figure III.2.5 – Courbe d'abandon montrant le nombre de participants pour chaque étape.

Sur la base des 966 étapes terminées, évaluer un débat prend 6,5 minutes en moyenne, soit 84,5 minutes pour l'expérimentation complète. Il est important de noter que la durée moyenne d'évaluation d'un débat est mesurée à partir de son affichage jusqu'à ce que le participant passe au suivant. De part la caractéristique « écologique » de l'expérimentation (sur le terrain), les participants n'étaient pas sous la supervision des expérimentateurs. De ce fait, nous n'avons pas pu relever les pauses et autres activités des participants : répondre au téléphone, par exemple. Ceci peut contribuer à expliquer le fort écart type (σ = 10,6 minutes) indiquant une importante variabilité entre les participants. Ainsi, les estimations de durées sont pessimistes car elles surestiment le temps qu'un participant consacre à l'évaluation d'un débat. En ignorant l'évaluation la plus longue pour chaque participant, la durée moyenne d'évaluation d'un débat chute à 5,2 minutes, avec une variabilité bien plus faible (σ = 4,6 minutes). Cette situation est davantage proche des durées qu'ont rapportées les participants que nous avons pu questionner *a posteriori*.

5. Statistiques concernant les 53 qui ont fini l'expérimentation.

2.2.2 Analyse qualitative des 121 participations

Mener une expérimentation en ligne apporte de nombreux avantages : les participants peuvent provenir du monde entier, peuvent être diversifiés et indépendants. Par contre, ce type d'expérimentation présente aussi des inconvénients par rapport aux expérimentations en laboratoire sous la supervision des expérimentateurs : certains participants ignorent les instructions, ne les comprennent pas, se lassent de la tâche et ne répondent pas consciencieusement ou font même des erreurs délibérées. Afin de prendre en compte ces éléments, nous avons défini des indicateurs pour vérifier la qualité des participations recueillies.

Durant la tâche ❶, les participants identifient les opinions des arguments. C'est une tâche subjective, en particulier lorsqu'on doit choisir entre les types $\mathscr{R}$ et $\mathscr{N}$, ou entre $\mathscr{C}$ et $\mathscr{N}$. De fait, deux personnes peuvent identifier des types complètement opposés. Toutefois, cela ne pose pas problème dans le contexte de cette expérimentation, car nous comparons la synthèse mentale de chaque participant (qu'il établit sur la base des types qu'il identifie) avec les résultats des algorithmes exécutés sur la base de ces mêmes types. En résumé, l'accord interpersonnel relatif à l'identification des types n'est pas requis pour comparer la perception humaine avec la validité sociale calculée par les algorithmes. Par conséquent, aucun indicateur de qualité n'est nécessaire pour la tâche ❶.

Par contre, les synthèses d'opinions recueillies durant la tâche ❷ peuvent être erronées et doivent donc être vérifiées. En effet, la prise en compte des participants qui n'ont pas compris les instructions ou les éléments de l'interface, ou bien qui ont positionné les réglettes quasi-aléatoirement pourrait biaiser les analyses. C'est pourquoi nous avons défini les quatre indicateurs d'erreur présentés dans le tableau III.2.6. Nous avons observé 21 % de synthèses d'opinions irrationnelles parmi les 5 647 évaluations (classe Évaluation, figure III.2.3). Le tableau III.2.6 détaille la proportion de chaque indicateur, par exemple l'indicateur ① identifie le cas d'un argument synthétisé comme réfuté bien qu'aucune de ses réponses ne le réfute ($\neg\mathscr{R}$). L'indicateur ② est la contrepartie de l'indicateur ①. La situation irrationnelle la plus fréquente (35 %) est reflétée par l'indicateur ③ lorsque des confirmations uniquement ($\neg\mathscr{N} \wedge \neg\mathscr{R}$) ou des réfutations uniquement ($\neg\mathscr{N} \wedge \neg\mathscr{C}$) sont synthétisées en tant que neutre. Enfin, l'indicateur ④ repère les arguments neutres ($\mathscr{N} \wedge \neg(\mathscr{R} \vee \mathscr{C})$) synthétisés en tant que confirmation ou réfutation.

Indicateur d'erreur	①	②	③	④
Synthèse de l'argument père	réfuté	confirmé	neutre	confirmé ou réfuté
Opinions des arguments fils	$\neg\mathscr{R}$	$\neg\mathscr{C}$	$\neg(\mathscr{R} \wedge \mathscr{C} \vee \mathscr{N})$	$\mathscr{N} \wedge \neg(\mathscr{R} \vee \mathscr{C})$
Proportion observée	28 %	24 %	35 %	12 %

Tableau III.2.6 – Synthèses d'opinions irrationnelles observées pour 21 % des 5 647 évaluations.

Le taux d'erreur moyen par participant est de 7 % (σ = 7 %), le participant qui s'est le plus trompé atteignant un taux d'erreur de 26 %. En posant l'hypothèse qu'un taux d'erreur au-delà de 20 % révèle une incompréhension de la part du participant, nous avons écarté les participations des 7 participants correspondants. Certains avaient pourtant terminé l'expérimentation… Sur la base des participations restantes, la prochaine section compare la perception humaine du consensus avec les trois algorithmes de validation sociale proposés.

2.2.3 Les algorithmes de validation sociale approximent-ils la perception humaine ?

L'analyse des 121 participations, en excluant les 7 irrationnelles, nous a permis d'évaluer à quel point les algorithmes de validation sociale approximent la perception humaine du consensus. Cette section compare les trois algorithmes avec les valeurs de synthèse d'opinions que les participants ont fournies. Par la suite, *ph* désigne la perception humaine, c'est-à-dire les valeurs de synthèse d'opinions fournies durant la tâche ❷. De plus, vs_1 fera référence au premier algorithme (section II.3.1.2, p. 57). L'algorithme de Cayrol et Lagasquie-Schiex (2005a) sera noté vs_2 (section II.3.1.3, p. 59) et notre extension de cet algorithme sera noté vs_3 (définition 4, p. 61). Notons que vs_1 et vs_3 sont en partie basés sur des informations inexistantes dans le corpus constitué, telles que l'expertise de l'annotateur et le nombre de références. Ainsi, ces aspects ne peuvent pas être évalués dans cette expérimentation.

Dans le but de comparer ph avec vs_i, nous avons dû choisir entre la prise en compte de chaque argument ayant des réponses (a_1 et a_2 dans la figure III.2.1, par exemple) et la prise en compte des arguments racine uniquement (par ex. : a_1). La première stratégie favorise les arguments non-racine qui surpassent clairement en nombre les arguments racine. Pourtant, la synthèse mentale des opinions pour un argument racine est plus difficile à réaliser que pour un argument non-racine, car cela requiert de synthétiser jusqu'à 34 arguments (tableau III.2.2). De ce fait, la comparaison obtenue serait biaisée car elle ne représenterait pas correctement la capacité des algorithmes à synthétiser un débat complet (à partir de la racine). C'est pourquoi nous avons opté pour la seconde alternative, qui est plus stricte. En excluant les participations écartées précédemment, nous avons extrait de la base de données 887 tuples tels que $\langle p, a, ph(p,a), vs_1(p,a), vs_2(p,a), vs_3(p,a)\rangle$ où :

- p est un participant ;
- a est un argument ;
- $ph(p,a)$ est la synthèse des opinions assignées à l'argument a par le participant p ;
- $vs_i(p,a)$ est le résultat de l'algorithme vs_i exécuté avec les opinions identifiées par le participant p pour l'argument a.

Nous avons recouru à une approche d'estimation de paramètres pour l'initialisation des algorithmes vs_i. Les valeurs ph et vs_i sont définies dans l'intervalle $[-1;1]$ représenté par l'échelle à 10 graduations de la figure III.2.6. Pour des raisons pratiques, le composant graphique simulant la réglette limite la sélection d'une valeur parmi les 10 graduations, d'où les valeurs discontinues observées pour *ph*. À l'opposé, les algorithmes vs_i calculent des valeurs continues. Cette différence entre les valeurs ph et vs_i sont considérées dans la suite de cette section.

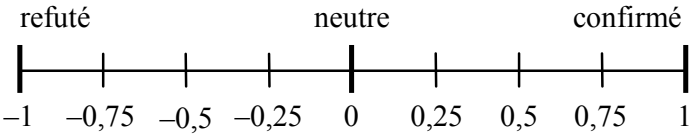


Figure III.2.6 – Encodage des synthèses d'opinions à partir d'une réglette à 10 graduations.

Pour comparer les algorithmes avec la perception humaine, nous avons tout d'abord calculé leurs différences $vs_i - ph$ et tracé la distribution des écarts obtenus (figure III.2.7). Nous posons l'hypothèse suivante : vs_i correspond exactement (erreur $x = 0$) à une position ph donnée si vs_i est davantage proche de la position ph que de n'importe quelle autre position. Étant donnée

la différence de 0,25 entre deux positions voisines, une erreur $x = 0$ est donc obtenue lorsque $|vs_i - ph| \leqslant \frac{0{,}25}{2} = 0{,}125$. Sur cette base, la figure III.2.7 représente la proportion y d'une différence donnée $x \pm 0{,}125$ pour les trois algorithmes : $y = p(x - 0{,}125 \leqslant vs_i - ph \leqslant x + 0{,}125)$. Idéalement, les algorithmes égaleraient la perception humaine, c'est-à-dire que $y(0) = 100\,\%$. En réalité, ils égalent la perception humaine dans approximativement $y(0) = 25\,\%$ des cas. De plus, la figure III.2.7 montre que les trois algorithmes fournissent des résultats similaires.

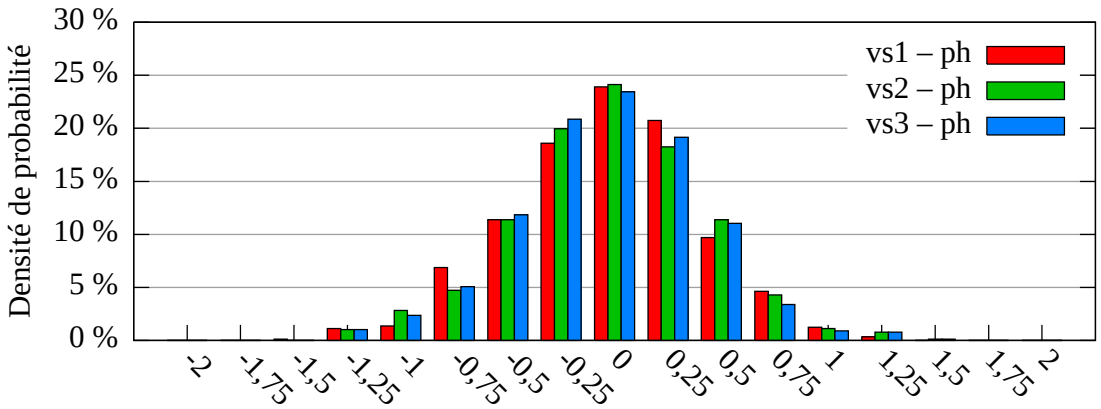


Figure III.2.7 – Différences entre les trois algorithmes vs_i et la perception humaine ph.

2.2.3.1 Algorithmes *versus* perception humaine : tests statistiques d'hypothèses

Afin de comparer les couples (ph, vs_i) à l'aide de tests statistiques d'hypothèses (Savy, 2006, ch. 4), nous avons tout d'abord vérifié la normalité de la distribution des erreurs $(vs_i - ph)$. En effet, la sélection du test statistique approprié (paramétrique ou non-paramétrique) dépend de cette distribution des erreurs. Nous avons recouru au test de Shapiro et Wilk (1965) pour vérifier que les trois séries (ph, vs_i) sont conformes à une distribution Normale. Le tableau III.2.7 montre les p-valeurs de significativité correspondant à l'hypothèse de normalité des séries (ph, vs_i).

	(ph, vs_1)	(ph, vs_2)	(ph, vs_3)
Valeurs p de significativité du test de normalité	0,0605	0,0007	0,0150

Tableau III.2.7 – Significativité du test de normalité de Shapiro-Wilk pour les couples (ph, vs_i).

Les valeurs $p\left(ph, vs_{\{2,3\}}\right) < \alpha$ rejettent cette hypothèse de normalité, où $\alpha = 0{,}05$ est le seuil communément admis (Hull, 1993). Par conséquent, les analyses doivent être réalisées avec des tests non-paramétriques.

Pour approfondir notre étude, nous avons utilisé le test *signed-rank* de Wilcoxon (1945) sur échantillons appariés. Ce test est la contrepartie non-paramétrique du t-test paramétrique de Student. Il permet de calculer la valeur de significativité $p \in [0;1]$ qui estime la probabilité que la différence entre deux méthodes est due au hasard [6]. En fait, on peut déduire que les deux méthodes sont statistiquement différentes (et que ce n'est pas le fruit du hasard) lorsque $p < \alpha$, où $\alpha = 0{,}05$ est communément utilisé (Hull, 1993). Autrement dit, plus $p \to 0$, plus il est probable que les deux méthodes soient différentes.

De plus, nous avons utilisé un second coefficient statistique, le coefficient de Pearson $r \in [-1;1]$ qui évalue le degré de corrélation linéaire entre deux méthodes (Savy, 2006, ch. 8). Une valeur $r \to 1$ reflète une relation linéaire positive ($y = x$) alors qu'une valeur $r \to -1$ reflète une relation linéaire négative $\left(y = \frac{1}{x}\right)$. Ainsi, plus $r \to 1$, plus les deux méthodes sont similaires.

Le tableau III.2.8 rapporte les valeurs de significativité p et de corrélation r. La première ligne concerne l'ensemble des débats (1 à 13). Les deux valeurs $p\left(ph, vs_{\{2,3\}}\right) < \alpha$ montrent que les différences entre les algorithmes vs_i et ph sont statistiquement significatives. De plus, comme $p(ph, vs_1) > p\left(ph, vs_{\{2,3\}}\right)$, l'algorithme vs_1 est moins différent de ph que vs_2 et vs_3 ne le sont. Puis, nous avons calculé la corrélation $r \approx 0{,}5$ qui indique une corrélation moyenne entre vs_i et ph. Notons que les différences entre les trois algorithmes ne semblent pas significatives car leurs résultats sont très proches les uns des autres.

Débats	Significativité p du test de Wilcoxon			Coefficient de corrélation r		
	(ph, vs_1)	(ph, vs_2)	(ph, vs_3)	(ph, vs_1)	(ph, vs_2)	(ph, vs_3)
1 à 13	0,0667	0,0475	0,0011	0,4859	0,4714	0,4905
2 à 13	0,0288	0,0626	0,0021	0,4954	0,4766	0,4971
3 à 13	0,0381	0,1970	0,0083	0,5161	0,4912	0,5146
4 à 13	0,1610	0,7319	0,0732	0,5316	0,5032	0,5281

Tableau III.2.8 – Significativité p du test de Wilcoxon et corrélation r pour les paires (ph, vs_i).

Certains participants nous ont averti que leurs évaluations au début de l'expérimentation devraient être éliminées car ils ont réalisé leurs erreurs après coup. En fait, l'application ne permet pas de modifier les évaluations des débats complétés, c'est pourquoi ils n'ont pas pu les modifier. De ce fait, nous avons voulu évaluer l'impact des premières étapes « d'échauffement ». À cet effet, les lignes 2 à 4 du tableau III.2.8 éliminent progressivement les étapes de 1 à 3. Il apparaît que c'est pour les débats de 4 à 13 que les valeurs p sont les plus importantes, dépassant les valeurs p calculées pour les débats de 1 à 13, de 2 à 13 et de 3 à 13. Cette observation semble confirmer le fait que les premières évaluations des participants sont erronées. Pour pousser l'analyse, la figure III.2.8 montre le graphique obtenu à partir des couples (vs_i, ph) pour les débats de 1 à 13.

Les points obtenus avec la série ph forment des lignes horizontales, car nous avons obtenu ces données à partir des réglettes similaires à la figure III.2.6. Pour chaque algorithme, nous avons calculé les régressions linéaires et obtenu les droites d'équation $\hat{y} = ax + b$. En présence d'un algo-

6. "The preliminary assumption, or null hypothesis H_0, will be that all the retrieval methods being tested are equivalent in terms of performance. The significance test will attempt to disprove this hypothesis by detecting a p-value, a measurement of the probability that the observed difference could have occurred by chance. Prior to the experiment, a significance level α is chosen, and if the p-value is less than α, one can conclude that the search methods are significantly different. A smaller value of α (which is often set to .05) implies a more conservative test. Alternatively, the p-value can merely be viewed as an estimate of the likelihood than [sic] two methods are different." (Hull, 1993)

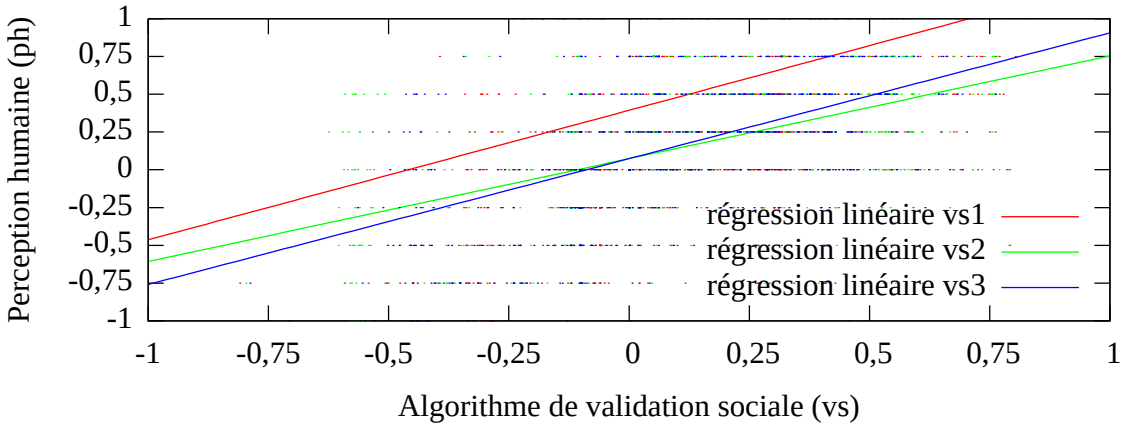


Figure III.2.8 – Couples (vs_i, ph) et leurs équations de droites correspondantes.

rithme qui fournirait exactement les mêmes résultats que la perception humaine, on obtiendrait $\hat{y} = x$. D'après la figure III.2.8 obtenue, les algorithmes vs_2 et vs_3 obtiennent des équations similaires, qui sont davantage proches de la droite idéale d'équation $\hat{y} = x$ que de vs_1.

Comme les algorithmes vs_2 et vs_3 semblent fournir des résultats très similaires d'après la figure III.2.8, nous avons vérifié si chaque algorithme est statistiquement différent des deux autres (tableau III.2.9). À cet effet, nous avons eu recours au test de Wilcoxon car la distribution des erreurs $\left(vs_i - vs_{j\neq i}\right)$ ne correspond pas à une distribution Normale, comme la première ligne du tableau III.2.9 le montre. Les valeurs de significativité obtenues $p \ll \alpha$ montrent une différence significative entre les algorithmes pris deux à deux (ligne 2).

Test	(vs_1, vs_2)	(vs_1, vs_3)	(vs_2, vs_3)
Significativité du test de normalité de Shapiro-Wilk	$2,7 \cdot 10^{-13}$	$8,6 \cdot 10^{-5}$	$1,1 \cdot 10^{-23}$
Significativité du test *signed-rank* de Wilcoxon	$1,5 \cdot 10^{-1}$	$2,3 \cdot 10^{-8}$	$1,2 \cdot 10^{-7}$

Tableau III.2.9 – Comparaison appariée des trois algorithmes vs_i.

