N° d'Ordre : D.U. 1142
EDSPIC : 199

Université Blaise Pascal Clermont-Ferrand II

Ecole Doctorale
Sciences Pour l'Ingénieur de Clermont-Ferrand

THÈSE

présentée par

Quentin DELACROIX

pour obtenir le grade de

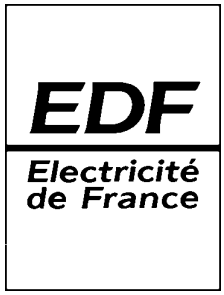
Docteur d'Université
spécialité Informatique

Un système pour la recherche plein texte et la consultation hypertexte de documents techniques

Soutenue publiquement le 8 juillet 1999 devant le jury composé de :

Mme Hélène BESTOUGEFF	Rapporteur
M. Jacques LE MAITRE	Rapporteur
M. Michel SCHNEIDER	Directeur de thèse
M. Lotfi LAKHAL	Examinateur
M. Jean-Loup LESNE	Examinateur
M. Alain QUILLOT	Président
M. Olivier VAILHEN	Examinateur

N° d'Ordre : D.U. 1142
EDSPIC : 199



GROUPE DES LABORATOIRES

Laboratoire d'Informatique (LIMOS)
Université Clermont-Ferrand II
Complexe scientifique des Cézeaux
F-63177 Aubière Cedex

Electricité de France
Groupe Des Laboratoires
21, Allée Privée
Carrefour Pleyel
F-93206 Saint-Denis Cedex 1

THÈSE

présentée

Quentin DELACROIX

pour obtenir le grade de

Docteur d'Université
spécialité Informatique

Un système pour la recherche plein texte et la consultation hypertexte de documents techniques

Soutenue publiquement le 8 juillet 1999 devant le jury composé de :

Mme Hélène BESTOUGEFF	Univ. Paris VII	Rapporteur
M. Jacques LE MAITRE	Univ. de Toulon et du Var	Rapporteur
M. Michel SCHNEIDER	Univ. Clermont-Ferrand II	Directeur de thèse
M. Lotfi LAKHAL	Univ. Clermont-Ferrand II	Examinateur
M. Jean-Loup LESNE	Electricité De France / GDL	Examinateur
M. Alain QUILLOT	Univ. Clermont-Ferrand II	Président du jury
M. Olivier VAILHEN	Electricité De France / DRD	Examinateur

"We commonly mistake data for information. Information starts with data, but data is not information – it is a source of information."

Ramesh Jain

mordu (d'informatique)
Personne passionnée d'informatique et ayant des connaissances techniques étendues.

extrait de la norme française Z 61-001
et de la norme internationnale ISO/CEI 2382-1:1993 :
"Technologie de l'information
Vocabulaire
Partie 1 : **Termes fondamentaux**"

L'achèvement de tout travail mené sur plusieurs années procure une grande satisfaction. Il est l'occasion de se remémorer les étapes passées et les personnes rencontrées.

Aussi, j'adresse mes sincères remerciements à mon directeur de thèse, Monsieur Michel Schneider, Professeur à l'Université Clermont-Ferrand II, dont les nombreuses interventions et les minutieuses relectures ont permis l'aboutissement de ce travail.

Je souhaite exprimer ma gratitude à Madame Hélène Bestougeff, Professeur à l'Université Paris VII et à Monsieur Jacques Le Maitre, Professeur à l'Université de Toulon et du Var, qui ont accepté d'évaluer ce travail afin d'en être les rapporteurs. Je les remercie d'avoir participé au jury, tout comme Monsieur Alain Quillot, Professeur à l'Université Clermont-Ferrand II, qui en fût le président ; Monsieur Lotfi Lakhal, Professeur à l'Université Clermont-Ferrand II, Monsieur Jean-Loup Lesne et Monsieur Olivier Vaihlen, Chefs de Service à Electricité De France, qui furent également membres de ce jury.

