

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	Examineur
M. Jean-Loup LESNE	Examineur
M. Alain QUILLOT	Président
M. Olivier VAILHEN	Examineur

N° d'Ordre : D.U. 1142

EDSPIC : 199



GRUPE 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	Examineur
M. Jean-Loup LESNE	Electricité De France / GDL	Examineur
M. Alain QUILLOT	Univ. Clermont-Ferrand II	Président du jury
M. Olivier VAILHEN	Electricité De France / DRD	Examineur

"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 internationale 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 Maître, 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 Joubé 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

Introduction	17
Chapitre 1 Le contexte de l'étude	19
1.1 L'information technique au sein du Groupe Des Laboratoires.	19
1.1.1 Le Groupe Des Laboratoires dans l'entreprise EDF.	19
1.1.1.1 Ses missions.	19
1.1.1.2 Ses effectifs et leur localisation.	20
1.1.1.3 Ses partenaires et prestataires.	20
1.1.2 Le Système d'Information technique du GDL.	20
1.1.2.1 Le système d'information du point de vue de l'assistance, de la surveillance et de la validation des contrôles.	20
1.1.2.2 Le système d'information du point de vue de la mise au point des interventions de contrôle et de la participation au Retour d'EXpérience.	21
1.1.2.3 Le système d'information du point de vue de l'organisation de l'entreprise et des méthodes de travail.	22
1.1.2.4 Travaux et projets dans lesquels le GDL est impliqué.	22
1.1.3 Faisabilité et limites d'un système d'aide à la consultation des documents techniques du GDL.	22
1.1.3.1 Faisabilité d'un stockage des documents techniques sous forme électronique.	23
1.1.3.2 Limites d'un traitement automatisé des concepts techniques utilisés au GDL.	23
1.2 Une problématique de la consultation des données du système d'information technique.	24
1.2.1 La production d'informations et l'utilisation d'informations.	24
1.2.1.1 Le décalage entre la production et l'utilisation.	24
1.2.1.2 Les perturbations entre la production et l'utilisation.	25
1.2.2 Les formats de codage de l'information électronique.	26
1.2.3 Les ressources multiples et les copies d'informations.	26
1.2.4 L'information et son référentiel.	27
1.2.5 Les restrictions d'accès à l'information.	28
1.3 Conclusion.	28
Chapitre 2 Techniques et outils pour la recherche et la consultation d'informations.	29
2.1 De la recherche de documents à la consultation d'informations : petite revue historique.	29
2.2 Effectuer la recherche et la consultation.	31
2.2.1 Deux processus étroitement liés.	31
2.2.2 Recherche par approches successives.	31
2.2.3 Recherche par accès directs.	32
2.2.3.1 Principes.	32

2.2.3.2	Caractérisation de l'information et de son contenant.	32
2.2.3.3	Formulation des requêtes.	32
2.2.3.4	Présentation des réponses.	33
2.2.4	Consultation de documents.	33
2.2.4.1	Le concept d'hypertexte.	34
2.2.4.2	Les liens typés.	34
2.2.4.3	L'hypertextualisation.	34
2.2.5	Stratégie des experts du domaine / Stratégie des experts en recherche d'informations.	34
2.2.6	Evaluation des systèmes de recherche d'informations.	35
2.3	Préparer la recherche d'informations.	35
2.3.1	Les difficultés d'analyse des textes.	35
2.3.1.1	Les ambiguïtés du langage naturel.	36
2.3.1.2	Réduction des ambiguës du langage naturel.	36
2.3.2	Approches et définitions relatives à l'indexation.	36
2.3.3	Recherche documentaire et recherche d'informations.	37
2.3.4	Recherche basée sur un condensé ou sur l'intégralité.	38
2.3.5	Indexation basée sur les caractéristiques textuelles.	38
2.4	Quelques techniques pour la construction de représentations.	39
2.4.1	Index de mots et concordance de chaînes (modèle booléen).	40
2.4.2	Le modèle vectoriel.	41
2.4.3	Le modèle probabiliste.	43
2.4.4	Utilisation du contexte des mots.	43
2.4.5	Approches utilisant les thésaurus et les réseaux sémantiques.	46
2.4.5.1	Un exemple de thésaurus : le WordNet.	46
2.4.5.2	Les chaînes lexicales.	47
2.5	Construction d'hypertextes.	47
2.5.1	Liens statiques et associations manuelles.	48
2.5.2	Liens statiques et associations automatiques.	49
2.5.2.1	Utilisation de la structure physique des documents et des références croisées.	49
2.5.2.2	Utilisation du modèle vectoriel affiné.	50
2.5.2.3	Utilisation du chaînage lexical.	53
2.5.3	Génération dynamique de liens.	56
2.6	Présentation de quelques systèmes.	57
2.7	Conclusion.	59