Les résultats obtenus dans cette section montrent que les trois algorithmes vs_i fournissent des résultats similaires. En termes statistiques, ils sont moyennement corrélés à la perception humaine ($r \approx 0,5$). La prochaine section définit et étudie des indicateurs centrés-individus pour évaluer le degré d'approximation de la perception humaine *ph* par les algorithmes de validation sociale vs_i.

2.2.3.2 Algorithmes *versus* perception humaine : indicateurs centrés-individus

En plus des résultats des tests statistiques rapportés dans la section précédente, nous définissons des indicateurs centrés-individus pour comparer vs_i et *ph*. Ces indicateurs permettent d'évaluer la capacité des algorithmes vs_i à égaler *ph* en termes de *distance*, *polarité* et *force*. Les résultats de ces trois indicateurs et leurs expressions sous forme prédicative sont rapportés pour chaque algorithme vs_i dans le tableau III.2.10. La fonction sgn : $x \mapsto \{-1; 0; 1\}$ retourne une valeur correspondant au signe de x, selon qu'il soit négatif (-1), nul (0) ou positif (1). Chaque indicateur existe en deux versions (stricte et tolérante), ils sont expliqués ci-dessous :

- la *distance* (D_i) mesure la différence entre vs_i et *ph* en fonction du nombre de positions qui les séparent sur la réglette représentée en figure III.2.6. L'indicateur D_1 (resp. D_2) reflète une séparation d'une (resp. de deux) position(s) entre le résultat de l'algorithme et la perception humaine ;
- la *polarité* (P_i) mesure la différence entre vs_i et *ph* en fonction de leur signe : types opinion ($\mathscr{R}$, $\mathscr{N}$, et $\mathscr{C}$ du tableau II.2.1, p. 49). En fait, P_1 est vrai lorsque vs_i et *ph* sont exactement du même type. L'indicateur P_2 est moins strict que P_1 étant donné que vs_i, *ph* ou les deux peuvent être nul(s) ;
- la *force* (F_i) mesure la capacité des algorithmes vs_i à égaler la même « zone » que *ph* sur la réglette représentée en figure III.2.6. Nous avons divisé cette réglette en trois zones dont les frontières sont exprimées par les prédicats F_i. Ces zones correspondent au consensus négatif, l'absence de consensus (indécision du groupe) et au consensus positif.

Indicateur	Prédicat	vs_1 (%)	vs_2 (%)	vs_3 (%)
D_1	$\lvert ph - vs_i \rvert \leqslant 0,25$	48,0	46,4	47,0
D_2	$\lvert ph - vs_i \rvert \leqslant 0,50$	77,1	76,3	77,3
P_1	$\text{sgn}(ph) = \text{sgn}(vs_i)$	66,1	65,4	65,4
P_2	$\text{sgn}(ph) = \text{sgn}(vs_i) \vee ph \cdot vs_i = 0$	84,8	84,7	84,6
F_1	$(ph, vs_i) \in \{[-1;-\frac{1}{3}[, [-\frac{1}{3};\frac{1}{3}],]\frac{1}{3};1[\}$	54,7	55,7	57,5
F_2	$(ph, vs_i) \in \{[-1;-\frac{2}{3}[, [-\frac{2}{3};\frac{2}{3}],]\frac{2}{3};1[\}$	73,5	73,7	74,2

Tableau III.2.10 – Comparaison entre la perception humaine (*ph*) et les algorithmes (vs_i).

Pour résumer les analyses réalisées dans cette section, la comparaison des algorithmes vs_i avec la perception humaine *ph* du consensus dans des débats argumentatifs montre que les trois algorithmes fournissent des résultats similaires du point de vue des tests d'hypothèses statistiques (tableau III.2.8) mais aussi de celui d'indicateurs centrés-individus (tableau III.2.10). Globalement, il existe une corrélation moyenne entre vs_i et *ph*. Les versions « tolérantes » des indicateurs centrés-individus montrent que vs_i approxime *ph* en termes de *distance*, *polarité* et *force* dans environ 80 % des cas. Bien que les trois algorithmes obtiennent des résultats similaires, ils peuvent être distingués en fonction de leur performance. À cet effet, le tableau III.2.11 rapporte les temps d'exécutions relatifs à chaque algorithme, en comparant la durée de calcul nécessaire pour calculer la validité sociale des arguments d'un échantillon qui en comprend 5 647.

	vs_1	vs_2	vs_3
Durée de calcul (en secondes)	66	13	17
Facteur multiplicateur correspondant (*baseline* : vs_2)	3,9	1,0	1,3

Tableau III.2.11 – Durées d'exécution des algorithmes vs_i pour 5 647 arguments.

L'algorithme le plus rapide est sans conteste vs_2. L'algorithme vs_3 est 30 % plus lent. Enfin, vs_1 est quasiment cinq fois plus lent que vs_2. Par conséquent, les algorithmes vs_2 et vs_3 sont les meilleures alternatives au regard de l'approximation de *ph* ainsi que de la durée d'exécution.

2.3 Discussion de l'expérimentation

La méthodologie d'expérimentation originale ainsi que l'analyse des données recueillies présentées dans ce chapitre peuvent susciter de nombreuses questions, nous discutons celles que nous avons identifiées dans cette section.

Une première interrogation peut porter sur le niveau d'Anglais des participants : ceux qui manquent de maîtrise ont pu éprouver des difficultés de compréhension lors de l'évaluation des débats. En effet, la plupart des participants ne sont pas anglophones natifs (tableau III.2.4) bien qu'ils déclarent une bonne maîtrise de cette langue (tableau III.2.5). De ce fait, certains participants ont certainement été déconcertés par la tâche ❶, qui consistait à identifier les opinions exprimées dans les arguments des débats. Toutefois, ce problème n'est pas critique pour deux raisons :

1. La première est énoncée dans la section III.2.2.2 : même si les individus identifient des opinions complètement différentes, ce n'est pas un problème car nous comparons leur perception humaine avec les résultats des algorithmes exécutés sur la base des *mêmes* types opinion ;
2. Deuxièmement, cette difficulté de compréhension due à la langue anglaise reflète la situation courante où des non anglophones natifs avec une maîtrise de la langue limitée sont confrontés à un Web où l'Anglais prédomine de nos jours.

Une seconde interrogation concerne l'ordonnancement des débats de façon à privilégier une difficulté progressive lors de leur évaluation. Une alternative aurait consisté à mélanger les débats présentés à chaque participant (*randomization*). Cette stratégie aurait certainement équilibré le nombre d'évaluations par débat, contrairement aux 121 évaluations obtenues pour le premier débat *versus* 53 pour le treizième. Toutefois, cette stratégie basée sur le mélange est incompatible avec une approche de difficulté progressive, condition qui nous semble être requise pour éviter l'abandon des participants dès les premières étapes de l'expérimentation. Par ce choix, nous allons à l'encontre d'une stratégie de type *high entrance barrier* recommandée par Reips (2007), visant à « provoquer tôt l'abandon pour garantir une participation soutenue après que le participant ait décidé de s'impliquer » [7]. Toutefois, un abandon dans notre expérimentation n'est pas en pure perte car les étapes évaluées durant de telles participations partielles sont quand même analysées.

Concernant la comparaison des résultats des algorithmes de validation sociale avec la perception humaine, trois points peuvent être discutés :

1. nous avons eu recours à une stratégie d'estimation de paramètres pour initialiser les algorithmes vs_i de façon à optimiser $vs_i \rightarrow ph$. Ainsi, les résultats rapportés dans ce chapitre concernent la comparaison entre ph et vs_i avec une configuration optimale des algorithmes ;
2. les composants graphiques simulant des réglettes (figure III.2.2) correspondent à des échelles ordinales, dont les étiquettes sont *réfuté*, *neutre* et *confirmé*. Nous avons dû encoder leurs valeurs pour réaliser la comparaison de la section III.2.2, les valeurs décimales assignées étant illustrées dans la figure III.2.6. Nous nous demandons si ces valeurs assignées correspondent

7. Texte original : "provoke dropout to happen early and ensure continued participation after someone makes the decision to stay." (Reips, 2007, p. 378)

à la perception des individus : est-ce que la différence de 25 % entre deux positions est consistante avec la perception humaine ? En fait, on pourrait estimer que la position qui précède *confirmé* représente un accord à 90 %, alors qu'il est actuellement encodé par 75 %. De ce fait, une alternative consisterait à encoder la réglette selon une distribution alternative (figure III.2.9), logarithmique par exemple. Cependant, nous n'avons à ce jour trouvé aucune étude supportant cette hypothèse ;

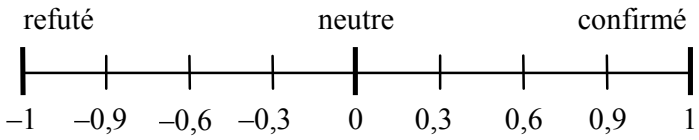


Figure III.2.9 – Encodage alternatif à l'encodage présenté en figure III.2.6.

3. le caractère généralisable de cette expérimentation « écologique » repose sur l'hypothèse de diversité des participants. Nous avons en effet supposé que les opinions identifiées seraient différentes — dues à la perception et à la sensibilité individuelles — durant la tâche ❶, fournissant alors des valeurs de synthèse d'opinions basées sur une large palette de combinaisons. Si chaque participant identifie les mêmes opinions que tous les autres, le fait d'avoir recruté 121 participants n'aurait aucun sens… Dans le but de vérifier la diversité des participants, nous avons calculé pour chaque débat le coefficient d'accord interpersonnel κ de Fleiss (1971) (section II.3.1.1, p. 57) sur la base des opinions identifiées. La figure III.2.10 présente les valeurs de κ résultantes, révélant un accord variable et globalement moyen entre les participants, en interprétant ces valeurs grâce au tableau II.3.2, p. 57.

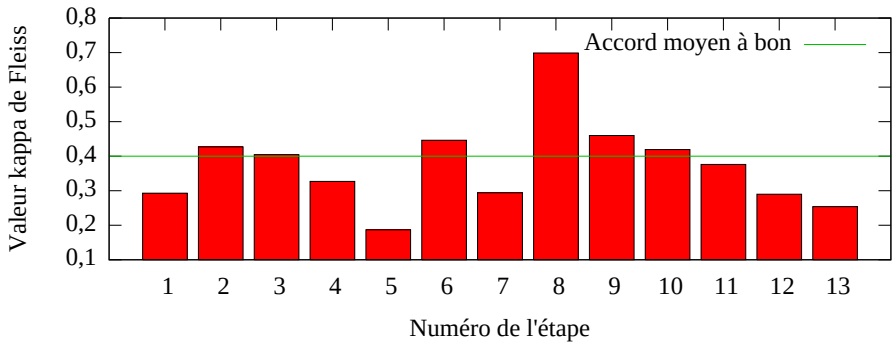


Figure III.2.10 – Accord interpersonnel sur l'identification des opinions (tâche ❶).

Ces résultats tendent à prouver que les participants ont identifié des opinions distinctes pour un même argument. Comme chaque participant s'est basé sur les types qu'il a lui-même identifiés pour les synthétiser (tâche ❷), les algorithmes de validation sociale ont par conséquent été exécutés avec de nombreuses combinaisons de paramètres en entrée (les différents types assignés aux arguments). Cela contribue à asseoir le caractère généralisable de nos résultats.

Une autre conclusion que nous pouvons tirer de la figure III.2.10 concerne le fait que les individus éprouvent des difficultés à identifier l'opinion d'un argument donné en lisant un débat. Cela signifie que l'interprétation des arguments est un problème critique pour les individus. Par

conséquent, les systèmes permettant de débattre au sein de fils de discussion devraient encourager les contributeurs à clairement exprimer l'opinion des arguments qu'ils rédigent. Ceci pourrait éviter aux lecteurs successifs de méprendre les opinions réelles des contributeurs. C'est une des observations qui nous ont incité à définir des types d'annotation (tableau II.2.1) dans notre modèle unifié (section II.2), afin que les annotateurs puissent clairement indiquer la sémantique de leurs annotations.

3 Expérimentation de la mesure de similarité basée sur l'usage des documents

"No one believes an hypothesis except its originator, but everyone believes an experiment except the experimenter."

William Ian Beardmore Beveridge (1908 — 2006)

À CONTRE-PIED de la mesure de similarité basée sur les contenus des documents, qui est somme toute classique dans le domaine de la Recherche d'Information, nous avons défini dans la section II.4 une mesure de similarité sur l'usage des documents. Son originalité est double. D'une part, elle reflète des liens d'usage entre les documents sur la base de leur structuration dans les EPA des membres organisationnels. D'autre part, elle ne requiert pas l'accès au contenu des documents ni une quelconque indexation.

Afin de vérifier que cette nouvelle mesure de similarité sur l'usage est bien différente de la mesure de similarité classique sur le contenu, nous présentons le résultat d'une expérimentation réalisée à partir d'un corpus hiérarchisé de documents médicaux. Elle complète l'observation empirique rapportée dans (Cabanac *et al.*, 2007a) et illustrée dans la figure II.6.1 (p. 80). Ce chapitre est organisé comme suit : nous exposons tout d'abord le protocole d'expérimentation que nous avons conçu avant d'analyser les résultats obtenus à l'aide d'outils statistiques. Enfin, un bilan de cette expérimentation clôt le présent chapitre.

3.1 Protocole d'expérimentation

Nous explicitions dans cette section l'hypothèse soumise à expérimentation, le corpus de documents retenu à cet effet, ainsi que la méthodologie adoptée pour analyser les résultats à l'aide de tests statistiques d'hypothèse.

3.1.1 Hypothèse : complémentarité entre similarité de contenu et d'usage

Ayant proposé une nouvelle mesure de similarité entre documents, nous souhaitons montrer par l'expérimentation que les résultats de cette mesure sont bien différents de ceux obtenus avec une similarité classique en Recherche d'Information, calculée à partir du contenu des documents. Le cas échéant, nous pourrons les considérer comme non-redondantes et indépendantes, justifiant de fait l'existence de la mesure de similarité sur l'usage proposée.

3.1.2 Constitution du corpus d'expérimentation

Le calcul de la similarité sur l'usage entre deux documents nécessite qu'ils soient organisés dans une arborescence. Afin de réaliser l'expérimentation, nous avons recouru à la collection de documents OHSUMED utilisée dans la plate-forme d'expérimentation TREC (Voorhees, 2001). Elle comprend 348 566 documents issus de la base de données MEDLINE qui regroupe 270 revues médicales sur une période de cinq ans (1987 – 1991). Nous avons retenu la collection OHSUMED car chaque document a été manuellement associé à une ou plusieurs catégorie(s) d'une hiérarchie, la *MEdical Subject Headings* (MeSH) de la *National Library of Medicine* des États-Unis en l'occurrence. Cette donnée correspond aux jugements des experts appelés *qrels* — pour *query-relevance* — dans TREC.

Pour des raisons de calculabilité, nous avons sélectionné un sous-ensemble de MeSH correspondant à la hiérarchie des maladies cardio-vasculaires. Cette hiérarchie comprend 146 nœuds et sa profondeur maximale est égale à six. De plus, la collection OHSUMED comprenant les documents est divisée en deux parties : collection d'entraînement et collection de test. Les $n = 4\,974$ documents que nous avons retenu pour cette expérimentation proviennent de l'intégralité de la collection d'entraînement.

3.1.3 Méthodologie

Afin de comparer les mesures de similarités sur le contenu *versus* sur l'usage des n documents retenus, nous les avons au préalable calculées comme suit :

1. Concernant la mesure de similarité sur le contenu, nous avons indexé les n documents de notre corpus d'expérimentation. Pour ce faire, la chaîne de traitement classique en Recherche d'Information (exposée en section II.2.2.2, p. 53) a été mise en œuvre. Enfin, nous avons représenté les n documents dans le modèle vectoriel (cf. section II.6.1.3.2, p. 87) afin de pouvoir calculer leur similarité sur le contenu $s_C\left(d_i, d_j\right)$ grâce à la fonction $\cos\left(\vec{d}_i, \vec{d}_j\right) \in [0;1]$;
2. Concernant la mesure de similarité sur l'usage, nous avons représenté la hiérarchie comprenant les n documents dans la structure de données multi-arbres (définition 5, p. 66). Le calcul de leur similarité sur l'usage $s_U\left(d_i, d_j\right)$ correspond à la fonction $\sigma_D \in [0;e]$ (section 4.7, p. 68) normalisée sur l'intervalle $[0;1]$. Cette normalisation vise à permettre la comparaison des valeurs de s_C et de s_U.

En s'appuyant sur la propriété de symétrie des mesures de similarité considérées, nous avons calculé les similarités de contenu $s_C\left(d_i, d_j\right)$ et d'usage $s_U\left(d_i, d_j\right)$ des documents de telle sorte que $1 \leqslant i < j \leqslant n$. Nous avons alors obtenu $N = \frac{n \cdot (n-1)}{2} = 12\,367\,851$ couples $\left(s_C\left(d_i, d_j\right), s_U\left(d_i, d_j\right)\right)$.

Afin de comparer les résultats des deux mesures, nous avons utilisé les tests statistiques d'hypothèse présentés dans la section III.2.2.3.1 (p. 109). Il s'agit des tests de Student (paramétrique) et de Wilcoxon (non-paramétrique) sur échantillons appariés, fournissant une valeur de significativité p permettant d'interpréter l'issue du test. Une valeur $p < \alpha = 0{,}05$ affirme une différence statistique significative entre les deux séries examinées. De plus, cette différence est d'autant plus avérée que $p \to 0$.

Par ailleurs, le coefficient $r \in [-1;1]$ de Pearson a permis d'évaluer le degré de corrélation entre les deux méthodes. Elles sont d'autant plus proches que $r \to 1$, à un coefficient multiplicateur $k \in \mathbb{R}_+$ près : quelle que soit X une série, $r(X, k \cdot X) = 1$. De plus une corrélation inverse $\left(\text{c'est-à-dire } x = y^{-1}\right)$ est mise en évidence par $r \to -1$. Enfin, $r \to 0$ est révélateur de deux séries non corrélées.

3.2 Vérification de l'hypothèse sous expérimentation

Avant de comparer les deux mesures de similarité à l'aide d'outils statistiques, nous avons représenté leurs résultats par des graphiques. Hull (1993) souligne le fait qu'ils permettent une première comparaison visuelle, notamment pour estimer si la distribution des erreurs entre les deux méthodes étudiées suit une loi Normale ou pas.

La figure III.3.1 représente la distribution des valeurs de similarités calculées à partir du contenu des documents (s_C). Les valeurs de similarité sont peu dispersées car elles appartiennent majoritairement à l'intervalle [0;0,2]. De plus, on remarque une forte concentration (40 %) sur la valeur zéro.

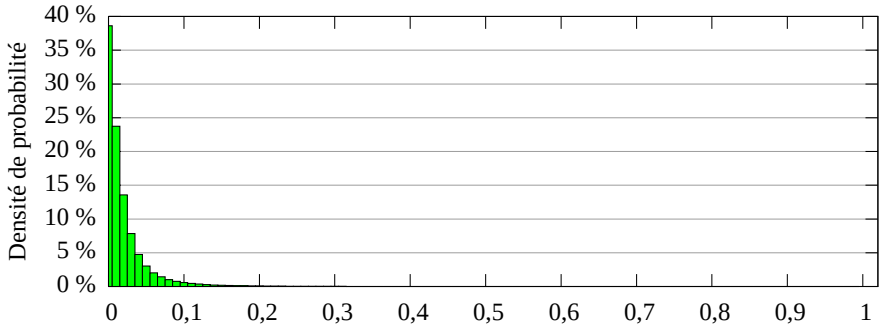


Figure III.3.1 – Distributions des valeurs de similarité s_C sur le contenu des documents.

La figure III.3.2 représente les valeurs de similarité sur l'usage des documents, l'échelle des abscisses [0;1] est la même alors que celle des ordonnées est différente : [0;12] *versus* [0;40]. On remarque que les valeurs de s_U sont davantage diversifiées sur tout l'axe des abscisses, contrairement aux valeurs de s_C qui sont regroupées au plus proche de zéro.