Je tiens aussi à remercier Monsieur Jean-Michel Barache, Directeur du Groupe Des Laboratoires d'EDF, ainsi que Messieurs Francis Pons, Bernhard Rotter, Jean-Loup Lesne, et Jean-François Joube qui m'ont accueilli au sein de leur équipe. Merci aussi à Monsieur Olivier Vaihlen pour son concours dans ce travail et pour ses conseils.

Pour leur collaboration lors de la mise en oeuvre de ce projet, je remercie Grégoire, Ludovic et Stéphane avec qui ce fût un grand plaisir de travailler.

Un grand merci à Jérôme pour ces observations avisées et ses encouragements renouvelés tout au long de cette thèse.

Je remercie également mes collègues du GDL à Saint-Denis, notamment Ghislain, Laurent et Laurent qui ont passé beaucoup de temps à m'expliquer leur métier au sein du GDL. Merci aussi à Benjamin, Olivier, Gérard...

Mes remerciements vont de même à mes camarades de Labo à Clermont-Ferrand : à Alain et Christophe, pour leurs relectures, leurs commentaires et les discussions, mais aussi pour les nombreux cafés et divertissements en tout genre ; à Stéphane, pour l'efficacité de ses interventions lors des phases de codage ; à Jérôme, Kitsana et Yahia, pour leurs conseils et leurs recommandations ; à Samah, pour son appui logistique et ses prescriptions ; à Lionel et Vincent, mais aussi à David, Julien, Nathalie, Pat... Merci pour votre sympathie, pour votre jovialité et pour les excellents moments partagés.

Enfin, je tiens à renouveler toute ma reconnaissance à mc pour son investissement personnel et pour son fidèle soutien.

La diversité des cultures et des personnalités rencontrées tout au long de cette thèse en firent une expérience très intéressante et formidablement enrichissante. Merci à tous.

Table des matières

Liste de tableaux

Table des illustrations

Introduction

La conservation sous forme électronique des diverses informations et documents que gère une entreprise est une étape indispensable au développement de son système d'information. Cependant, cette disposition n'est pas suffisante pour assurer la meilleure valorisation de l'information. Même lorsque toutes les informations sont gérées par des systèmes informatiques, un acteur du système d'information peut rencontrer des difficultés d'accès et de consultation. Partant de cette constatation, nous nous sommes intéressés aux problèmes que peuvent poser la recherche et la consultation des informations techniques dans un contexte industriel.

Parmi les solutions mises en oeuvre pour palier à ces problèmes, nous constatons que ni les systèmes de recherche de documents, ni les systèmes plus récents de recherche d'informations basés sur le Web ne donnent pleinement satisfaction. Les premiers retrouvent les documents mais n'en permettent pas une consultation aisée. Les seconds offrent une consultation conviviale des documents mais ne sont pas adaptés aux caractéristiques particulières des documents d'entreprise.

Les difficultés de gestion de l'information et des documents en milieu industriel sont nombreuses. Citons la diversité des systèmes de stockage, le respect des permissions d'accès, la gestion des versions, la mise à jour des références croisées entre documents. Le stockage sous forme électronique des documents pose en lui-même des problèmes supplémentaires. Par exemple, les nombreux formats de stockage (`.doc`, `.dat`, ...) sont spécifiques aux contenus des documents (textes standards, signaux multidimentionnels, ...) et nécessitent des applications spécifiques. De plus ces applications évoluent fréquemment et des problèmes de compatibilité entre versions peuvent survenir.

Les travaux présentés dans cette thèse ont pour objectif d'analyser ces diverses difficultés et de suggérer des améliorations pour les processus de recherche et de consultation d'informations techniques en contexte industriel. Ils considèrent aussi bien les attentes et les besoins des utilisateurs et des producteurs d'informations que les préoccupations des administrateurs des systèmes informatiques.