Chapitre 3 Le système RECITAL..... 61

3.1	Un modèle des données de l'entreprise.	61
3.1.1	Introduction.	62
3.1.2	La classe d'objets DONNEE.	62
3.1.2.1	Considérations matérielles sur les données.	62
3.1.2.2	Considérations organisationnelles sur les données.	64
3.1.2.3	Considérations temporelles sur les données.	64
3.1.2.4	Autres considérations.	65
3.1.2.5	Synthèse de la classe DONNEE.	66
3.1.3	La classe d'objets ACTEUR.	66
3.1.4	Les classes d'objets INFORMATION et DONNEE NUMERIQUE.	68
3.1.4.1	UNITE DE DONNEES NUMERIQUES et ENSEMBLE DE DONNEES NUMERIQUES.	68

3.1.4.2 ENSEMBLE DE DONNEES NUMERIQUES TEXTUELLES et ENSEMBLE DE DONNEES NUMERIQUES SANS TEXTE.	69
3.1.5 Synthèse du modèle proposé.	70
3.2 Les fonctionnalités et les propriétés.	70
3.2.1 La fonctionnalité de recherche d'informations.....	70
3.2.2 La fonctionnalité de consultation d'informations.	71
3.2.3 Les fonctionnalités communes à la recherche et à la consultation d'informations.	71
3.2.3.1 Rechercher et consulter depuis le poste de travail de chaque acteur.	71
3.2.3.2 Rechercher et consulter via une interface unique.	71
3.2.3.3 Rechercher et consulter dans l'ensemble des bases d'informations.	71
3.2.3.4 Rechercher et consulter en respectant l'accessibilité de chaque unité de données.....	71
3.2.3.5 Filtrer les informations.....	72
3.2.3.6 Assurer la disponibilité immédiate des informations.	73
3.2.4 Les fonctionnalités pour la recherche d'informations.	74
3.2.4.1 Se connecter à un guichet unique.....	74
3.2.4.2 Rechercher dans l'ensemble des bases d'informations.	74
3.2.4.3 Caractériser le contenant et le contenu de l'information recherchée.....	74
3.2.5 Les fonctionnalités pour la consultation d'informations.	74
3.2.5.1 Accéder directement à l'information correspondant aux critères de la recherche.	74
3.2.5.2 Utiliser un nommage uniforme pour les unités de données.....	75
3.2.5.3 Consulter des données de façon uniforme.	75
3.2.5.4 Consulter aisément les informations associées.	75
3.2.5.5 Utiliser les informations consultées.....	75
3.2.6 Les fonctionnalités pour l'administration.	75
3.2.6.1 Indiquer les bases d'informations à prendre en compte.	75
3.2.6.2 Indiquer comment associer les données.....	76
3.2.6.3 Suivre le déroulement des processus de RECITAL.....	76
3.2.7 Les propriétés de RECITAL.	76
3.2.7.1 Dissociation entre administration des bases d'informations et administration de RECITAL.	76
3.2.7.2 Compatibilité avec les formats de données.....	76
3.2.7.3 Réactivité.	76
3.2.7.4 Interopérabilité avec les autres systèmes informatiques.	76
3.2.7.5 Modularité et relative indépendance de RECITAL.	77
3.2.7.6 Indépendance du système d'information vis à vis de RECITAL.	77
3.2.7.7 Evolutivité.....	77
3.2.8 Synthèse.....	77
3.3 L'architecture et le fonctionnement.	78
3.3.1 L'organisation des services.	78
3.3.2 L'architecture trois-tiers.	79
3.3.3 Le fonctionnement de RECITAL.	82
3.3.4 Services assurés par les éléments du poste de travail.	85
3.4 L'utilisation.....	85
3.4.1 Se connecter.....	86
3.4.2 Définir les préférences.....	87
3.4.3 Préparer la question de recherche d'informations.	87
3.4.3.1 Caractériser l'information recherchée.	87