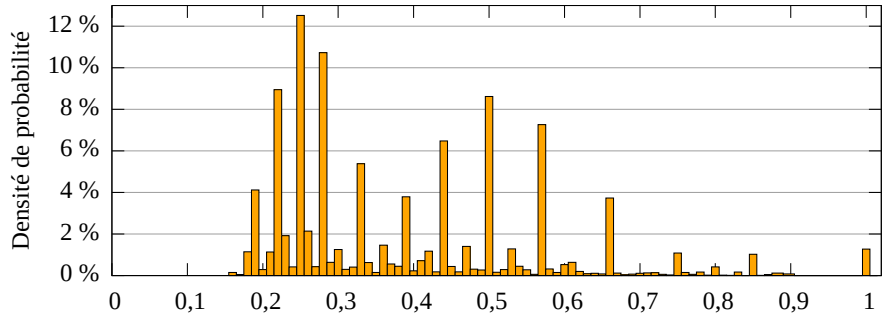


Figure III.3.2 – Distributions des valeurs de similarité s_U sur l'usage des documents.

Enfin, la figure III.3.3 illustre la différence entre les valeurs de similarité de contenu (s_C) et les valeurs de similarité d'usage (s_U). On remarque que la distribution des différences n'est pas centrée sur zéro, elle ne suit vraisemblablement pas une loi Normale. Enfin, la différence $s_C - s_U$ est majoritairement négative, signifiant que $s_C < s_U$ conformément à nos observations précédentes.

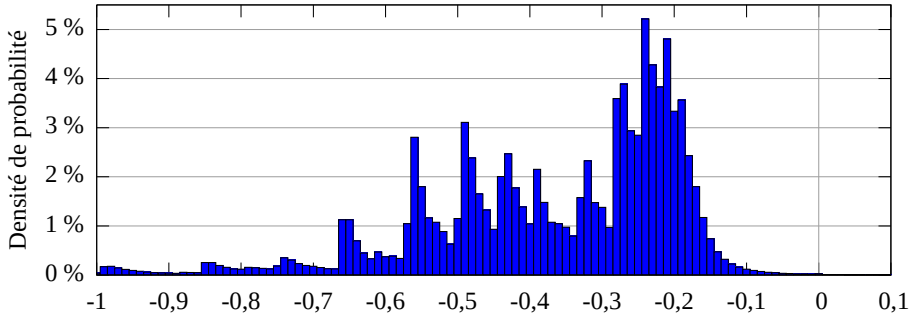


Figure III.3.3 – Distribution de la différence $s_C - s_U$ entre les deux mesures de similarité.

Pour compléter ces observations, le tableau III.3.1 présente des statistiques élémentaires réalisées sur les deux séries originales (s_C et s_U) ainsi que sur leur différence ($s_C - s_U$). Notons que si les deux méthodes étaient identiques, la moyenne ainsi que l'écart-type de leur différence tendraient vers zéro, ce qui n'est pas le cas ici.

Variable	Moyenne	Écart type	Minimum	Maximum
s_C	0,024	0,035	0,000	1,000
s_U	0,392	0,172	0,166	1,000
$s_C - s_U$	−0,368	0,170	−1,000	0,797

Tableau III.3.1 – Statistiques élémentaires sur les trois séries s_C, s_U et $s_C - s_U$.

Le tableau III.3.1 montre que la similarité d'usage minimale (0,166) est supérieure à zéro, ce qui est normal car tous les documents sont dans une même branche de la hiérarchie MeSH. Par ailleurs, la gamme de la similarité d'usage (s_U) est la plus étendue car son écart type (0,172) est plus élevé que celui associé à la similarité de contenu (0,035).

Concernant l'hypothèse expérimentale, le t-test de Student ainsi que le test de Wilcoxon

concluent tous deux indiscutablement à une différence statistiquement significative des deux séries, avec une valeur de significativité $p = 0,000$. Le coefficient de Pearson $r = 0,154$ reflète également l'absence de corrélation entre ces deux séries.

3.3 Bilan de l'expérimentation de la mesure de similarité basée usage

Afin de démontrer l'indépendance de la mesure de similarité sur l'usage par rapport à la mesure de similarité sur le contenu, nous avons proposé un protocole d'expérimentation reposant sur 4 974 documents issus de la collection de test OHSUMED. Les tests statistiques d'hypothèse concluent à une différence statistique significative entre les résultats de ces deux mesures de similarité. À la lumière de ces résultats, ces deux mesures ne sont pas identiques ou similaires, mais bien complémentaires sur le corpus que nous avons testé.

4 Implantation de la contribution : le prototype TafAnnote pour améliorer les activités documentaires

"What I hear, I forget. What I see, I remember. What I do, I understand."
Confucius (551 av. J.-C. — 479 av. J.-C.)

FÉDÉRER ET AMÉLIORER les activités documentaires sont les objectifs au cœur de notre contribution (partie II). Les deux chapitres précédents ont mis en œuvre des méthodologies d'expérimentation et d'analyse pour valider les propositions originales que nous avons préalablement exposées. Le présent chapitre vise à étayer cette démarche globale de validation scientifique. Pour ce faire, nous détaillons l'implantation de la contribution au sein du prototype de recherche TafAnnote, dont nous avons entrepris le développement en Master Recherche (Cabanac, 2005). Par la réalisation effective de cette application, nous avons souhaité apporter une « preuve de concept » visant à démontrer la faisabilité technique de nos propositions.

Ce chapitre est organisé comme suit. La première section présente un aperçu du prototype, son architecture faisant l'objet de la deuxième section. Puis, nous détaillons les fonctionnalités du système qui reprennent les éléments du modèle unifié et des processus intégrés (niveaux « microscopique » et « macroscopique »). Enfin, nous soulignons les limites du prototype avant de dresser un bilan de ce développement.

4.1 Description du prototype TafAnnote

L'approche retenue repose sur la fédération des six activités documentaires au sein d'un seul système. En phase de conception, deux alternatives nous sont apparues :

1. développer une application *autonome* qui remplacerait toutes celles de l'usager. Cette alter-

native ne contraint pas le développeur qui réalise l'application de bout en bout, en retenant les technologies de son choix, notamment. Par contre, l'usager doit renoncer aux pratiques acquises avec ses anciennes applications et se former à la nouvelle. De fait, il risque de résister au changement à cause de l'important effort d'adaptation requis par cette alternative ;

2. développer un composant additionnel *intégré* dans l'une des applications — appelée hôte — que l'usager maîtrise déjà afin d'étendre ses fonctionnalités. Cette alternative contraint le développeur aux technologies et restrictions imposées par l'application hôte. Par contre, l'usager est moins dérouté car, étant déjà familier de l'application existante, seules les nouvelles fonctionnalités sont à acquérir. Dans un premier temps, il peut même en ignorer une grande partie (par ex. : l'assistance lors de la réorganisation de son EPA) et n'utiliser que les plus utiles pour son activité (par ex. : l'annotation pour mémoriser des bribes de documents).

Dans l'objectif de limiter la résistance au changement, nous avons retenu la seule alternative viable : l'application intégrée. Le choix de l'application hôte est technologiquement contraint par son caractère extensible. Sur le plan humain, elle doit déjà être maîtrisée par les usagers. Ces contraintes nous ont conduit à retenir le navigateur de l'usager comme application hôte, Microsoft Internet Explorer et Mozilla Firefox étant les leaders du marché [1]. L'étude de l'extensibilité des deux navigateurs (Cabanac, 2005) a identifié un plus fort potentiel chez Firefox, cette observation est encore d'actualité (Paramelle, 2008). C'est pourquoi le prototype TafAnnote est un composant additionnel [2] pour le navigateur Firefox. Il s'y intègre sous la forme d'une barre d'outils (figure III.4.1) donnant accès à ses fonctionnalités.

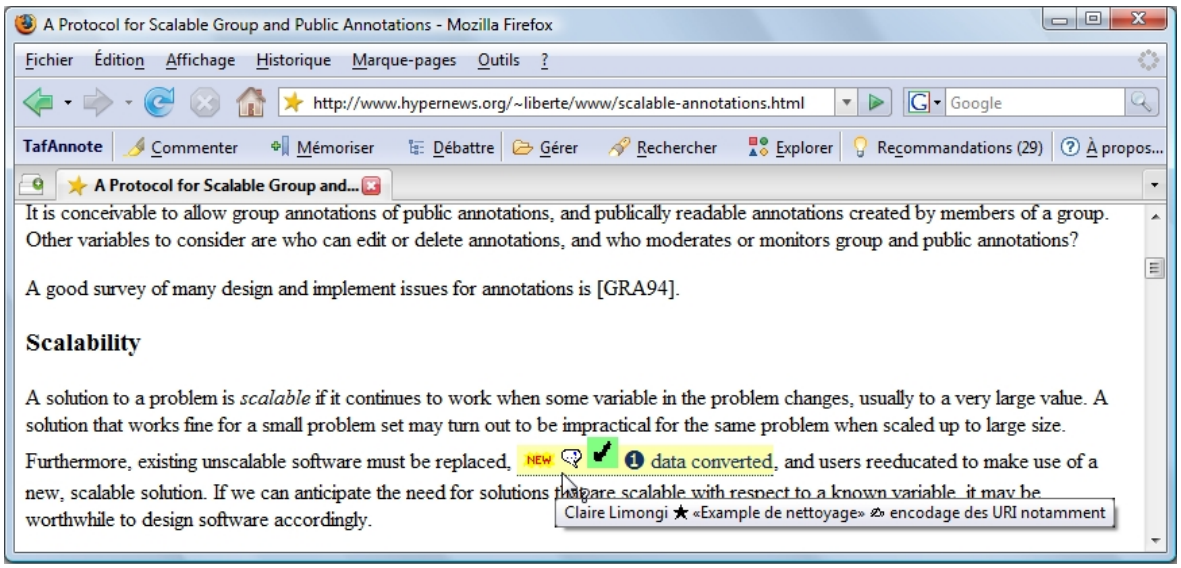


Figure III.4.1 – Barre d'outils de TafAnnote intégrée dans le navigateur Firefox. La page visualisée contient une annotation, sur laquelle l'usager a positionné le pointeur de la souris pour obtenir des informations sur son auteur et son contenu.

Nous exposons dans la section suivante l'architecture de notre prototype avant de détailler ses fonctionnalités dans la section III.4.3.

1. En juin 2008, Firefox était crédité de 42,6 % de parts de marché, contre 52 % pour Internet Explorer selon le W3C, cf. `http://www.w3schools.com/browsers/browsers_stats.asp`.
2. Une « extension », dans le vocable Firefox, cf. `http://developer.mozilla.org/fr/docs/Extensions`.

4.2 Architecture du prototype TafAnnote

TafAnnote repose sur l'architecture client-serveur schématisée dans la figure III.4.2.

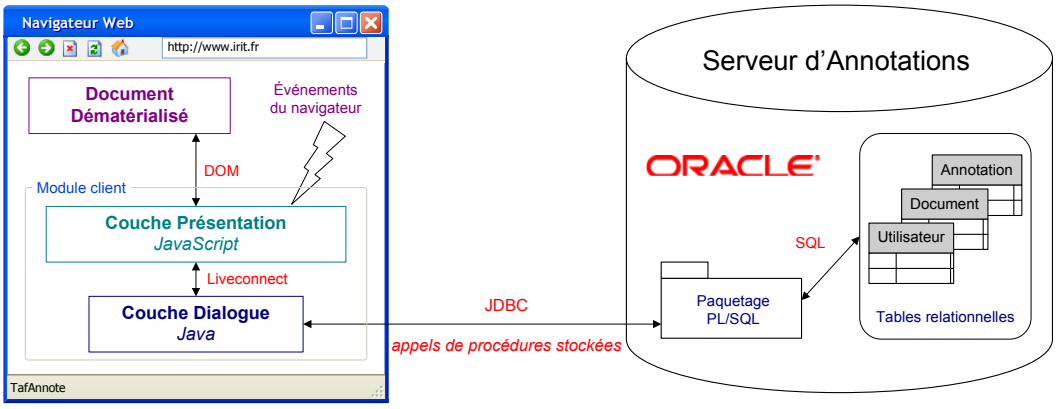


Figure III.4.2 – Architecture générale du prototype « preuve de concept » TafAnnote.

Le module *client* s'insère dans Firefox pour étendre ses fonctionnalités, notamment par la gestion et la création d'annotations collectives sur les documents au format HTML. Ce dernier établit un « pont » entre deux éléments principaux. D'une part, il manipule le document dématérialisé affiché par l'usager selon les événements du navigateur desquels il est à l'écoute. D'autre part, il échange des données au travers du réseau avec le module *serveur* détaillé dans la section suivante.

4.2.1 TafAnnote : module serveur

Le module serveur est une base de données (BD) relationnelles reposant sur le SGBD Oracle $10g2$; il implante le modèle objet unifié proposé (chapitre II.2). Concernant l'échange de données entre le client et le serveur, nous avons identifié deux approches :

1. l'implantation d'un mapping objet-relationnel via un *framework* de persistance tel qu'Hibernate. Dans ce cas, les instances des classes métier du modèle unifié sont manipulées du côté client, puis automatiquement transformées en tuples relationnels ou opérations correspondantes dans le Langage de Modification des Données (LMD) relationnel lors de la validation des transactions. Les requêtes sont alors stockées dans le module client et envoyées au module serveur à chaque exécution ;
2. la définition d'une interface de programmation (API) spécifiant les services fournis par le module serveur. Chaque service possède un nom ainsi que des paramètres en entrée et en sortie, son code étant stocké dans le module serveur. Cette approche s'implante en Oracle en définissant des procédures et fonctions (les services) rassemblées dans un paquetage PL/SQL (l'API). Le module client utilise les services ainsi créés en transmettant les données en entrée des procédures stockées uniquement, leur code étant déjà présent du côté serveur.

Nous avons retenu la seconde alternative pour trois raisons. Premièrement, elle favorise un couplage faible entre les modules client et serveur. Par exemple, la BD relationnelle peut être remplacée par une BD objet sans adaptation du code du module client, dès lors que les services sont maintenus. Deuxièmement, les transferts réseaux sont minimaux car seuls le nom de la procédure

et ses paramètres en entrée sont transmis, contrairement à l'envoi des n requêtes nécessaires avec l'approche mapping (sans procédure). Enfin, comme le code des services réside du côté serveur, la BD peut précalculer le plan d'exécution des requêtes associées afin de les accélérer.

4.2.2 TafAnnote : module client

Le module client est intégré à Firefox au travers d'une barre d'outils, il comprend deux parties : la couche « présentation » et la « couche dialogue ». Les composants additionnels de ce navigateur doivent être développés en XUL (XML User Interface Language) pour la partie interface, et en JavaScript pour réaliser les traitements associés. Nous avons donc spécifié les composants de la barre d'outils (boutons et icônes) en XUL. Par contre, nous avons été confronté à deux limites concernant ces technologies : l'impossibilité d'établir une connexion à la BD et l'impossibilité d'afficher des composants graphiques de type arborescence (pour représenter les fils de discussion). Pour contourner ces limites, nous avons recouru à un langage invocable à partir de JavaScript grâce à LiveConnect : Java, associé à JDBC (Java Database Connectivity) pour communiquer avec la BD et à sa bibliothèque standard de composants graphiques Swing. En outre, comme Java est un langage portable, TafAnnote peut être exécuté sur tout système d'exploitation. Concrètement, le scénario suivant illustre l'inter-opération de ces technologies.

1. Un individu lance son navigateur, ce dernier charge la barre d'outils de TafAnnote en mémoire centrale (couche présentation). Une fois chargé, le navigateur appelle la fonction JavaScript d'initialisation de TafAnnote qui se met alors à l'écoute des événements du navigateur (par ex. : chargement d'une page) et instancie la couche dialogue Java. Cette dernière établit alors une connexion avec la BD et se met en attente.
2. L'individu tape une URL dans son navigateur, qui récupère alors le document associé. La couche présentation est notifiée de cet événement et demande les annotations ancrées sur l'URL transmise à la couche dialogue. Cette dernière, qui communique avec la BD, invoque le service BD associé, obtient les annotations et les retourne à la couche présentation. Enfin, la couche présentation identifie leurs points d'ancrage exprimés en XPointer et y insère les annotations en modifiant le DOM (Document Object Model) du document récupéré.
3. En parallèle avec le point précédent, TafAnnote indexe le document dès qu'il est récupéré par le navigateur. Son contenu est transmis à la couche dialogue qui réalise les tâches suivantes détaillées dans la section II.2.2.2 (p. 53) : segmentation, élimination des mots vides, lemmatisation et pondération des termes. Les couples (terme, poids) obtenus sont ensuite envoyés à la BD pour intégration à l'index global. Cette approche originale d'indexation décentralisée met à profit les ressources de calcul des postes clients, ce qui décharge le serveur de transferts réseaux (pour récupérer les documents) et de calculs importants.
4. Toute fenêtre de TafAnnote est un composant graphique Java : par exemple, la sélection d'une annotation dans le document ouvre une fenêtre permettant de la consulter et éventuellement de la modifier ou d'y répondre.

Après avoir détaillé l'architecture client-serveur et les technologies sous-tendant TafAnnote, nous présentons ses fonctionnalités qui ont été spécifiées dans la partie II.

4.3 Fonctionnalités issues du modèle unifié et des processus

Le prototype TafAnnote fédère les activités documentaires des membres organisationnels et leur permet de visualiser et d'explorer le capital documentaire qui en résulte. Il est téléchargeable à partir de la page Web dédiée « `http://www.irit.fr/~Guillaume.Cabanac/TafAnnote` ». Une fois installé, l'individu s'inscrit pour obtenir son login et son mot de passe ou utilise le compte de test fourni à des fins d'évaluation. L'usager a alors accès aux différentes fonctionnalités que nous exposons dans les sections suivantes.

4.3.1 Niveau « microscopique » : amélioration des activités documentaires

4.3.1.1 Visualisation et création d'annotations collectives

L'usager de TafAnnote visualise les annotations en contexte, au sein des documents qu'il consulte. Par exemple, une annotation est présente sur le document de la figure III.4.1, le pictogramme « NEW » indique au lecteur qu'il n'a jamais vu cette annotation : c'est soit sa première visite du document, soit l'annotation a été créée après sa dernière visite. Le pictogramme « » renseigne le lecteur sur le fait qu'un commentaire est associé à l'annotation ; il peut en lire les premiers mots en survolant l'annotation avec la souris. L'infobulle qui s'affiche alors contient également le nom de l'annotateur (Claire Limongi dans l'exemple) et son expertise (★ représente le niveau d'expertise le plus faible). Enfin, un clic sur l'annotation affiche une fenêtre comprenant l'ensemble des données de l'annotation.

Créer une annotation nécessite de choisir entre les trois boutons *Commenter*, *Mémoriser* et *Débattre* représentant les trois objectifs modélisés (section II.2.1.2, p. 48). La sélection d'une partie du document en tant que point d'ancrage est facultative car on peut annoter tout ou partie de ce dernier. Pour chaque catégorie d'annotation, une fenêtre contenant uniquement les informations nécessaires est présentée à l'usager. À titre d'exemple, la figure III.4.3 montre la création d'une annotation argumentative (bouton *Débattre*).