D'une part, les besoins des acteurs du système d'information sont sans cesse plus pressants. Un utilisateur voudra toujours disposer d'une information où qu'il soit, où qu'elle soit, au plus tôt. Qu'il se déplace au sein de différents sites de l'entreprise, ou qu'il soit en visite chez un partenaire, l'information nécessaire doit être accessible rapidement afin de prendre les bonnes décisions. Quand aux producteurs d'informations, ils préfèrent ne pas avoir à se soucier, lors de la rédaction de leurs documents, des contraintes liées au stockage, à la recherche et à la consultation qui seront effectuées ultérieurement.

D'autre part, les préoccupations des administrateurs de systèmes informatiques orientent leurs choix vers des logiciels nécessitant une maintenance minimum, perturbant le moins possible les

utilisateurs et les autres systèmes informatiques, et économisant la bande passante des réseaux locaux et distants.

Afin de faciliter la recherche et la consultation d'informations dans un contexte industriel, nous proposons un système nommé RECITAL (Recherche Et Consultation de l'Information Technique Aux Laboratoires) qui concilie la plupart des points évoqués ci-dessus. RECITAL combine la recherche en texte intégral à la consultation hypertexte des documents. Il offre ainsi une recherche sur l'intégralité des textes contenus dans les documents de l'entreprise et permet la consultation de ces documents dans un environnement hypertexte. Les documents pris en compte par RECITAL peuvent être répartis sur des serveurs de fichiers en réseau ou dans des bases de données. La recherche tient compte de l'identité de l'utilisateur afin de ne lui présenter que des documents qu'il a le droit de consulter. La consultation des documents se fait depuis n'importe quel poste de travail en réseau disposant d'un client universel hypertexte (typiquement un butineur Web). Les formats électroniques dans lesquels sont stockés les documents sont éventuellement convertis par RECITAL dans les formats universels compatibles avec le client universel. Des liens hypertextes vers les documents cités en référence, s'ils sont présents sur le réseau et si l'utilisateur a le droit d'y accéder, sont ajoutés dynamiquement par RECITAL lors de la consultation. Ainsi, l'utilisateur peut aisément accéder à d'autres documents liés à celui qu'il consulte.

Organisation du mémoire.

Le Chapitre 1 présente le contexte de notre étude et fait ressortir la nécessité de considérer les diverses difficultés rencontrées lors de la recherche et de la consultation d'informations techniques en milieu industriel. Une problématique plus générale de ces activités est alors présentée conjointement à l'exposé de nos objectifs.

Le Chapitre 2 fait état de systèmes de recherche d'informations et de consultation de documents. Les principales techniques pour l'indexation de textes et la construction d'hypertextes sont notamment présentées.

Le Chapitre 3 décrit le système RECITAL que nous proposons pour effectuer la recherche et la consultation des informations contenues dans des documents techniques de l'entreprise. Nous suggérons d'abord un modèle de données adapté à notre contexte et à nos objectifs. Les fonctionnalités et les propriétés de RECITAL sont alors explicitées. L'architecture d'ensemble et le fonctionnement général de RECITAL sont ensuite présentés. Les étapes de l'utilisation de RECITAL sont décrites avant de développer les processus mis en oeuvre lors de l'indexation des textes et de l'hypertextualisation des documents. Enfin, les possibilités offertes pour l'administration de RECITAL sont évoquées.

Le Chapitre 4 présente une maquette partielle de RECITAL. Les structures de données nécessaires à la mise en oeuvre des processus de recherche plein texte et de construction dynamique d'hypertextes sont proposées. Leur construction, leur utilisation et leur mise à jour sont expliquées. La mise en œuvre des fonctionnalités innovantes de RECITAL est ensuite illustrée à travers des scénarios d'utilisation.

La Conclusion de ce mémoire fait le bilan des travaux réalisés. Enfin, des évolutions du système RECITAL sont proposées.