3.4.3.2 Choisir les préférences de présentation de la réponse de RECITAL.	89
3.4.4 Obtenir la liste des informations ayant les caractéristiques précisées.	91
3.4.5 Choisir un élément dans la liste des liens.	91
3.4.6 Consulter un élément choisi.	93
3.4.7 Détailler un élément choisi.	93
3.5 Les processus d'indexation, de recherche et d'hypertextualisation.	94
3.5.1 Approche retenue.	94
3.5.2 Présentation générale orientée processus.	96
3.5.2.1 L'indexation des données.	96
3.5.2.2 La recherche de données.	96
3.5.2.3 L'hypertextualisation de documents.	96
3.5.3 Présentation détaillée orientée données et métadonnées.	97
3.5.3.1 Les données échangées avec les bases d'informations.	97
3.5.3.2 Les métadonnées internes au fonctionnement.	98
3.5.3.3 Les données échangées avec l'utilisateur.	105
3.5.3.4 Les données échangées avec l'administrateur.	106
3.6 L'administration.	106
3.6.1 Indications des données à prendre en compte.	106
3.6.2 Paramètres pour la localisation des références de documents et des noms d'acteurs.	108
3.6.3 Paramètres pour l'ajustement des performances.	108
3.6.4 Paramètres pour les options des utilisateurs.	109
3.6.5 Les messages pour la maintenance.	109
Chapitre 4 Une maquette pour RECITAL	111
4.1 Implémentation des structures de données.	111
4.1.1 Organisation inter-structures.	111
4.1.2 Organisation intra-index.	114
4.1.2.1 Arbre binaire de recherche.	114
4.1.2.2 Fonction de transposition.	115
4.1.2.3 Implémentation des arbres binaires de caractères transposés.	117
4.1.3 Organisation des principales structures.	118
4.1.3.1 La classe <i>arborescence image</i>	118
4.1.3.2 Les classes <i>index de mots</i>	120
4.1.3.3 Les classes <i>index de fonctions</i>	123
4.1.3.4 Les classes <i>table</i>	124
4.1.3.5 La classe <i>métadonnées pour l'hypertextualisation</i>	126
4.1.4 Conservation des structures.	127
4.1.5 Bilan.	128
4.1.5.1 Complexité.	128
4.1.5.2 Codage.	129
4.2 Les scénarios illustrant les principales fonctionnalités.	131
4.2.1 Fonctionnalités à démontrer.	131
4.2.2 Rappel du schéma de fonctionnement de RECITAL.	132
4.2.3 Etat initial commun à tous les scénarios.	132
4.2.3.1 Les acteurs.	132
4.2.3.2 Le système informatique.	133
4.2.4 Scénario 1 (S1).	136
4.2.4.1 Rappel de la fonctionnalité à démontrer.	136