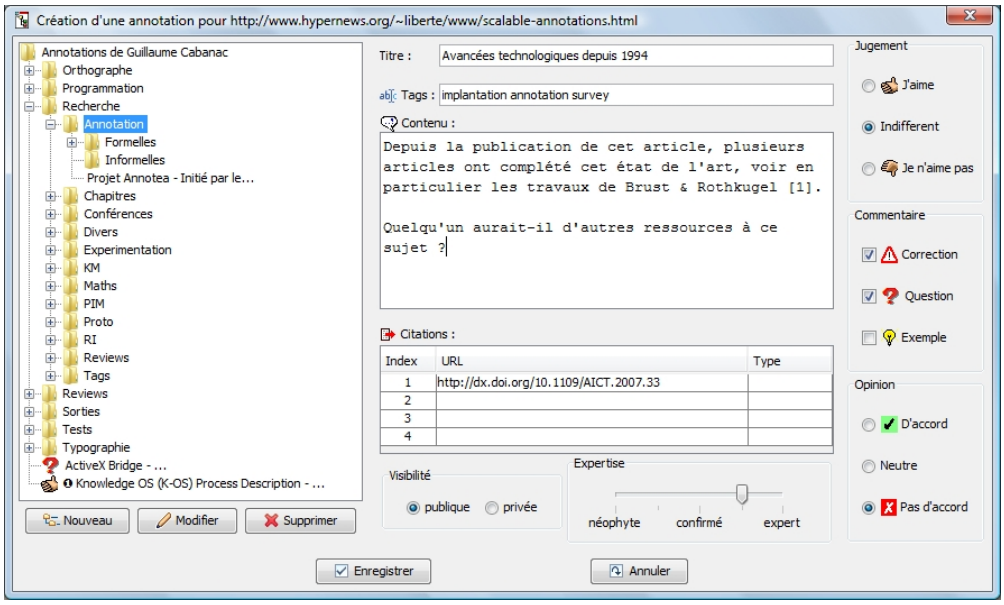


Figure III.4.3 – Création d'une annotation argumentative avec TafAnnote.

Dans cet exemple, l'usager peut décider de conserver son annotation dans son EPA. Il sélectionne alors le répertoire adéquat (/Recherche/Annotation) dans le composant de gauche. Il fournit son titre, les *tags* qui la caractérisent, un commentaire et une référence. De plus, il donne une visibilité publique à cette annotation pour laquelle il se déclare avoir une expertise élevée. Il indique également que son commentaire comprend une question et une correction. Notons que les pictogrammes présents dans cette interface sont également utilisés pour afficher les annotations en contexte. Ainsi, un lecteur peut identifier visuellement toutes les questions, par exemple. Enfin, il positionne son opinion (réfutation) car il a choisi de créer une annotation argumentative. La figure III.4.4 illustre l'affichage de la page annotée, où on distingue la nouvelle annotation sur son point d'ancrage « good survey ».

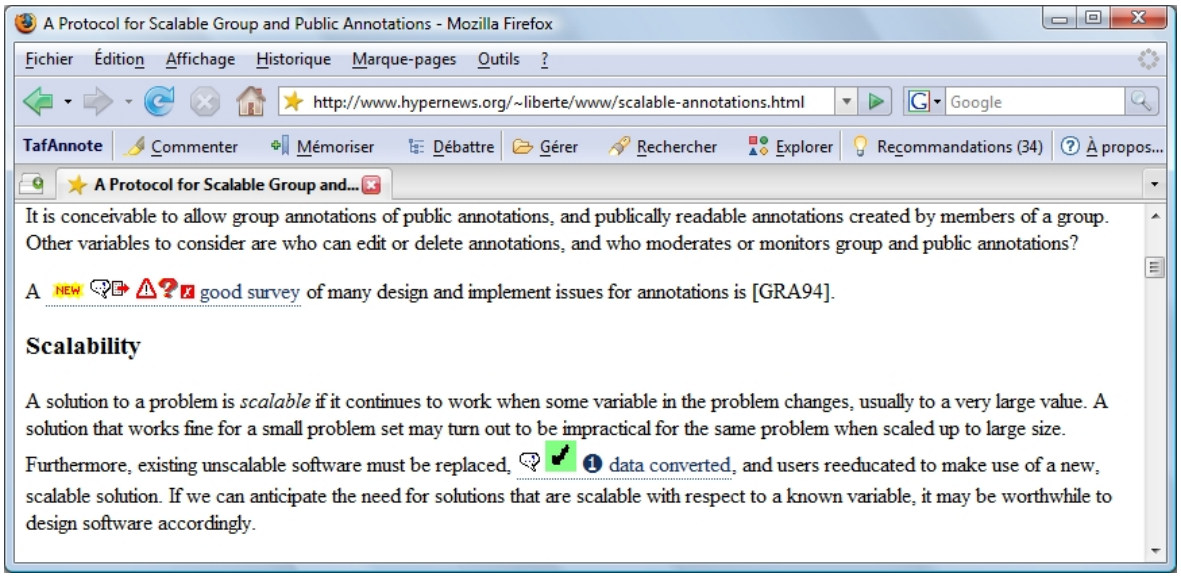


Figure III.4.4 – Visualisation en contexte de l'annotation créée avec TafAnnote.

Notons que l'intégralité du point d'ancrage (du début à la fin) est clairement visualisable, contrairement à d'autres systèmes d'annotation qui insèrent un pictogramme uniquement au début du point d'ancrage, compliquant de fait son interprétation (cf. section I.3.6, p. 38). Enfin, le pictogramme « ❶ » de la seconde annotation indique qu'elle a obtenu une réponse. Le lecteur peut y accéder en cliquant dessus, ce qui a pour effet d'afficher une fenêtre similaire à celle de la figure III.4.3 dont le composant de gauche (EPA) est remplacé par le fil de discussion. Au travers de cette fenêtre, le lecteur peut consulter chaque argument et y répondre.

4.3.1.2 Gestion des annotations collectives : l'Espace Personnel d'Annotations (EPA)

Une fois que l'usager est identifié, il a accès à son EPA via le bouton *Gérer* de la barre d'outils. La fenêtre correspondante (figure III.4.5) comprend trois onglets, permettant de consulter ses annotations selon différents points de vue : selon leur classement original, classées par type ou classées par *tag*. L'usager peut consulter chacune de ses annotations pour éventuellement les modifier, il peut également les supprimer ou visualiser les documents annotés. Enfin, au sein du premier onglet, l'usager peut réorganiser manuellement son EPA en déplaçant (*drag and drop*) les annotations ou répertoires où bon lui semble.

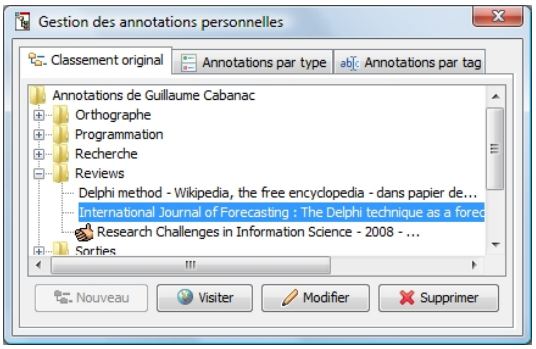


Figure III.4.5 – EPA de l'usager, les annotations sont également visualisables par type et par tag.

TafAnnote offre également une fonctionnalité de recherche dans les annotations, grâce au bouton *Rechercher* de la barre d'outils. La recherche est initiée à partir d'une requête (termes et connecteurs booléens) dont on peut filtrer les résultats selon les noms d'annotateurs spécifiés.

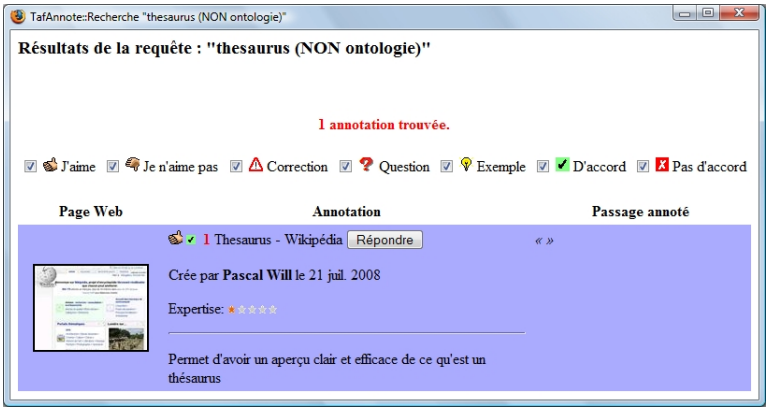


Figure III.4.6 – Résultat d'une recherche dans le corpus des annotations.

La figure III.4.6 montre le résultat de la requête « thesaurus (NON ontologie) » : une seule annotation-résultat. Les résultats peuvent être filtrés par type. Les informations relatives à chaque annotation sont accompagnées d'une miniature du document annoté pour en offrir un aperçu.

4.3.1.3 Recommandations lors de la navigation

Au cours de la navigation de l'usager, le processus NAVI (section II.5.5, p. 75) lui recommande des documents liés par l'usage, issus des EPA des autres membres organisationnels. Les recommandations sont accessibles en cliquant sur le bouton *Recommandations (n)* de la barre d'outils, où *n* représente le nombre de recommandations suggérées. Sur l'exemple de la figure III.4.7, l'usager a obtenu dix recommandations dont trois au moins sont nouvelles (indicateur rouge). Les recommandations sont ordonnées selon leur score, chacune correspond à un document qui a été annoté par les personnes mentionnées. Un clic sur une recommandation affiche le document associé dans le navigateur. Les recommandations dont l'indicateur est vert sont réitérées : le document associé a été recommandé à plusieurs reprises durant la navigation de l'usager. Dans ce cas, le score présenté est obtenu en faisant la somme de tous les scores précédents.

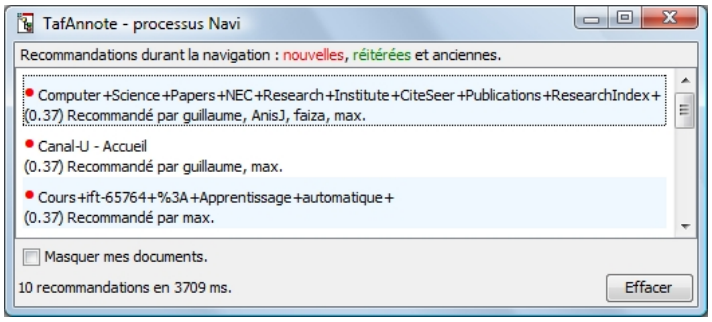


Figure III.4.7 – Recommandations émises durant la navigation de l'usager.

4.3.2 Niveau « macroscopique » : exploration du capital organisationnel

Les EPA représentent de véritables mines d'informations constituées par et pour les membres organisationnels. Afin d'en permettre la visualisation et l'exploration, le chapitre II.6 a défini une interface multi-facettes que nous avons intégrée dans la barre d'outils de TafAnnote (Moreau, 2008). Ainsi, le bouton *Explorer* donne accès à une fenêtre comprenant quatre onglets, un pour chacune des vues représentées en figure II.6.4 (p. 85). De fait, l'usager voit en permanence dans quelle vue il se situe. Cette section présente le développement réalisé à l'aide d'un scénario réaliste : Pierre est affecté à un projet de développement Java, il utilise l'interface multi-facettes pour se documenter à ce sujet. Au lancement, il obtient la vue 1 en mode « classification » (figure III.4.8).

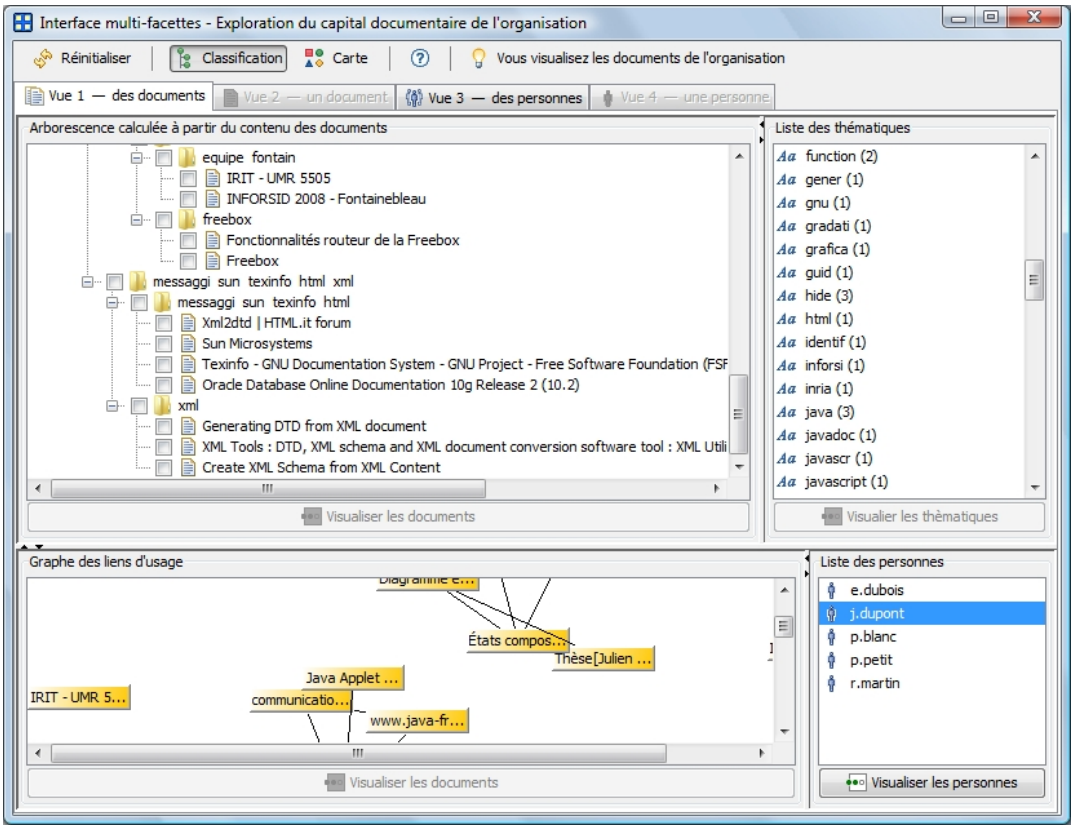


Figure III.4.8 – Vue 1 (classification) montrant tous les documents de l'organisation.

La première facette représente la hiérarchie des thématiques des documents de l'organisation obtenue par classification ascendante hiérarchique (section II.6.1.3.3, p. 88). Les trois autres facettes complètent la vue 1 en montrant ces mêmes documents reliés selon leur usage, ainsi que les différents membres organisationnels qui les possèdent. Grâce aux boutons de la barre d'outils, l'usager peut changer de technique de visualisation, en passant du mode « classification » au mode « carte ». Ce dernier mode affiche une carte auto-organisatrice permettant de visualiser sur un seul écran les thématiques les plus présentes dans l'organisation. La carte est divisée en zones correspondant à des documents. Chaque zone est étiquetée avec les thématiques de ses documents, sa couleur de fond est d'autant plus claire qu'elle regroupe beaucoup de documents. Enfin, l'usager peut « forer » cette carte en cliquant sur une zone, il raffine ainsi sa recherche jusqu'à obtenir un seul document, de proche en proche.

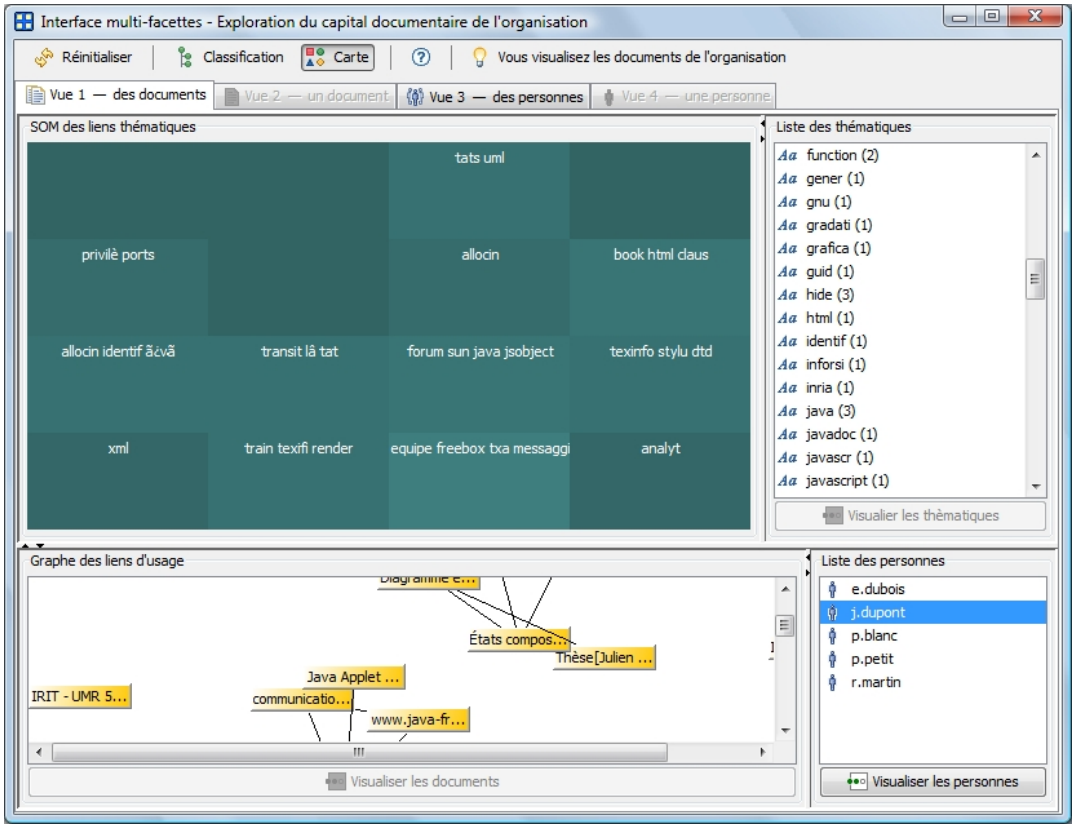


Figure III.4.9 – Vue 1 (carte) montrant tous les documents de l'organisation.

Désireux de consulter en priorité les documents que son chef de projet a jugé bon de conserver, Pierre accède à la fiche de ce dernier (figure III.4.10). Elle contient son identité (Jean Dupont), les groupes auxquels il appartient, les thématiques de ses documents, son EPA ainsi que les personnes qui partagent les mêmes thématiques que lui. Pierre repère le répertoire JAVA dans l'EPA de Jean, il sélectionne en particulier le document de Jean intitulé *Liveconnect Java Javascript*.

Pierre accède alors à la vue 2 (figure III.4.11) qui présente la fiche du document, en détaillant son titre, son URL, les chemins dans les EPA des personnes qui l'ont stocké et ses thématiques. Pierre a également accès aux documents utilisés avec celui-ci ainsi qu'aux logins de leurs propriétaires.

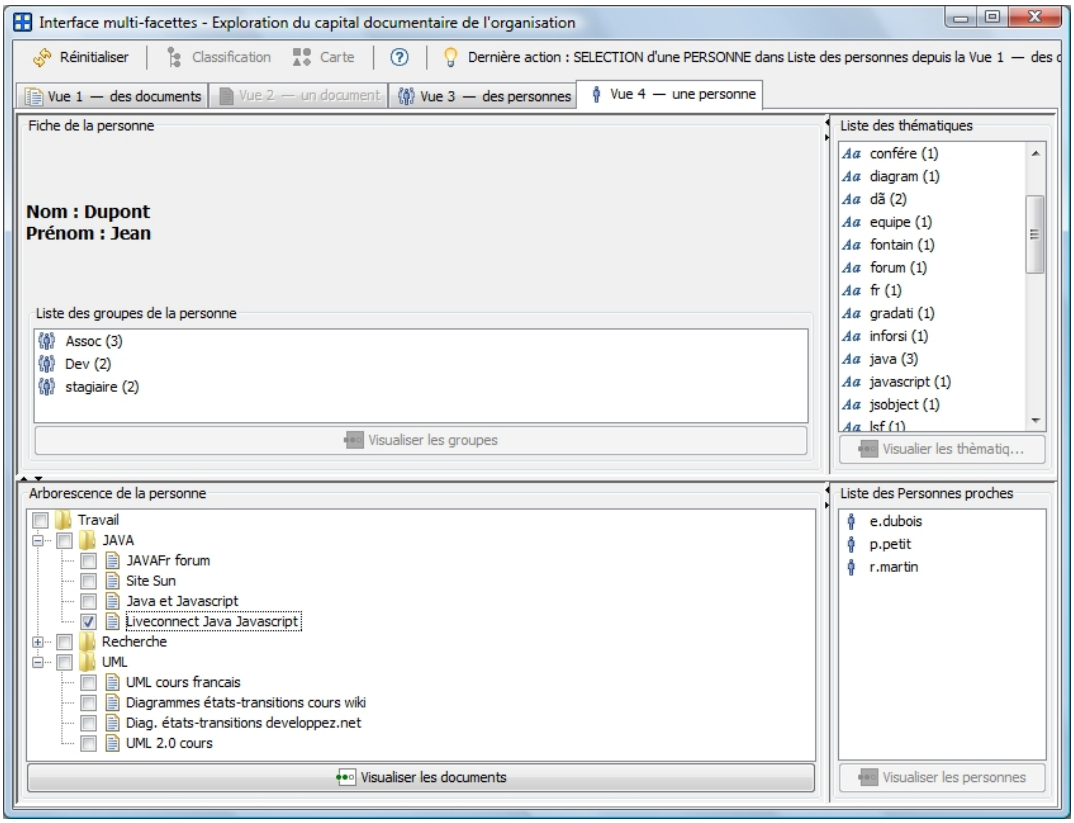


Figure III.4.10 – Vue 4 de l'interface montrant la fiche du membre « Jean Dupont ».