4.2.4.2	Synoptique du scénario 1.	136
4.2.4.3	Détail de l'implémentation.	137
4.2.4.4	Déroulement d'un scénario S1.	138
4.2.4.5	Détails d'un scénario S1.	141
4.2.4.6	Synthèse : aspect démontré par le scénario 1.	143
4.2.5	Scénario 2 (S2).	143
4.2.5.1	Rappel de la fonctionnalité à démontrer.	144
4.2.5.2	Synoptique du scénario 2.	144
4.2.5.3	Détail de l'implémentation.	145
4.2.5.4	Déroulement d'un scénario S2.	147
4.2.5.5	Détails d'un scénario S2.	149
4.2.5.6	Synthèse : aspect démontré par le scénario 2.	154
4.2.6	Scénario 3 (S3).	154
4.2.6.1	Rappel de la fonctionnalité à démontrer.	154
4.2.6.2	Synoptique du scénario 3.	154
4.2.6.3	Synthèse : aspect démontré par le scénario 3.	154
4.2.7	Scénario 4 (S4).	154
4.2.7.1	Rappel de la fonctionnalité à démontrer.	154
4.2.7.2	Synoptique du scénario 4.	155
4.2.7.3	Détail de l'implémentation.	156
4.2.7.4	Déroulement d'un scénario S4.	158
4.2.7.5	Détails d'un scénario S4.	160
4.2.7.6	Synthèse : aspect démontré par le scénario 4.	163
Conclusion		165
Références bibliographiques		167
Références vers l'internet		177
Glossaire		179
Annexes		187

Liste de tableaux

Tableau 2-1 : typologie des modes de requêtes [Lefevre 97 p. 75].	33
Tableau 2-2 : typologie des modes d'indexation [Lefevre 97 p. 69].	39
Tableau 3-1 : les principaux ensembles de données et de métadonnées organisés en structures.	99
Tableau 4-2 : récapitulatif des structures par catégories de données concernées.	112

Table des illustrations

Figure 1-1 : les flux de données entre les activités de contrôle.	21
Figure 1-2 : le cycle de production / utilisation de l'information.	25
Figure 2-1 : exemple de construction d'une matrice de co-occurrences.	45
Figure 2-2 : exemple de relations entre termes du WordNet.	47
Figure 2-3 : les relations de similarité entre les articles d'une encyclopédie.	50
Figure 2-4 : les liens entre les articles ayant une similarité > 0,30.	51
Figure 2-5 : les liens entre les articles ayant une similarité > 0,50.	51
Figure 2-6 : représentation des similarités entre les paragraphes d'un même document.	51
Figure 2-7 : représentation des similarités entre les paragraphes de deux documents.	51
Figure 2-8 : représentation montrant que le document de Février 1989 est un assemblage des deux documents de Janvier 1988.	52
Figure 2-9 : représentation montrant que le document <i>USA</i> aborde un sujet détaillé dans le document <i>JFK</i>	52
Figure 2-10 : un exemple de matrice des densités de chaînes.	54
Figure 2-11 : un exemple de matrice des adjacences de paragraphes.	55
Figure 3-1 : les attributs de la classe DONNEE.	63
Figure 3-2 : première représentation de l'exemple "ligne".	63
Figure 3-3 : seconde représentation de l'exemple "ligne".	63
Figure 3-4 : les caractéristiques organisationnelles de la classe DONNEE.	64
Figure 3-5 : les caractéristique temporelles de la classe DONNEE.	65
Figure 3-6 : la classe INFO_SUP.	66
Figure 3-7 : la classe d'objets DONNEE et les caractéristiques associées.	66
Figure 3-8 : la classe d'objets ACTEUR.	67
Figure 3-9 : un exemple de sous-classe d'ACTEUR : la classe AUTEUR.	67
Figure 3-10 : la classe d'objets INFORMATION.	68
Figure 3-11 : la classe d'objets DONNEE NUMERIQUE.	68
Figure 3-12 : UNITE DE DONNEES NUMERIQUES et ENSEMBLE DE DONNEES NUMERIQUES.	69
Figure 3-13 : les CARACTERES de l'ALPHABET composent les MOTS des ENSEMBLES DE DONNEES NUMERIQUES TEXTUELLES.	70
Figure 3-14 : le filtrage des informations effectué par RECITAL, point de vue de l'utilisateur de RECITAL.	73
Figure 3-15 : le filtrage des informations effectué par RECITAL, point de vue de l'administrateur des informations.	73
Figure 3-16 : les activités et les flux de données entre les 3 entités Acteur, RECITAL et Base d'informations.	79
Figure 3-17 : l'architecture client-serveur du réseau.	80
Figure 3-18 : la répartition géographique des composants et des modules de RECITAL.	80
Figure 3-19 : l'organisation fonctionnelle des modules de RECITAL.	81
Figure 3-20 : les flux de données et de métadonnées au sein de l'architecture fonctionnelle de RECITAL.	82
Figure 3-21 : diagramme séquentiel du fonctionnement de RECITAL.	83