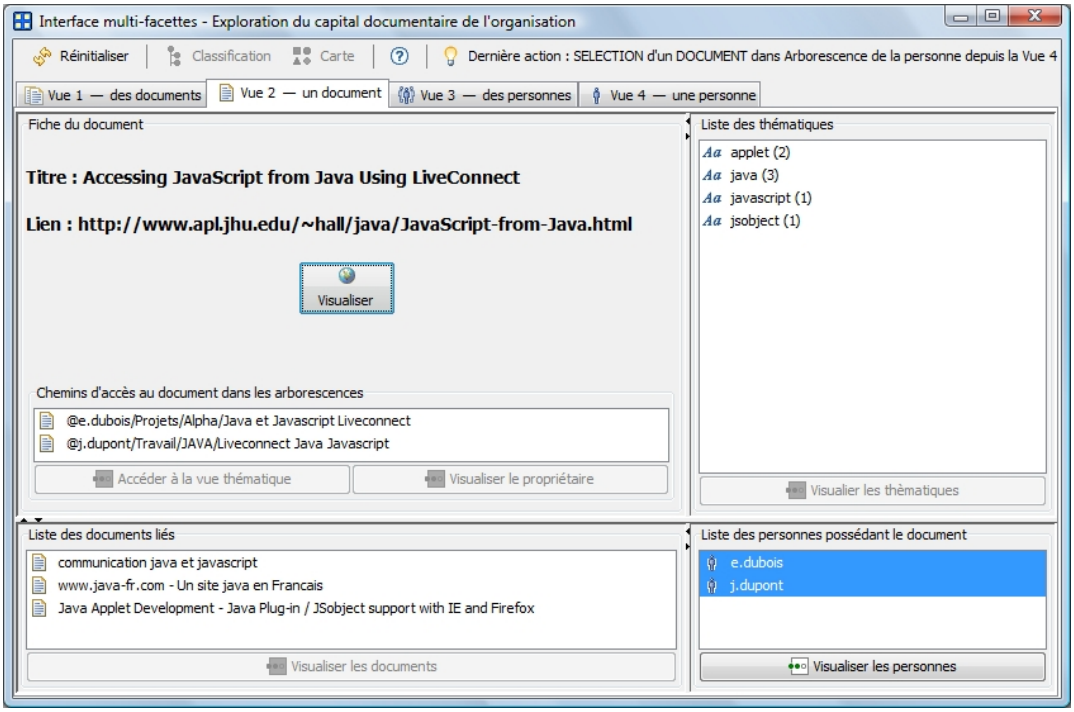


Figure III.4.11 – Vue 2 de l'interface montrant la fiche du document sélectionné.

En sélectionnant deux personnes dans la quatrième facette de l'interface, Pierre obtient la vue 3 représentée en figure III.4.12. Il peut alors connaître les thématiques et groupes que les deux personnes sélectionnées partagent, puis continuer l'exploration du capital documentaire.

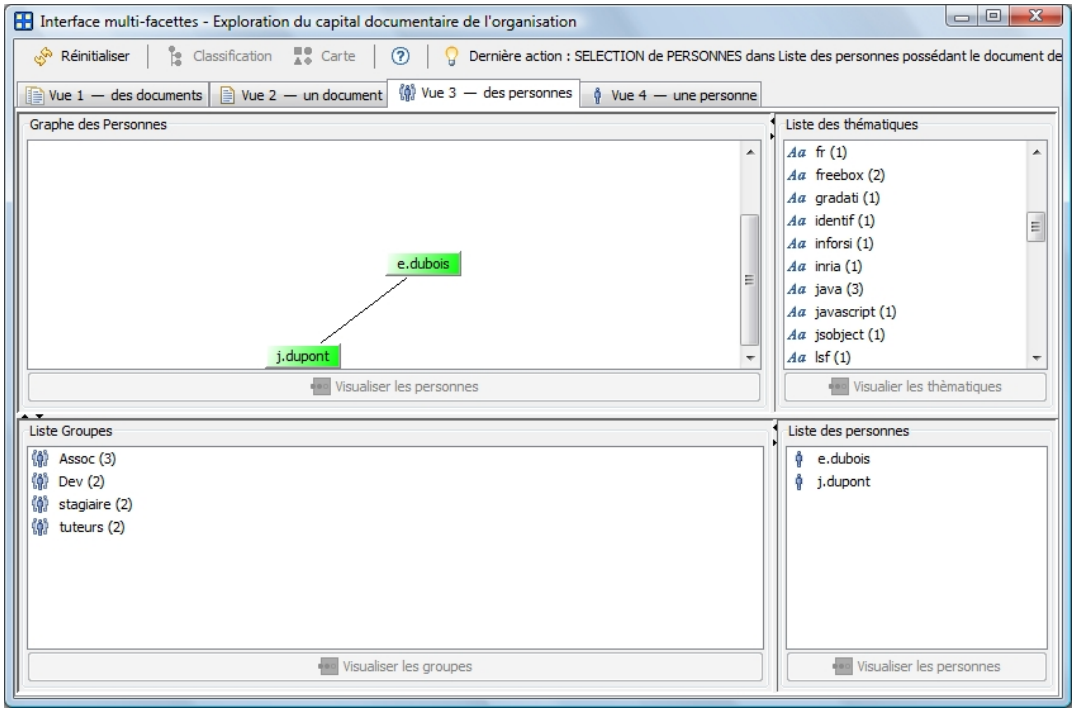


Figure III.4.12 – Vue 3 montrant les deux personnes possédant le document sélectionné.

À l'heure actuelle, l'interface multi-facettes est en phase finale de développement. Aussi, il reste à implanter la fonctionnalité de recherche textuelle dans les vues. Par ailleurs, nous prévoyons d'intégrer une fonctionnalité de zoom dans les graphes combinée à l'affichage d'une minicarte pour voir quelle partie du graphe l'on explore, ces éléments étant d'ores et déjà développés indépendamment (Pallavidino, 2008).

4.4 Discussion et retours d'expérience avec TafAnnote

Lors de la conception de l'architecture proposée, nous avons tenté d'anticiper des problèmes liés au passage à l'échelle. En particulier, la centralisation des données sur un seul serveur l'expose à une charge importante. Afin de limiter cette dernière, nous avons notamment adopté une stratégie d'indexation distribuée. Ainsi, l'indexation n'est pas réalisée par le serveur — approche centralisée qui l'aurait transformé en goulet d'étranglement — mais sur chacun des postes clients. TafAnnote obtient le contenu du document à indexer à partir du navigateur (qui l'a préalablement téléchargé pour l'afficher) puis réalise l'indexation et en transmet le résultat au serveur.

Nous avons entrepris une évaluation informelle de la performance du prototype TafAnnote, en termes d'impact sur le poste client et de charge pour le serveur. Pour le poste client, les paramètres considérés sont la charge du processeur (pour l'indexation) et les transferts réseau (récupération des annotations, des recommandations, envoi des indexes, etc.). Pour le serveur, nous pouvons considérer la taille de la base de données en fonction du nombre d'usagers.

Nous avons installé TafAnnote sur un serveur Sun Fire v440 (mis sur le marché en octobre 2003) doté de 16 Go de RAM, 292 Go de disque dur en SCSI 320 et 4 processeurs cadencés à 1,28 GHz. Ce serveur est extrêmement sollicité pour des expérimentations de chercheurs en Recherche d'Information, ce qui en fait un candidat intéressant reflétant un contexte réaliste d'utilisation. La BD est gérée par un SGBD Oracle 10*g*2. Huit personnes utilisent au quotidien TafAnnote depuis quatre mois, représentant 65 000 documents indexés, soit 400 Mo stockés en base de données. D'après les observations que nous avons réalisées, la charge processeur du serveur est minime ($\leqslant 2\,\%$). Les usagers n'ont pas rapporté un allongement du temps d'affichage des pages sur leur poste avec TafAnnote, moyennant les débits testés (connexions réseau du laboratoire ainsi que haut-débit ADSL à Toulouse et à Lyon). Nous envisageons de compléter ces observations avec un nombre croissant d'usagers et d'en évaluer l'impact sur les modules serveur et client avec des indicateurs adéquats.

4.5 Limites du prototype TafAnnote

La conception et le développement de TafAnnote ont requis la maîtrise de connaissances dans de nombreux domaines de l'informatique dont la modélisation de système d'informations, l'algorithmique, les bases de données et l'interaction homme-machine, principalement. Sa réalisation a fait intervenir de nombreuses technologies et langages : SQL et PL/SQL pour le module serveur implanté avec une base Oracle ; XUL, JavaScript, Java avec JDBC et Swing pour le module client.

La raison d'être de ce prototype est d'implanter l'ensemble des propositions constituant la contribution présentée en partie II. Toutefois, afin de nous concentrer sur les points les plus délicats eu égard au développement (intégration dans Firefox et interface multi-facettes, notamment) nous en avons mis de côté certaines, qui se répartissent en deux catégories :

1. certaines propositions ont déjà fait l'objet d'une implantation dans des travaux antérieurs : les processus RÉORG (Cabanac, 2002; Chevalier, 2002) et RECO (Chevalier, 2002) notamment. Nous envisageons naturellement de les intégrer à TafAnnote ;
2. d'autres ont été simplifiées ou remises à une date ultérieure. Par exemple, le processus PROTODOC qui consiste à créer un proto-document à partir des annotations sélectionnées sera implanté de façon similaire à la recherche dans les annotations (figure III.4.6) qui crée un document à partir des annotations correspondant à la requête de l'usager. Par ailleurs, nous avons restreint les droits d'accès au choix booléen « public » ou « privé » alors que la gestion de ces droits est plus fine dans le modèle unifié (notion de groupe, notamment). Enfin, l'envoi de recommandations manuelles n'est pas encore implanté.

À court terme, nous envisageons d'intégrer à TafAnnote ces éléments qui ne sont ni des problématiques de recherche, ni des défis techniques car ils ne présentent pas de difficulté particulière. Par ailleurs, concernant la version actuelle de TafAnnote, nous avons identifié les limites relatives aux points suivants :

– **formats des documents.** L'implantation d'une fonctionnalité d'annotation est hautement dépendante du format des documents annotables, notamment pour l'expression des points d'ancrage. Certains systèmes d'annotations étudiés en section I.3.2.3 (p. 34) contournent cette difficulté en faisant une capture d'écran du document, le lecteur étant ensuite libre

d'annoter l'image obtenue comme bon lui semble. C'est notamment le cas des systèmes ScreenCrayons (Olsen *et al.*, 2004) et OAS (Harmon, 2007). Toutefois, cette approche ne permet pas de partager les annotations, le système ne stockant que image annotée du document, au lieu de son point d'ancrage. Nous avons choisi l'approche avec point d'ancrage pour pouvoir partager les annotations, mais aussi pour prendre en compte la dimension évolutive des documents. En fait, TafAnnote restitue les annotations sur les passages non impactés par des modifications, dans le cas contraire l'annotation est orpheline ou trompeuse (section I.3.2.2.1) et est restituée en pied de document. Actuellement, TafAnnote ne supporte que l'annotation de documents bien formés, dans le langage HTML. Pour prendre en compte davantage de formats de documents, il faudra concevoir de nouvelles techniques d'ancrage spécifiques, telle que celle que nous avons définie pour annoter les éléments d'un entrepôt de données (Cabanac *et al.*, 2006c,d, 2007d, 2009b).

- **visualisation des annotations.** Actuellement, nous restituons une annotation en créant un élément HTML `span` autour du point d'ancrage spécifié. Toutefois cette stratégie est limitée lorsque le point d'ancrage débute dans un élément HTML et finit dans un autre. Dans ce cas, il faudrait décomposer le `span` original en plusieurs `span` englobant le texte annoté dans chacune des balises concernées par le point d'ancrage. Par ailleurs, cette approche ne permet pas de créer des annotations qui se chevauchent, par exemple : une annotation sur un texte qui est déjà annoté.
- **restrictions dues à la plate-forme cible.** Nous avons retenu le navigateur Firefox car il permet d'y adjoindre des extensions développées en XUL et JavaScript qui sont portables sur tout système. Or, ce n'est pas le cas d'Internet Explorer qui impose un développement avec la technologie COM (Component Object Model) spécifique aux systèmes d'exploitation Microsoft. Afin d'adapter la version actuelle de TafAnnote au navigateur Internet Explorer, nous envisageons d'explorer une alternative, basée sur la prochaine version du JRE 6 (Java Runtime Environment) prévue pour l'automne 2008. Celle-ci devrait implanter une nouvelle spécification de *LiveConnect*[3] qui permettrait d'invoquer du code Java (la couche dialogue du module client) depuis tout navigateur.

Le prototype de recherche TafAnnote est actuellement fonctionnel, il est d'ores et déjà utilisé par des usagers réels pour épauler leurs activités documentaires au quotidien. Nous souhaitons continuer son développement afin d'en améliorer les fonctionnalités existantes, tout en l'utilisant en tant que plate-forme logicielle de validation de nouvelles propositions de recherche.

3. "Formerly Mozilla-specific LiveConnect functionality, such as the ability to call static Java methods, instantiate new Java objects and reference third-party packages from JavaScript, is now available in all browsers. The new LiveConnect specification is forthcoming." cf. `https://jdk6.dev.java.net/plugin2/#LIVECONNECT`.

Conclusion générale

"Work is play when it's something you like."

Andy Warhol (1928 — 1987)

Synthèse des propositions

Dans le contexte des organisations modernes, nous avons exposé les activités documentaires réalisées au quotidien par les « travailleurs du savoir ». Ces activités formant le cycle de vie du document (Sellen et Harper, 2003, p. 203) ont été illustrées sur support papier comme électronique, au travers de nombreux systèmes (logiciels). À la lumière de cet état de l'art, nous avons identifié plusieurs problématiques. D'une part, la grande diversité de systèmes que les usagers doivent maîtriser pour réaliser les six activités documentaires implique un éparpillement de leurs données, ainsi qu'une surcharge cognitive de leur part. D'autre part, les activités documentaires sont cloisonnées et linéaires, alors que les individus se comportent tout autrement. De ce fait, chaque système est très adapté pour une seule activité, mais il ignore les autres. Cette situation conduit à une représentation partielle des usagers (un traitement de texte « connaît » ses usagers en tant que rédacteurs, mais jamais en tant que chercheurs d'information) et à une assistance forcément sous-efficiente. Enfin, les informations introduites dans l'organisation — au prix de coûteux efforts d'élicitation des besoins en information, de recherche, de filtrage, d'analyse, de consolidation, de structuration et de maintenance d'information dans les Espaces Personnels d'Information (EPI) — ne sont jamais valorisées, formant ainsi un capital à haute valeur ajoutée, mais paradoxalement en sommeil.

Afin de proposer une solution originale à ces problématiques, nous avons exposé dans ce mémoire une approche basée sur la fédération des activités documentaires dans une architecture multi-utilisateurs. Son but vise à améliorer les activités quotidiennes de chaque usager, en lui faisant bénéficier du capital de l'organisation et vice versa, sur le principe du donnant-donnant. La conception du système favorise également un enrichissement mutuel des activités : la tâche de navigation, incluse dans l'activité de recherche d'information, exploite le résultat de l'activité de

classement, par exemple. Considérer l'organisation comme un capital à haute valeur ajoutée pour que chacun de ses membres en tire bénéfice est un aspect au cœur de notre contribution. Ainsi, nous exploitons l'ensemble des Espaces Personnels d'Annotations (EPA) de façon non intrusive, afin d'aider chaque individu pour améliorer le retour sur investissement global.

L'architecture que nous proposons est basée sur un concept fédérateur qui est transversal aux activités documentaires : l'annotation collective, en tant que trace de l'activité intellectuelle des travailleurs du savoir. Cette démarche fait suite aux observations de Kidd (1994) confirmées plus récemment par Sellen et Harper (2003, p. 63) : ce sont les annotations des individus qui contiennent la réelle valeur ajoutée du document. Ce modèle est complété par les six processus représentés dans la figure II.5.1 (p. 72). Ils visent à aider l'usager dans les tâches de réorganisation de son EPA (RÉORG), d'exploitation de documents (ADAPTAFFICHAGE), de création de proto-documents à partir des annotations formulées (PROTODOC), de navigation ou de recherche en recommandant des documents provenant d'autres membres organisationnels (NAVI et RECO). Par ailleurs, la VUE UNIFIÉE proposée permet la visualisation et l'exploration de l'organisation selon deux dimensions (les individus et les documents) afin d'en obtenir une meilleure connaissance.

Concernant la validation de la contribution proposée dans ce mémoire, nous avons porté une grande importance au développement du prototype TafAnnote afin de démontrer la faisabilité de nos propositions. Ce dernier prend en charge actuellement une majorité des processus présentés dans ce mémoire ; il est en constant développement pour intégrer l'ensemble des propositions à court terme. Le prototype TafAnnote peut être téléchargé et installé à partir du site Web dédié « `http://www.irit.fr/~Guillaume.Cabanac/TafAnnote` ». Par ailleurs, deux propositions majeures de notre contribution ont fait l'objet de validations scientifiques :

1. nous avons établi un protocole d'expérimentation « écologique » afin de comparer la perception humaine du consensus dans les débats argumentatifs de 121 participants avec les résultats des algorithmes de validation sociale proposés. Les résultats obtenus montrent que les algorithmes approximent la perception humaine dans 80 % des cas ;
2. nous avons comparé deux types de mesure de similarité inter-documents : sur le contenu (classique en Recherche d'Information) *versus* sur l'usage (proposée dans ce mémoire). Les résultats obtenus montrent que ces deux mesures sont statistiquement différentes. Nous exploitons la complémentarité de ces mesures dans l'interface multi-facettes d'accès au capital documentaire de l'organisation, en restituant les deux similarités pour montrer à la fois les liens de thématique et d'usage.

Champs d'application de notre approche

Nous avons focalisé notre approche sur les activités documentaires, notamment dans le contexte du Web. Toutefois, tout ou partie de la contribution présentée dans ce mémoire peut se décliner dans divers autres contextes, où le support électronique remplace le support papier. Pour de tels contextes de nos jours omniprésents, l'annotation électronique collective répond à deux types de préoccupations. D'une part, elle procure les fonctionnalités de commentaire, de mémorisation et de reformulation en contexte, disponibles sur papier et transposées sur support électronique. D'autre part, l'aspect collectif permet le débat en contexte et l'échange d'idées. De

plus, la dématérialisation des annotations offre de nombreuses possibilités d'exploitation de ces traces des activités humaines, prenant dans nos travaux la forme de processus intégrés.

Parmi les champs d'application possibles, les systèmes d'informations permettant la gestion de documents patrimoniaux peuvent bénéficier de nos propositions. En effet, la conservation des traces des lecteurs et de leurs interactions au travers des annotations collectives pourrait offrir à de futurs lecteurs une valeur ajoutée au contenu des documents. Dans le domaine médical, les divers praticiens manipulant les dossiers patient électroniques pourraient y intégrer des remarques et discuter collectivement et en contexte de différents points liés aux informations du dossier (diagnostic, traitement, etc.). Un autre champ d'application concerne la gestion de documentations techniques, dans le domaine de l'aéronautique par exemple. L'annotation pourrait représenter un objet médiateur permettant, entre autres, aux usagers de la documentation de fournir des retours aux concepteurs.

En complément des différents champs d'application mentionnés jusqu'alors, nous avons considéré un contexte moins « orienté document textuel » : celui des systèmes d'aide à la prise de décision. Dans ce contexte, la construction d'entrepôts de données à partir des bases de données opérationnelles de l'organisation vise à aider les usagers à réaliser diverses analyses. Ces dernières sont supportées par des « tables multidimensionnelles » (TM) qui sont des tableaux à double entrée. L'annotation de ces documents électroniques particuliers fait défaut aux systèmes actuels, imposant aux analystes d'imprimer les TM pour les annoter. Or, Foshay *et al.* (2007) soulignent l'intérêt que revêt l'adjonction de contenus provenant des analystes. Afin de répondre au besoin d'annoter les TM et de faire communiquer les analystes en contexte, nous avons étendu l'annotation collective argumentative en « annotation décisionnelle » dans (Cabanac *et al.*, 2006c,d, 2007d, 2009b). Une telle annotation permet par exemple aux décideurs d'annoter les éléments de l'entrepôt de données (TM et schéma de l'entrepôt) afin de conserver la trace d'une prise de décision.

Perspectives de recherche

Outre les divers champs d'application mentionnés précédemment, nous avons identifié de nombreuses perspectives aux travaux présentés dans ce mémoire.

Une première perspective concerne l'évaluation globale de l'architecture proposée. Dans un premier temps, nous envisageons de l'expérimenter avec une équipe de recherche du laboratoire. Le retour d'expérience des enseignants-chercheurs spécialistes de leurs domaines, ainsi que des nouveaux arrivants néophytes (étudiants en stage de master 2 notamment) fournira une première évaluation qualitative, à l'image de celle rapportée par (Millen et Fontaine, 2003). De tels résultats pourront être approfondis par des évaluations quantitatives, évaluant les contraintes imposées par notre application et les gains qui en découlent.

Même si le contexte applicatif d'une organisation limite quelque peu le problème du passage à l'échelle, nous souhaitons améliorer l'architecture client-serveur proposée afin de garantir sa performance dans des contextes d'utilisation davantage contraints tels que le Web. Pour ce faire, nous pourrons considérer l'emploi de réseaux pair-à-pair associés à une grille de calcul pour la gestion des données.

Par ailleurs, afin de généraliser notre approche, nous envisageons de spécifier des techniques

d'ancrage permettant la prise en compte d'un vaste panel de formats de documents. Dans ce domaine, nous avons d'ores et déjà défini une technique d'ancrage sur les éléments d'entrepôts de données (Cabanac *et al.*, 2007d). En prolongement de ces travaux, nous pourrons également considérer la restitution correcte des annotations sur des ressources évolutives (cas des versions de documents notamment) qui nécessitent la recherche de techniques « robustes » d'ancrage.

Au regard de l'adoption de notre approche, nous pourrons envisager d'autres processus pour améliorer les activités documentaires :

- Concernant l'activité ① de Recherche d'Information, nous avons principalement considéré dans ce mémoire la modalité de navigation (processus NAVI et VUE UNIFIÉE). Eu égard à la modalité duale d'interrogation, nous avons identifié dans (Cabanac *et al.*, 2007b) des pistes de recherche utilisant les annotations collectives. D'une part, la phase d'indexation de tout ou partie des documents pourrait exploiter le contenu et les débats des annotations collectives afin d'enrichir la représentation des granules documentaires. D'autre part, les annotations argumentatives collectives pourraient être exploitées en entrée d'un processus de fouille d'opinions (Liu, 2007).
- Concernant l'activité ② de rédaction de document, nous pouvons améliorer le travail collectif en favorisant les modifications en contexte ainsi que les débats associés, contrairement au principe du wiki présentant un espace de discussion externe au document discuté. Cela pourrait encourager les lecteurs à s'impliquer davantage, tout en limitant les « guerres d'annulation » (*revert wars*) observées par Kittur *et al.* (2007) sur Wikipedia, où chaque rédacteur supprime constamment les modifications d'autres contributeurs.

À plus long terme, nous souhaitons poursuivre nos travaux axés sur la relation individus-documents : la fédération des activités documentaires au cœur de notre contribution fournit un cadre de recherche particulièrement riche. L'observation des activités documentaires des usagers permet d'ores et déjà l'enrichissement mutuel de ces activités pour aider chaque usager. En complément, l'étude des interactions entre usagers (contact, envoi d'information, par exemple) permettra d'améliorer la représentation que le système construit des usagers pour leur porter une assistance judicieuse. Une telle représentation pourra être formalisée au travers de profils auxquels nous devrons associer des processus de maintenance afin de s'assurer de la prise en compte effective de l'évolution des intérêts, des pratiques documentaires, des relations sociales, etc. de l'individu modélisé.

Enfin, nous envisageons de relâcher les contraintes de notre cadre d'étude. En effet, le contexte de l'organisation implique la considération d'individus travaillant dans des groupes préétablis : équipes, composantes, etc. Une perspective de recherche consiste, dans un contexte moins structuré, à l'identification de communautés ou de groupes d'intérêt à partir, par exemple, de leurs pratiques documentaires.

Bibliographie

« Le but d'une lecture intelligente est votre instruction. Cela fera mieux que de vous aider à passer le temps ; la lecture changera la nature de vos relations avec autrui ; elle déterminera en vous des perceptions plus rapides, de nouveaux concepts et de nouvelles formes de pensée, car sa fonction principale est de vous éveiller. Et grâce à la lecture vous découvrirez en vous-même et dans le monde des possibilités nouvelles. »

Howard Phillips Lovecraft (1890 — 1937)

Abe, M. et Hori, M. (2003). Robust Pointing by XPath Language: Authoring Support and Empirical Evaluation. *In SAINT'03: Proceedings of the 2003 Symposium on Applications and the Internet*, page 156, Washington, DC, USA. IEEE Computer Society. Cité 1 fois, p. 31.

Abrams, D., Baecker, R. et Chignell, M. (1998). Information Archiving with Bookmarks: Personal Web Space Construction and Organization. *In CHI'98: Proceedings of the conference on Human factors in computing systems*, pages 41–48, New York, NY, USA. ACM Press. Cité 2 fois, p. 21 et 75.

Ackerman, M. S., Wulf, V. et Pipek, V. (2003). *Sharing expertise: Beyond knowledge management*. MIT Press, Cambridge, MA, USA. Cité 1 fois, p. 147.

Adler, A., Gujar, A., Harrison, B. L., O'Hara, K. et Sellen, A. (1998). A Diary Study of Work-Related Reading: Design Implications for Digital Reading Devices. *In CHI'98: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 241–248, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co. Cité 1 fois, p. 10.

Adler, M. J. et van Doren, C. (1972). *How to read a book*. Simon & Shuster, NY. Cité 3 fois, p. 10, 24 et 27.

Agosti, M. (1996). An Overview of Hypertext. *In* Agosti, M. et Smeaton, A. F., éditeurs : *Information Retrieval and Hypertext*, chapitre 2, pages 27–47. Kluwer Academic Publishers, Dordrecht. Cité 2 fois, p. 16 et 75.

Agosti, M. et Ferro, N. (2005). Annotations as Context for Searching Documents. *In CoLIS'05: Proceedings of the 5th International Conference on Conceptions of Library and Information Sciences*, volume 3507 de *LNCS*, pages 155–170. Springer. Cité 1 fois, p. 39.

Agosti, M. et Ferro, N. (2006). Search Strategies for Finding Annotations and Annotated Documents: The FAST Service. *In FQAS'06: Proceedings of the 7th International Conference on Flexible Query Answering Systems*, volume 4027 de *LNCS*, pages 270–281. Springer. Cité 1 fois, p. 100.

Agosti, M. et Ferro, N. (2007). A Formal Model of Annotations of Digital Content. *ACM Trans. Inf. Syst.*, 26(1):3. Cité 3 fois, p. 32, 39 et 100.

Agosti, M., Ferro, N. et Orio, N. (2005). Annotating Illuminated Manuscripts: an Effective Tool for Research and Education. *In JCDL'05: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 121–130, New York, NY, USA. ACM Press. Cité 2 fois, p. 32 et 37.

Azouaou, F., Desmoulins, C. et Mille, D. (2003). Formalismes pour une mémoire de formation à base d'annotations : articuler sémantique implicite et explicite. *In Actes de la conférence EIAH 2003*, pages 43–54, Paris, France. INRP. Cité 1 fois, p. 25.

Baber, C., Cross, J., Yang, F. et Smith, P. (2005). Supporting Shared Analysis for Mobile Investigators. *In* Boujut (2005), pages 11–20. Cité 1 fois, p. 29.

Baeza-Yates, R. A. et Ribeiro-Neto, B. A. (1999). *Modern information retrieval*. ACM Press/Addison-Wesley. Cité 3 fois, p. 54, 75 et 87.

Ballay, J.-F. (2002). Nous sommes tous des travailleurs du savoir. *L'Expansion Management Review*, 107:94–101. Cité 1 fois, p. 8.

Barcellini, F., Détienne, F. et Burkhardt, J.-M. (2007). Annotation et éléments discursifs dans les discussions en ligne de projets *Open Source* : analyse des pratiques de citation électronique. *In* Salembier et Zacklad (2007), chapitre 4, pages 91–108. Cité 1 fois, p. 29.

Bargeron, D., Gupta, A. et Brush, A. J. B. (2001). A Common Annotation Framework. Rapport technique MSR-TR-2001-108, Microsoft Research, Redmond, USA. Cité 3 fois, p. 31, 32 et 37.

Berners-Lee, T., Hendler, J. et Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5):34–43. Cité 1 fois, p. 29.

Boardman, R. et Sasse, M. A. (2004). "Stuff Goes into the Computer and Doesn't Come Out": a Cross-tool Study of Personal Information Management. *In CHI'04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 583–590, New York, NY, USA. ACM. Cité 1 fois, p. 21.

Bottoni, P., Civica, R., Levialdi, S., Orso, L., Panizzi, E. et Trinchese, R. (2004). MADCOW: a Multimedia Digital Annotation System. *In AVI'04: Proceedings of the working conference on Advanced visual interfaces*, pages 55–62, New York, NY, USA. ACM Press. Cité 1 fois, p. 37.

Bottoni, P., Levialdi, S. et Rizzo, P. (2003). An Analysis and Case Study of Digital Annotation. *In* Bianchi-Berthouze, N., éditeur : *DNIS'03: Proceedings of the 3rd International Workshop on Databases in Networked Information Systems*, volume 2822 de *LNCS*, pages 216–230. Cité 1 fois, p. 37.

Bouchet, H. (2004). La cybersurveillance sur les lieux de travail. Rapport technique, CNIL, Paris, France. electronic edition `http://www.ladocumentationfrancaise.fr/rapports-publics/044000175/`. Cité 2 fois, p. 86 et 92.

Boujut, J.-F., éditeur (2005). *Proceedings of the International Workshop on Annotation for Collaboration – Methods, Tools and Practices*. CNRS. Cité 3 fois, p. 142, 143 et 146.

Boujut, J.-F., Darses, F. et Guibert, S. (2007). Etude des annotations en situation collaborative de conception mécanique. *In* Salembier et Zacklad (2007), chapitre 6, pages 127–152. Cité 1 fois, p. 29.

Boulanger, C., Decortis, F. et Leclercq, P. (2007). Annotations et architecture. *In* Salembier et Zacklad (2007), chapitre 5, pages 109–126. Cité 1 fois, p. 29.

Bouthors, V. et Dedieu, O. (1999). Pharos, a Collaborative Infrastructure for Web Knowledge Sharing. *In* Abiteboul, S. et Vercoustre, A.-M., éditeurs : *ECDL'99 : Proceedings of the 3rd European Conference on Research and Advanced Technology for Digital Libraries*, volume 1696 de *LNCS*, pages 215–233. Springer. Cité 3 fois, p. 32, 34 et 36.

Bouvin, N. O. (1999). Unifying Strategies for Web Augmentation. *In HYPERTEXT'99: Proceedings of the tenth ACM Conference on Hypertext and hypermedia : returning to our diverse roots*, pages 91–100, New York, NY, USA. ACM. Cité 1 fois, p. 36.

Bouvin, N. O., Zellweger, P. T., Grønbæk, K. et Mackinlay, J. D. (2002). Fluid annotations through open hypermedia: using and extending emerging web standards. *In WWW'02: Proceedings of the 11th international conference on World Wide Web*, pages 160–171, New York, NY, USA. ACM Press. Cité 5 fois, p. 31, 33, 52, 72 et 153.

Boyer, M., Canut, M.-F., Chevalier, M., Péninou, A. et Sèdes, F. (2007). Cartographie de l'organisation : une approche topologique des connaissances. *In EGC'07 : actes des 7e journées Extraction et Gestion des Connaissances*, volume RNTI-E-9 de *Revue des Nouvelles Technologies de l'Information*, pages 557–568. Cépaduès. Cité 3 fois, p. 20, 81 et 83.

Bringay, S., Barry, C. et Charlet, J. (2004). Les documents et les annotations du dossier patient hospitalier. *Information - Interaction - Intelligence*, 4(1):191–211. Cité 1 fois, p. 25.

Bringay, S., Barry, C. et Charlet, J. (2005). A specific tool of Annotations for the Electronic Health Record. *In* Boujut (2005), pages 21–30. Cité 1 fois, p. 37.

Bringay, S., Barry, C. et Charlet, J. (2007). Un modèle pour les annotations du dossier patient informatisé. *In* Salembier et Zacklad (2007), chapitre 2, pages 47–68. Cité 1 fois, p. 29.

Brush, A. J. B. (2002). Annotating Digital Documents for Asynchronous Collaboration. Technical report 02-09-02, Department of Computer Science and Engineering, University of Washington, USA. Cité 2 fois, p. 31 et 36.

Brush, A. J. B., Bargeron, D., Grudin, J., Borning, A. et Gupta, A. (2002). Supporting Interaction Outside of Class: Anchored Discussions vs. Discussion Boards. *In CSCL'02: Proceedings of the international conference on Computer support for collaborative learning*, pages 425–434. Cité 1 fois, p. 36.

Brust, M. R. et Rothkugel, S. (2007). On Anomalies in Annotation Systems. *In AICT'07: Proceedings of the 3rd International Conference on Telecommunications*, page 3 (8 pp). IEEE. Cité 1 fois, p. 38.

Cabanac, G. (2002). Interface de classification de signets Web. Rapport de stage de DUT, IRIT, Université Toulouse 3, France. Stage encadré par Max Chevalier et Christine Julien, `ftp://ftp.irit.fr/IRIT/SIG/2002_DUT_C.pdf`**.** Cité 2 fois, p. 75 et 134.

Cabanac, G. (2005). Annotation de ressources électroniques sur le Web : formes et usages. Rapport de Master 2 Recherche, IRIT, Université Toulouse 3, France. `ftp://ftp.irit.fr/IRIT/SIG/2005_M2R_C.pdf`**.** Cité 3 fois, p. 37, 123 et 124.

Cabanac, G. (2008a). Annotation collective dans le contexte RI : définition d'une plate-forme pour expérimenter la validation sociale. ***In CORIA/RJCRI'08 : 3e Rencontres Jeunes Chercheurs en Recherche d'Information*****, pages 385–392. Université de Rennes 1.** Cité 1 fois, p. 99.

Cabanac, G. (2008b). Interface multi-facettes d'accès au capital documentaire de l'organisation. ***In INFORSID'08 : 26e congrès de l'INFormatique des Organisations et Systèmes d'Information et de Décision*****, pages 69–84. Éditions Inforsid.** Cité 1 fois, p. 79.

Cabanac, G., Chevalier, M., Chrisment, C. et Julien, C. (2005). A Social Validation of Collaborative Annotations on Digital Documents. ***In*** **Boujut (2005), pages 31–40.** Cité 1 fois, p. 57.

Cabanac, G., Chevalier, M., Chrisment, C. et Julien, C. (2006a). L'architecture CoMED pour la gestion collective de documents électroniques dans l'organisation. ***In*** **Zreik, K. et Vanoirbeek, C., éditeurs :** ***CIDE'06 : 9e Colloque International sur le Document Électronique*****, pages 237–252, Paris, France. Europia.** Cité 1 fois, p. 51.

Cabanac, G., Chevalier, M., Chrisment, C. et Julien, C. (2006b). Validation sociale d'annotations collectives : argumentation bipolaire graduelle pour la théorie sociale de l'information. ***In INFORSID'06 : 24e congrès de l'INFormatique des Organisations et Systèmes d'Information et de Décision*****, pages 467–482. Éditions Inforsid.** Cité 2 fois, p. 57 et 61.

Cabanac, G., Chevalier, M., Chrisment, C. et Julien, C. (2007a). An Original Usage-based Metrics for Building a Unified View of Corporate Documents. ***In*** **Wagner, R., Revell, N. et Pernul, G., éditeurs :** ***DEXA'07: Proceedings of the 18th International Conference on Database and Expert Systems Applications*****, volume 4653 de** ***LNCS*****, pages 202–212. Springer.** Cité 5 fois, p. 65, 80, 88, 117 et 154.

Cabanac, G., Chevalier, M., Chrisment, C. et Julien, C. (2007b). Collective Annotation: Perspectives for Information Retrieval Improvement. ***In RIAO'07: Proceedings of the 8th conference on Information Retrieval and its Applications*****. CID.** Cité 4 fois, p. 48, 57, 61 et 140.

Cabanac, G., Chevalier, M., Chrisment, C. et Julien, C. (2007c). Valoriser et intégrer les activités documentaires de l'organisation grâce à l'annotation collective de documents électroniques. ***In VSST'07 : actes du 5e colloque Veille Stratégique, Scientifique & Technologique.*** Cité 1 fois, p. 51.

Cabanac, G., Chevalier, M., Chrisment, C. et Julien, C. (2008a). Exploiting the Annotation Practice for Personal and Collective Information Management. ***In INFORSID/PeCUSI'08 : 2e atelier Prise en Compte de l'Usager dans les Systèmes d'Information*****, pages 55–66. Inforsid.** Cité 1 fois, p. 51.

Cabanac, G., Chevalier, M., Chrisment, C. et Julien, C. (2008b). Exploiting the Annotation Practice for Personal and Collective Information Management. ***In*** **Ebersold, S., Front, A., Lopistéguy, P. et Nurcan, S., éditeurs :** ***CAiSE/MoDISE-EUS'08: International Workshop on Model Driven Information Systems Engineering: Enterprise, User and System Models*****, volume 341 de** ***CEUR Workshop Proceedings*****, pages 67–78. CEUR-WS.** Cité 1 fois, p. 51.

Cabanac, G., Chevalier, M., Chrisment, C. et Julien, C. (2009a). Visualisation et exploration du capital documentaire d'une organisation au travers d'une interface multi-facettes. ***Ingénierie des Systèmes d'Information*****. à paraître.** Cité 1 fois, p. 80.

Cabanac, G., Chevalier, M., Chrisment, C., Julien, C., Soulé-Dupuy, C. et Tchienehom, P. (2008c). Web Information Retrieval: Towards Social Information Search Assistants. ***In*** **Kidd, T. et Chen, I., éditeurs :** ***Social Information Technology: Connecting Society and Cultural Issues*****, chapitre 16, pages 218–252. IGI Global.** `http://www.igi-global.com/downloads/pdf/KC218.pdf`**.** Cité 1 fois, p. 16.

Cabanac, G., Chevalier, M., Ravat, F. et Teste, O. (2006c). Modèle conceptuel pour bases de données multidimensionnelles annotées. ***In*** **Ritschard, G. et Djeraba, C., éditeurs :** ***EGC*****, volume RNTI-E-6 de** ***Revue des Nouvelles Technologies de l'Information*****, pages 119–124. Cépaduès.** Cité 2 fois, p. 135 et 139.

Cabanac, G., Chevalier, M., Ravat, F. et Teste, O. (2006d). Méta-modélisation des bases de données multidimensionnelles annotées. ***In Revue des Nouvelles Technologies de l'Information (RNTI-B-2) - Entrepôts de Données et Analyse en ligne (EDA'06)*****, pages 39–54, Toulouse, France. Cépaduès.** Cité 2 fois, p. 135 et 139.

Cabanac, G., Chevalier, M., Ravat, F. et Teste, O. (2007d). An Annotation Management System for Multidimensional Databases. ***In*** **Song, I.-Y., Eder, J. et Nguyen, T. M., éditeurs :** ***DaWaK'07: Proceedings of the 9th International Conference on Data Warehousing and Knowledge Discovery*****, volume 4654 de** ***LNCS*****, pages 89–98. Springer.** Cité 3 fois, p. 135, 139 et 140.

Cabanac, G., Chevalier, M., Ravat, F. et Teste, O. (2009b). Decisional Annotations: Integrating and Preserving Decision-Makers' Expertise in Multidimensional Systems. ***In*** **Nguyen, T. M., éditeur :** ***Complex Data Warehousing and Knowledge Discovery for Advanced Retrieval Development: Innovative Methods and Applications*****, Advances in Data Warehousing and Mining, chapitre 4. IGI Global. à paraître.** Cité 2 fois, p. 135 et 139.

Cadiz, J. J., Gupta, A. et Grudin, J. (2000). Using Web Annotations for Asynchronous Collaboration Around Documents. *In CSCW'00: Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 309–318, New York, NY, USA. ACM Press. Cité 3 fois, p. 31, 32 et 36.

Carter, S., Churchill, E., Denoue, L., Helfman, J. et Nelson, L. (2004). Digital Graffiti: Public Annotation of Multimedia Content. *In CHI'04: CHI'04 extended abstracts on Human factors in computing systems*, pages 1207–1210, New York, NY, USA. ACM. Cité 1 fois, p. 37.

Catlin, T., Bush, P. et Yankelovich, N. (1989). InterNote: Extending a Hypermedia Framework to Support Annotative Collaboration. *In HYPERTEXT'89: Proceedings of the second annual ACM conference on Hypertext*, pages 365–378, New York, NY, USA. ACM. Cité 1 fois, p. 36.

Cayrol, C. et Lagasquie-Schiex, M.-C. (2004). Bipolarité en argumentation. Report 2004-07-R, IRIT, Toulouse, France. Cité 1 fois, p. 100.

Cayrol, C. et Lagasquie-Schiex, M.-C. (2005a). Gradual Valuation for Bipolar Argumentation Frameworks. *In* Godo, L., éditeur : *ECSQARU'05: Proceedings of the 8th European Conference on Symbolic and Qualitative Approaches to Reasoning and Uncertainty*, volume 3571 de *LNCS*, pages 366–377. Springer. Cité 5 fois, p. 57, 59, 60, 61 et 108.

Cayrol, C. et Lagasquie-Schiex, M.-C. (2005b). Graduality in Argumentation. *J. Artif. Intell. Res.*, 23:245–297. Cité 2 fois, p. 57 et 59.

Chen, C. (2006). *Information visualization: Beyond the horizon.* Springer, 2nd édition. Cité 2 fois, p. 19 et 88.

Chevalier, M. (2002). *Interface adaptative pour l'aide à la recherche d'information sur le Web*. Thèse de doctorat, Université Toulouse 3, France. Cité 6 fois, p. 65, 74, 75, 76, 97 et 134.

Chevalier, M. et Julien, C. (2003). Interface adaptative et coopérative pour l'aide à la Recherche d'Information sur le Web. *Information - Interaction - Intelligence*, 3(2):47–73. Cité 1 fois, p. 74.

Chevalier, M., Julien, C. et Soulé-Dupuy, C. (2008). Profils usagers pour la recherche d'information — Pertinence de leur usage ? *In* Lopistéguy et Tricot (2008), pages 81–93. Cité 2 fois, p. 18 et 93.

Chevalier, M. et Verlhac, M. (2000). ISIDOR: A Visualisation Interface for Advanced Information Retrieval. *In ICEIS'00: Proceedings of the 2th International Conference on Enterprise Information Systems*, pages 414–418. Cité 1 fois, p. 20.

Churchill, E. F., Trevor, J., Bly, S., Nelson, L. et Cubranic, D. (2000). Anchored Conversations: Chatting in the Context of a Document. *In CHI'00: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 454–461, New York, NY, USA. ACM Press. Cité 1 fois, p. 36.

Ciaccia, A. (2008). Recherche d'informations sur le Web — Analyse cognitive du rôle des connaissances de l'utilisateur. *In* Lopistéguy et Tricot (2008), pages 67–79. Cité 1 fois, p. 16.

Clark, J. et DeRose, S., éditeurs (1999). *XML Path Language (XPath) – Version 1.0*. W3C. Cité 1 fois, p. 30.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.*, 20:37–46. Cité 2 fois, p. 56 et 57.

Davis, J. R. et Huttenlocher, D. P. (1994). Conote: small group annotation experiment. `http://web.archive.org/web/19990422160552/http:/dri.cornell.edu/pub/davis/annotation.html`. Cité 1 fois, p. 32.

Davis, J. R. et Huttenlocher, D. P. (1995). Shared annotation for cooperative learning. *In CSCL'95: The first international conference on Computer support for collaborative learning*, pages 84–88, Mahwah, NJ, USA. Lawrence Erlbaum Associates, Inc. Cité 1 fois, p. 36.

de Tocqueville, A. (1835). *La Démocratie en Amérique*, volume 1. deuxième partie. Cité 1 fois, p. 62.

Denjean, P. (1989). *Interrogation d'un Système Vidéotex Arborescent : l'indexation des textes*. Thèse de doctorat, Université Toulouse 3. Cité 1 fois, p. 54.

Denoue, L. (2000). *De la création à la capitalisation des annotations dans un espace personnel d'informations*. Thèse de doctorat, Université de Savoie, France. Cité 4 fois, p. 26, 28, 30 et 33.

Denoue, L. (2005). Yawas for Firefox. `http://lists.w3.org/Archives/Public/www-annotation/2005JanJun/0010.html`. Cité 1 fois, p. 37.

Denoue, L. et Vignollet, L. (2000a). An annotation tool for Web browsers and its applications to information retrieval. *In RIAO'00: Proceedings of the 6th international conference on Information Retrieval and its Applications*, pages 180–196, Paris, France. CID. Cité 1 fois, p. 37.

Denoue, L. et Vignollet, L. (2000b). L'importance des annotations – Application à la classification des documents du Web. *Document numérique*, 4(1-2):37–57. Cité 2 fois, p. 32 et 34.

DeRose, S., Daniel, R., Grosso, P., Maler, E., Marsh, J. et Walsh, N. (2002). *XML Pointer Language (XPointer)*. W3C. Cité 3 fois, p. 30, 49 et 52.

Desmontils, E., Jacquin, C. et Simon, L. (2004). Dinosys: An Annotation Tool for Web-Based Learning. *In* Liu, W., Shi, Y. et Li, Q., éditeurs : *ICWL'04*, volume 3143 de *LNCS*, pages 59–66. Springer. Cité 1 fois, p. 37.

Dmitriev, P. A., Eiron, N., Fontoura, M. et Shekita, E. (2006). Using Annotations in Enterprise Search. *In WWW'06: Proceedings of the 15th international conference on World Wide Web*, pages 811–817, New York, NY, USA. ACM Press. Cité 1 fois, p. 17.

Donin, N. et Theureau, J. (2007). Annotation de la partition par le musicien et (re)distribution de son attention en situation de répétition. *In* Salembier et Zacklad (2007), chapitre 8, pages 173–204. Cité 1 fois, p. 29.

Doumat, R., Egyed-Zsigmong, E. et Pinon, J.-M. (2008). Réutilisation des traces d'utilisation dans un système d'annotation de manuscrits. *In* Lopistéguy et Tricot (2008), pages 43–53. Cité 1 fois, p. 37.

Dousset, B. (2003). *Intégration de méthodes interactives de découverte de connaissances pour la veille stratégique*. Habilitation à diriger des recherches, Université Toulouse 3, France. Cité 1 fois, p. 20.

Drucker, P. F. (1959). *Landmarks of tomorrow: A report on the new "post-modern" world*. Transaction. Cité 1 fois, p. 7.

Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in a nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.*, 77:321–357. Cité 2 fois, p. 59 et 60.

Durand Degranges, C. (2003). Célébrations narcissiques. *WebLettres*, 103. `http://www.weblettres.net/spip/article.php3?id_article=113`. Cité 1 fois, p. 24.

Eades, P. (1984). A Heuristic for Graph Drawing. *Congressus Numerantium*, 42:149–160. Cité 1 fois, p. 88.

Fekete, J.-D. et Plaisant, C. (2002). Interactive Information Visualization of a Million Items. *In INFOVIS'02: Proceedings of the IEEE Symposium on Information Visualization*, page 117, Washington, DC, USA. IEEE Computer Society. Cité 2 fois, p. 19 et 153.

Feldman, S. (2004). The high cost of not finding information. *KM World magazine*, 13(3):electronic edition `http://www.kmworld.com/Articles/PrintArticle.aspx?ArticleID=9534`. Cité 4 fois, p. 16, 17, 22 et 74.

Fleiss, J. L. (1971). Measuring Nominal Scale Agreement Among Many Raters. *Psychol. Bull.*, 76(5):378–382. Cité 2 fois, p. 57 et 114.

Fleiss, J. L., Levin, B. et Paik, M. C. (2003). The Measurement of Interrater Agreement. *In* Fleiss, J. L., Levin, B. et Paik, M. C., éditeurs : *Statistical Methods for Rates and Proportions*, chapitre 18, pages 598–626. John Wiley & Sons, Inc., 3 édition. Cité 2 fois, p. 57 et 157.

Fogli, D., Fresta, G. et Mussio, P. (2004). On Electronic Annotation and Its Implementation. *In AVI'04: Proceedings of the working conference on Advanced visual interfaces*, pages 98–102, New York, NY, USA. ACM Press. Cité 1 fois, p. 37.

Fogli, D., Fresta, G., Mussio, P., Marcante, A. et Padula, M. (2005). Annotation in cooperative work: from paper-based to the web one. *In* Boujut (2005), pages 1–10. Cité 1 fois, p. 29.

Foshay, N., Mukherjee, A. et Taylor, A. (2007). Does data warehouse end-user metadata add value? *Commun. ACM*, 50(11):70–77. Cité 1 fois, p. 139.

Fraenkel, A. S. et Klein, S. T. (1999). Information Retrieval from Annotated Texts. *J. Am. Soc. Inf. Sci.*, 50(10): 845–854. Cité 3 fois, p. 24, 39 et 76.

Freyne, J., Farzan, R., Brusilovsky, P., Smyth, B. et Coyle, M. (2007). Collecting Community Wisdom: Integrating Social Search & Social Navigation. *In IUI'07: Proceedings of the 12th international conference on Intelligent user interfaces*, pages 52–61, New York, NY, USA. ACM Press. Cité 1 fois, p. 37.

Frommholz, I. et Fuhr, N. (2006). Probabilistic, Object-oriented Logics for Annotation-based Retrieval in Digital Libraries. *In JCDL'06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 55–64, New York, NY, USA. ACM Press. Cité 2 fois, p. 39 et 100.

Fruchterman, T. M. J. et Reingold, E. M. (1991). Graph Drawing by Force-directed Placement. *Softw. Pract. Exper.*, 21(11):1129–1164. Cité 1 fois, p. 88.

Furnas, G. W., Landauer, T. K., Gomez, L. M. et Dumais, S. T. (1987). The Vocabulary Problem in Human-System Communication. *Commun. ACM*, 30(11):964–971. Cité 2 fois, p. 18 et 76.

Furnas, G. W. et Zacks, J. (1994). Multitrees: Enriching and Reusing Hierarchical Structure. *In CHI'94: Conference companion on Human factors in computing systems*, page 223, New York, NY, USA. ACM Press. Cité 1 fois, p. 66.

Gabrielli, S. et Law, A. (2003). Annotation in the Wild: Benefits of Linking Paper to Digital Media. *In CHI'03: CHI'03 extended abstracts on Human factors in computing systems*, pages 890–891, New York, NY, USA. ACM. Cité 1 fois, p. 37.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin, Boston. Cité 1 fois, p. 9.

Golovchinsky, G., Price, M. N. et Schilit, B. N. (1999). From Reading to Retrieval: Freeform Ink Annotations as Queries. *In SIGIR'99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25, New York, NY, USA. ACM Press. Cité 1 fois, p. 39.

GRAHL software (2004). Pdf annotator. http://www.ograhl.com/en/pdfannotator. Cité 1 fois, p. 37.

Greengard, S. (1999). Getting rid of the paper chase. *Workforce*, 78(11):69. Cité 1 fois, p. 8.

Grønbæk, K., Sloth, L. et Ørbæk, P. (1999). Webvise: Browser and Proxy Support for Open Hypermedia Structuring Mechanisms on the World Wide Web. *Comp. Netw.*, 31(11-16):1331–1345. Cité 1 fois, p. 36.

Guzdial, M., Rick, J. et Kerimbaev, B. (2000). Recognizing and Supporting Roles in CSCW. *In CSCW'00: Proceedings of the conference on Computer supported cooperative work*, pages 261–268, New York, NY, USA. ACM Press. Cité 1 fois, p. 17.

Hammond, T., Hannay, T. et Lund, B. (2004). The Role of RSS in Science Publishing: Syndication and Annotation on the Web. *D-Lib Magazine*, 10(12):electronic edition http://dx.doi.org/10.1045/december2004--hammond. Cité 2 fois, p. 53 et 74.

Hammond, T., Hannay, T., Lund, B. et Scott, J. (2005). Social Bookmarking Tools (I): A General Review. *D-Lib Magazine*, 11(4):electronic edition http://dx.doi.org/10.1045/april2005--hammond. Cité 4 fois, p. 16, 18, 49 et 53.

Hansen, F. A. (2006). Ubiquitous annotation systems: technologies and challenges. *In HYPERTEXT'06: Proceedings of the 17th conference on Hypertext and hypermedia*, pages 121–132, New York, NY, USA. ACM Press. Cité 1 fois, p. 39.

Harmon, T. (2007). Web-based Annotation and Collaboration. *In WEBIST'07: Proceedings of the 3rd International Conference on Web Information Systems and Technologies*. INSTICC. Cité 3 fois, p. 29, 37 et 135.

Heck, R. M. et Luebke, S. M. (1999). HyperPass: An Annotation System for the Classroom. Cité 2 fois, p. 31 et 36.

Heck, R. M., Luebke, S. M. et Obermark, C. H. (1999). A Survey of Web Annotation Systems. http://www.math.grin.edu/~rebelsky/Blazers/Annotations/Summer1999/Papers/survey_paper.html. Cité 1 fois, p. 36.

Hendricksen, C. (1997). DocReview: A Document Reviewing Tool for the WWW. http://depts.washington.edu/bkn/public/PURL/info.html. Cité 1 fois, p. 36.

Herman, I., Melançon, G. et Marshall, M. S. (2000). Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):24–43. Cité 2 fois, p. 19 et 88.

Hertzum, M. et Pejtersen, A. M. (2000). The information-seeking practices of engineers: searching for documents as well as for people. *Inf. Process. Manage.*, 36(5):761–778. Cité 2 fois, p. 80 et 83.

Hinds, P. J. et Pfeffer, J. (2003). Why Organizations Don't "Know What They Know": Cognitive and Motivational Factors Affecting the Transfer of Expertise. *In* Ackerman *et al.* (2003), chapitre 1, pages 3–26. Cité 2 fois, p. 17 et 79.

Holtman, K. (1996). The Futplex System. *In Proceedings of the ERCIM workshop on CSCW and the Web*, page 3, Sankt Augustin, Germany. GMD. Cité 1 fois, p. 36.

Hull, D. (1993). Using Statistical Testing in the Evaluation of Retrieval Experiments. *In SIGIR'93: Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 329–338, New York, NY, USA. ACM Press. Cité 3 fois, p. 109, 110 et 119.

Huysman, M. et de Wit, D. (2003). A Critical Evaluation of Knowledge Management Practices. *In* Ackerman *et al.* (2003), chapitre 2, pages 27–55. Cité 1 fois, p. 71.

iMarkup Solutions Inc. (2000). imarkup client. http://www.imarkup.com/products/imarkup_client.asp. Cité 1 fois, p. 36.

Jackson, H. J. (2002). *Marginalia: Readers writing in books*. Yale University Press. Cité 3 fois, p. 23, 24 et 27.

Jaczynski, M. et Trousse, B. (1998). WWW Assisted Browsing by Reusing Past Navigations of a Group of Users. *In* Smyth, B. et Cunningham, P., éditeurs : *EWCBR*, volume 1488 de *LNCS*, pages 160–171. Springer. Cité 1 fois, p. 67.

Jardine, N. et van Rijsbergen, C. J. (1971). The Use of Hierarchic Clustering in Information Retrieval. *Information Storage and Retrieval*, 7(5):217–240. Cité 2 fois, p. 75 et 89.

Johnson, B. et Shneiderman, B. (1991). Tree-maps: a space-filling approach to the visualization of hierarchical information structures. *In VIS'91: Proceedings of the 2nd conference on Visualization*, pages 284–291, Los Alamitos, CA, USA. IEEE Computer Society Press. Cité 1 fois, p. 20.

Jones, W. (2007). How People Keep and Organize Personal Information. *In* Jones et Teevan (2007b), chapitre 3, pages 35–56. Cité 4 fois, p. 11, 21, 70 et 75.

Jones, W., Phuwanartnurak, A. J., Gill, R. et Bruce, H. (2005). Don't Take My Folders Away!: Organizing Personal Information to Get Things Done. *In CHI'05 extended abstracts on Human factors in computing systems*, pages 1505–1508, New York, NY, USA. ACM Press. Cité 3 fois, p. 21, 50 et 70.

Jones, W. et Teevan, J. (2007a). Introduction. *In* Jones et Teevan (2007b), chapitre 1, pages 3–20. Cité 1 fois, p. 11.

Jones, W. et Teevan, J. (2007b). *Personal information management*. University of Washington Press, WA, USA. Cité 4 fois, p. 11, 148, 149 et 151.

Kahan, J., Koivunen, M.-R., Prud'Hommeaux, E. et Swick, R. R. (2002). Annotea: an open RDF infrastructure for shared Web annotations. *Comp. Netw.*, 32(5):589–608. Cité 11 fois, p. 30, 31, 32, 33, 34, 37, 38, 49, 55, 73 et 153.

Karacapilidis, N. et Papadias, D. (2001). Computer supported argumentation and collaborative decision making : the HERMES system. *Inf. Syst.*, 26(4):259–277. Cité 1 fois, p. 59.

Karouach, S. (2003). *Visualisations interactives pour la découverte de connaissances : concepts, méthodes et outils*. Thèse de doctorat, Université Toulouse 3, France. Cité 1 fois, p. 20.

Kaye, J. J., Vertesi, J., Avery, S., Dafoe, A., David, S., Onaga, L., Rosero, I. et Pinch, T. (2006). To Have and to Hold: Exploring the Personal Archive. *In CHI'06: Proceedings of the conference on Human Factors in computing systems*, pages 275–284, New York, NY, USA. ACM Press. Cité 2 fois, p. 12 et 89.

Keller, R. M., Wolfe, S. R., Chen, J. R., Rabinowitz, J. L. et Mathe, N. (1997). A bookmarking service for organizing and sharing URLs. *Comput. Netw. ISDN Syst.*, 29(8-13):1103–1114. Cité 1 fois, p. 34.

Khoo, C. S., Luyt, B., Ee, C., Osman, J., Lim, H.-H. et Yong, S. (2007). How users organize electronic files on their workstations in the office environment: a preliminary study of personal information organization behaviour. *Information Research*, 11(2):electronic edition `http://informationr.net/ir/12--2/paper293.html`. Cité 2 fois, p. 21 et 70.

Kidd, A. (1994). The marks are on the knowledge worker. *In CHI'94: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 186–191, New York, NY, USA. ACM. Cité 11 fois, p. 1, 3, 7, 10, 12, 19, 21, 24, 47, 48 et 138.

Kittur, A., Suh, B., Pendleton, B. A. et Chi, E. H. (2007). He Says, She Says: Conflict and Coordination in Wikipedia. *In CHI'07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 453–462, New York, NY, USA. ACM. Cité 1 fois, p. 140.

Klas, C.-P. et Fuhr, N. (2000). A new Effective Approach for Categorizing Web Documents. *In Proceedings of the 22th BCS-IRSG Colloquium on IR Research*. Cité 2 fois, p. 75 et 87.

Kleiner, I. (2000). From Fermat to Wiles: Fermat's Last Theorem Becomes a Theorem. *Elemente der Mathematik*, 55(1):19–37. Cité 1 fois, p. 24.

Kohonen, T. (2001). *Self-organizing maps*. Springer-Verlag, Secaucus, NJ, USA, 3rd édition. Cité 2 fois, p. 20 et 89.

Lagus, K., Kaski, S. et Kohonen, T. (2004). Mining massive document collections by the WEBSOM method. *Inf. Sci.*, 163(1-3):135–156. Cité 2 fois, p. 20 et 153.

LaLiberte, D. et Braverman, A. (1995). A protocol for scalable group and public annotations. *In Proceedings of the Third International World-Wide Web conference on Technology, tools and applications*, pages 911–918, New York, NY, USA. Elsevier North-Holland, Inc. Cité 1 fois, p. 36.

Lanagan, J. et Smeaton, A. F. (2007). SportsAnno: What Do You Think? *In RIAO'2007: Proceedings of the 8th conference on Information Retrieval and its Applications*. CID. Cité 1 fois, p. 37.

Landis, J. R. et Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174. Cité 2 fois, p. 57 et 157.

Liu, B. (2007). Opinion Mining. *In* Liu, B., éditeur : *Web data mining: Exploring hyperlinks, contents, and usage data*, Data-Centric Systems and Applications, chapitre 11, pages 411–447. Springer. Cité 2 fois, p. 49 et 140.

Lober, W. B., Trigg, L. J., Bliss, D. et Brinkley, J. M. (2001). IML: An image markup language. *In Proceedings of the American Medical Informatics Association Symposium*, pages 403–407. Cité 1 fois, p. 36.

Lopistéguy, P. et Tricot, A., éditeurs (2008). *INFORSID/PeCUSI'08 : 2e atelier Prise en Compte de l'Usager dans les Systèmes d'Information*. Éditions Inforsid. Cité 1 fois, p. 145.

Lortal, G., Lewkowicz, M. et Todirascu-Courtier, A. (2006). Annotations: A Way to Capture Experience. *In* Gabrys, B., Howlett, R. J. et Jain, L. C., éditeurs : *KES'06: Proceedings of the 10th International Conference on Knowledge-Based Intelligent Information and Engineering Systems*, volume 4251 de *LNCS*, pages 1131–1138. Springer. Cité 1 fois, p. 37.

Lortsch, D. (1910). *Histoire de la Bible en France et fragments relatifs à l'histoire générale de la Bible*. Société biblique britannique et étrangère, Paris. Cité 1 fois, p. 24.

Lund, B., Hammond, T., Flack, M. et Hannay, T. (2005). Social Bookmarking Tools (II): A Case Study – Connotea. *D-Lib Magazine*, 11(4):electronic edition `http://dx.doi.org/10.1045/april2005--lund`. Cité 3 fois, p. 16, 18 et 37.

Lutters, W. G., Ackerman, M. S. et Zhou, X. (2007). Group Information Management. *In* Jones et Teevan (2007b), chapitre 14, pages 236–248. Cité 1 fois, p. 92.

Ma, F. et Murota, M. (2006). Development of Web Browser Extension for Cross Sectional Search of History, Bookmarks and ScrapBook. *In Proceedings of the 22th Annual Conference of Japan Society for Educational Technology*, pages 459–460. Cité 1 fois, p. 37.

Maarek, Y. S. et Ben-Shaul, I. (1996). Automatically Organizing Bookmarks per Contents. *Computer Networks and ISDN Systems*, 28(7-11):1321–1333. Cité 2 fois, p. 75 et 89.

Manning, C. D., Raghavan, P. et Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press. Cité 2 fois, p. 54 et 75.

Marais, H. et Bharat, K. (1997). Supporting cooperative and personal surfing with a desktop assistant. *In UIST'97: Proceedings of the 10th annual ACM symposium on User interface software and technology*, pages 129–138, New York, NY, USA. ACM. Cité 1 fois, p. 34.

Margolis, M. et Resnick, P. (1999). Third Voice: Vox Populi Vox Dei? *First Monday*, 4(10). Cité 3 fois, p. 30, 32 et 36.

Marshall, C. C. (1997). Annotation: from paper books to the digital library. *In DL'97: Proceedings of the second ACM international conference on Digital libraries*, pages 131–140, New York, NY, USA. ACM. Cité 3 fois, p. 24, 26 et 27.

Marshall, C. C. (1998). Toward an ecology of hypertext annotation. *In Hypertext'98: Proceedings of the 9th conference on Hypertext and hypermedia*, pages 40–49, New York, NY, USA. ACM Press. Cité 10 fois, p. 9, 10, 24, 26, 27, 28, 29, 39, 49 et 63.

Marshall, C. C. et Brush, A. J. B. (2004). Exploring the relationship between personal and public annotations. *In JCDL'04: Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, pages 349–357, New York, NY, USA. ACM Press. Cité 1 fois, p. 39.

Mille, D. (2005). *Modèles et outils logiciels pour l'annotation sémantique de documents pédagogiques*. Thèse de doctorat, Université Joseph Fournier, Grenoble, France. Cité 1 fois, p. 25.

Millen, D. R., Feinberg, J. et Kerr, B. (2006). Dogear: Social Bookmarking in the Enterprise. *In CHI'06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 111–120, New York, NY, USA. ACM Press. Cité 1 fois, p. 18.

Millen, D. R. et Fontaine, M. A. (2003). Improving Individual and Organizational Performance through Communities of Practice. *In GROUP'03: Proceedings of the international conference on Supporting group work*, pages 205–211, New York, NY, USA. ACM Press. Cité 1 fois, p. 139.

Mock, K. (2004). Teaching with Tablet PC's. *J. Comput. Sci. Coll.*, 20(2):17–27. Cité 1 fois, p. 37.

Montaner, M., López, B. et de la Rosa, J. L. (2003). A Taxonomy of Recommender Agents on the Internet. *Artif. Intell. Rev.*, 19(4):285–330. Cité 1 fois, p. 18.

Moreau, B. (2008). Interface multi-facettes d'accès au capital documentaire d'une organisation. Rapport de stage d'IUP ISI L3, Université Toulouse 3. Stage encadré par Guillaume Cabanac et Max Chevalier, `http://www.irit.fr/~Guillaume.Cabanac/stagiaires/2008_L3_IUP_ISI_M.pdf`. Cité 1 fois, p. 130.

Mothe, J., Chrisment, C., Dousset, B. et Alaux, J. (2003). DocCube: Multi-dimensional visualisation and exploration of large document sets. *JASIST*, 54(7):650–659. Cité 1 fois, p. 20.

Mozdev (2003). Annozilla. `http://annozilla.mozdev.org`. Cité 3 fois, p. 33, 34 et 37.

Murphy, P. K., Long, J. F., Holleran, T. A. et Esterly, E. (2003). Persuasion online or on paper: A new take on an old issue. *Learn. Instr.*, 13(5):511–532. Cité 1 fois, p. 18.

NCSA (1993). Ncsa mosaic for microsoft windows user's guide. `http://www.state.sd.us/state/help/mosaic/WM8_1.htm`. Cité 1 fois, p. 36.

Noël, S. et Robert, J.-M. (2004). Empirical Study on Collaborative Writing: What Do Co-authors Do, Use, and Like? *Comput. Supported Coop. Work*, 13(1):63–89. Cité 1 fois, p. 17.

O'Hara, K. et Sellen, A. (1997). A Comparison of Reading Paper and On-Line Documents. *In CHI'97: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 335–342, New York, NY, USA. ACM Press. Cité 3 fois, p. 19, 39 et 55.

Olsen, D. R. J., Taufer, T. et Fails, J. A. (2004). ScreenCrayons: Annotating Anything. *In UIST'04: Proceedings of the 17th annual ACM symposium on User Interface Software and Technology*, pages 165–174, New York, NY, USA. ACM Press. Cité 3 fois, p. 29, 37 et 135.

O'Reilly, T. (2005). What Is Web 2.0 – Design Patterns and Businness Models for the Next Generation of Software. Available online `http://tim.oreilly.com/lpt/a/6228`. Cité 1 fois, p. 39.

Ovsiannikov, I. A., Arbib, M. A. et McNeill, T. H. (1999). Annotation technology. *Int. J. Hum.-Comput. Stud.*, 50(4):329–362. Cité 8 fois, p. 25, 26, 28, 30, 32, 34, 36 et 38.

Pallavidino, L. (2008). Vue unifiée d'un ensemble documentaire. Rapport de stage de DUT, Université Toulouse 3. Stage encadré par Guillaume Cabanac et Max Chevalier, `http://www.irit.fr/~Guillaume.Cabanac/stagiaires/2008_L2_IUT_Info_Pal.pdf`. Cité 1 fois, p. 133.

Pang, B., Lee, L. et Vaithyanathan, S. (2002). Thumbs up?: Sentiment Classification using Machine Learning Techniques. *In EMNLP'02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing*, pages 79–86, Morristown, NJ, USA. ACL. Cité 1 fois, p. 49.

Paramelle, R. (2008). Développement d'un composant additionnel d'annotation pour Internet Explorer. Rapport de stage de DUT, Université Toulouse 3. Stage encadré par Guillaume Cabanac et Max Chevalier, `http://www.irit.fr/~Guillaume.Cabanac/stagiaires/2008_L2_IUT_Info_Par.pdf`. Cité 2 fois, p. 98 et 124.

Pédauque, R. T. (2006). *Le document à la lumière du numérique*. C&F éditions, Caen, France. Cité 2 fois, p. 1 et 15.

Pédauque, R. T. (2007). *La redocumentarisation du monde*. Cépaduès-éditions, Toulouse, France. Cité 1 fois, p. 1.

Phelps, T. A. et Wilensky, R. (2000). Robust Intra-document Locations. *Comp. Netw.*, 33(1-6):105–118. Cité 2 fois, p. 31 et 36.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137. Cité 1 fois, p. 54.

Price, M. N., Schilit, B. N. et Golovchinsky, G. (1998). XLibris: The Active Reading Machine. *In CHI'98: CHI 98 conference summary on Human factors in computing systems*, pages 22–23, New York, NY, USA. ACM Press. Cité 3 fois, p. 10, 36 et 39.

Radev, D. R. (1999). Topic shift detection – finding new information in threaded news. Technical Report CUCS-026-99, Columbia University, Manhattan, NY, USA. Cité 1 fois, p. 63.

Reips, U.-D. (2002). Standards for Internet-Based Experimenting. *Exp. Psychol.*, 49(4):243–256. Cité 2 fois, p. 99 et 102.

Reips, U.-D. (2007). The methodology of Internet-based experiments. *In* Joinson, A. N., McKenna, K. Y. A., Postmes, T. et Reips, U.-D., éditeurs : *The Oxford Handbook of Internet Psychology*, chapitre 24, pages 373–390. Oxford University Press, New York, NY, USA. Cité 4 fois, p. 99, 102, 106 et 113.

Reips, U.-D. et Lengler, R. (2005). The *Web Experiment List*: A Web service for the recruitment of participants and archiving of Internet-based experiments. *Behav. Res. Meth.*, 37(2):287–292. Cité 1 fois, p. 105.

Resnick, P. et Varian, H. R. (1997). Recommender systems. *Commun. ACM*, 40(3):56–58. Cité 1 fois, p. 65.

Röscheisen, M., Mogensen, C. et Winograd, T. (1994). Shared web annotations as a platform for third-party value-added, information providers: Architecture, protocols, and usage examples. Technical report CSDTR/DLTR, Stanford University, Stanford, CA, USA. Cité 3 fois, p. 33, 34 et 36.

Rucker, J. et Polanco, M. J. (1997). Siteseer: personalized navigation for the web. *Commun. ACM*, 40(3):73–76. Cité 3 fois, p. 20, 65 et 76.

S, V. et RKVS, R. (2002). Annotation tool for the semantic web. *In Proceedings of the WWW2002 International Workshop Real world RDF and Semantic Web applications*, Hawaii, USA. Cité 1 fois, p. 37.

Salembier, P. et Zacklad, M. (2007). *Annotations dans les documents pour l'action*. Management des savoirs. Lavoisier, Paris, France. Cité 4 fois, p. 142, 143, 145 et 152.

Salton, G., Wong, A. et Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. *Commun. ACM*, 18(11):613–620. Cité 1 fois, p. 87.

Savy, N. (2006). *Probabilités et statistiques pour modéliser et décider : tests, validation, régression, plans d'expérience*. Statistiques. Ellipses, France. Cité 2 fois, p. 109 et 110.

Schickler, M. A., Mazer, M. S. et Brooks, C. (1996). Pan-browser Support for Annotations and Other Meta-Information on the World Wide Web. *Comp. Netw.*, 28(7-11):1063–1074. Cité 1 fois, p. 36.

Sellen, A. et Harper, R. (1997). Paper as an Analytic Resource for the Design of New Technologies. *In CHI'97: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326, New York, NY, USA. ACM. Cité 3 fois, p. 8, 24 et 27.

Sellen, A. J. et Harper, R. H. (2003). *The myth of the paperless office*. MIT Press, Cambridge, MA, USA. Cité 12 fois, p. 2, 7, 8, 9, 10, 19, 24, 28, 43, 137, 138 et 153.

Sellen, A. J., Murphy, R. et Shaw, K. L. (2002). How Knowledge Workers Use the Web. *In CHI'02: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 227–234, New York, NY, USA. ACM. Cité 1 fois, p. 21.

Seltzer, W. (1999). Annotation engine. http://cyber.law.harvard.edu/projects/annotate.html. Cité 1 fois, p. 36.

Shapiro, S. S. et Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611. Cité 1 fois, p. 109.

Shirky, C. (2003). A Group Is Its Own Worst Enemy. `http://shirky.com/writings/group_enemy.html`. Cité 1 fois, p. 92.

Shirky, C. (2008). *Here comes everybody: The power of organizing without organization*. The Penguin Press, London, UK. Cité 1 fois, p. 55.

Shum, S. B. et Sumner, T. (2001). JIME: An Interactive Journal for Interactive Media. *First Monday*, 6(2). Cité 1 fois, p. 36.

Surowiecki, J. (2005). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*. Anchor Books, New York. Cité 1 fois, p. 99.

Swarts, J. (2004). Cooperative Writing: Achieving Coordination Together and Apart. *In SIGDOC'04: Proceedings of the 22nd annual international conference on Design of communication*, pages 83–89, New York, NY, USA. ACM Press. Cité 1 fois, p. 17.

Teevan, J., Capra, R. et Quiñones, M. P. (2007). How People Find Personal Information. *In* Jones et Teevan (2007b), chapitre 2, pages 22–34. Cité 1 fois, p. 76.

Thiel, U., Brocks, H., Frommholz, I., Dirsch-Weigand, A., Keiper, J., Stein, A. et Neuhold, E. J. (2004). COLLATE – A collaboratory supporting research on historic European films. *Int. J. Digit. Libr.*, 4(1):8–12. Cité 1 fois, p. 36.

Tuffery, M. (1984). *Système Documentaire Base de Données Textuelles : le projet ETOILE*. Thèse de doctorat, Université Toulouse 3. Cité 1 fois, p. 54.

Twardy, C. (2004). Argument Maps Improve Critical Thinking. *Teaching Philosophy*, 27(2):95–116. Cité 1 fois, p. 100.

Uren, V., Cimiano, P., Iria, J., Handschuh, S., Vargas-Vera, M., Motta, E. et Ciravegna, F. (2006). Semantic annotation for knowledge management: Requirements and a survey of the state of the art. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4(1):14–28. Cité 1 fois, p. 29.

Vasudevan, V. et Palmer, M. (1999). On Web Annotations: Promises and Pitfalls of Current Web Infrastructure. *In HICSS'99: Proceedings of the 32nd Annual Hawaii International Conference on System Sciences*, volume 2, page 2012 (9 pages), Washington, DC, USA. IEEE Computer Society. Cité 4 fois, p. 32, 36, 38 et 73.

Voorhees, E. M. (2001). Overview of TREC 2001. *In TREC'01: Proceedings of the 10th Text REtrieval Conference*. Cité 1 fois, p. 118.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biom. Bull.*, 1(6):80–83. Cité 1 fois, p. 110.

Wiles, A. (1995). Modular elliptic curves and Fermat's Last Theorem. *Ann. Math.*, 141(3):443–551. Cité 1 fois, p. 24.

Wolfe, J. (2008). Annotations and the collaborative digital library: Effects of an aligned annotation interface on student argumentation and reading strategies. *International Journal of Computer-Supported Collaborative Learning*, 3(2):141–164. Cité 1 fois, p. 28.

Wolfe, J. L. (2000). Effects of Annotations on Student Readers and Writers. *In DL'00: Proceedings of the 5th ACM conference on Digital libraries*, pages 19–26, New York, NY, USA. ACM. Cité 3 fois, p. 28, 39 et 49.

Wolfe, J. L. et Neuwirth, C. M. (2001). From the Margins to the Center: The Future of Annotation. *J. Bus. Tech. Commun.*, 15(3):333–371. Cité 5 fois, p. 24, 27, 28, 38 et 104.

Wu, H. et Gordon, M. D. (2004). Collaborative Filing in a Document Repository. *In SIGIR'04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 518–519, New York, NY, USA. ACM Press. Cité 1 fois, p. 17.

Wu, H., Gordon, M. D. et DeMaagd, K. (2004). Document Co-Organization in an Online Knowledge Community. *In CHI'04: CHI'04 extended abstracts on Human factors in computing systems*, pages 1211–1214, New York, NY, USA. ACM Press. Cité 1 fois, p. 17.

Yang, Y., Akers, L., Klose, T. et Barcelon Yang, C. (2008). Text mining and visualization tools – impressions of emerging capabilities. *World Patent Information*, 30(4):280–293. Cité 2 fois, p. 19 et 88.

Zacklad, M. (2007). Annotation : attention, association, contribution. *In* Salembier et Zacklad (2007), chapitre 1, pages 29–46. Cité 1 fois, p. 29.

Zacklad, M., Lewkowicz, M., Boujut, J.-F., Darses, F. et Détienne, F. (2003). Formes et gestion des annotations numériques collectives en ingénierie collaborative. *In* Dieng, R., éditeur : *IC 2003*, pages 207–224, France. PUG. Cité 1 fois, p. 29.

Zheng, Q., Booth, K. et McGrenere, J. (2006). Co-Authoring with Structured Annotations. *In CHI'06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 131–140, New York, NY, USA. ACM Press. Cité 1 fois, p. 17.

Liste des figures

Liste des tableaux

Index