Figure 3-22 : diagramme d'activités d'une utilisation de RECITAL.....	86
Figure 3-23 : l'acteur se connecte au réseau.....	86
Figure 3-24 : l'acteur se connecte à RECITAL.....	87
Figure 3-25 : l'utilisateur caractérise l'information recherchée.....	89
Figure 3-26 : choix du type d'élément associé à chaque lien.....	90
Figure 3-27 : choix de l'ordre de présentation des liens.....	90
Figure 3-28 : choix du niveau de détail des commentaires.....	90
Figure 3-29 : validation des données.....	91
Figure 3-30 : réponse de RECITAL.....	91
Figure 3-31 : choix pour la consultation par éléments de données.....	92
Figure 3-32 : choix pour la consultation de l'unité de données en format universel.....	92
Figure 3-33 : choix pour la consultation de l'unité de données en format natif.....	92
Figure 3-34 : choix d'un élément de données à consulter.....	93
Figure 3-35 : choix du constituant à détailler.....	94
Figure 3-36 : diagramme d'activités d'un exemple de fonctionnement de RECITAL.....	95
Figure 3-37 : séquence des échanges avec les bases d'informations.....	97
Figure 3-38 : séquence des échanges avec les annuaires réseaux.....	98
Figure 3-39 : exemple d'ensembles de mots indexés.....	102
Figure 3-40 : diagramme d'activités pour la création des métadonnées.....	103
Figure 3-41 : diagramme d'activités pour la mise à jour des métadonnées.....	104
Figure 3-42 : exemple de syntaxe pour les références de documents.....	108
Figure 4-1 : vue synthétique des référencements entre structures.....	113
Figure 4-2 : exemple d'un arbre binaire dont les noeuds sont les caractères des mots.....	115
Figure 4-3 : définition d'une transposition.....	115
Figure 4-4 : un exemple de transposition, T1.....	116
Figure 4-5 : une structure d'index utilisant un arbre binaire dont les noeuds sont des caractères transposés.....	117
Figure 4-6 : le tableau de transposition pour T1.....	117
Figure 4-7 : un arbre binaire de caractères transposés utilisant les listes chaînées.....	118
Figure 4-8 : représentation UML de la classe <i>arborescence image</i>	119
Figure 4-9 : l' <i>arborescence image</i> (structure 1b).....	119
Figure 4-10 : représentation UML des classes index des <i>nom_de_l'arborescence_image</i> et index des <i>mots</i>	120
Figure 4-11 : l'index des <i>nom_de_l'arborescence_image</i> (structure 2).....	121
Figure 4-12 : organisation de la structure 3 - schéma 1.....	122
Figure 4-13 : organisation de la structure 3 - schéma 2.....	122
Figure 4-14 : l'index des <i>mots</i> contenus dans les fichiers (structure 3).....	123
Figure 4-15 : représentation UML des classes d'index de fonctions.....	124
Figure 4-16 : l'index des <i>fonction_de_mot</i> (structure 4).....	124
Figure 4-17 : représentation UML de la classe table de <i>nom_d'acteur</i>	125
Figure 4-18 : la table des <i>nom_d'acteur</i> (structure 5).....	125
Figure 4-19 : représentation UML de la classe table de <i>référence_de_document</i>	125
Figure 4-20 : la table des <i>référence_de_document</i> (structure 8).....	126
Figure 4-21 : représentation UML de la classe <i>métadonnées pour l'hypertextualisation</i>	127
Figure 4-22 : un ensemble de <i>métadonnées pour l'hypertextualisation</i> (structure 11).....	127
Figure 4-23 : les principales propriétés des objets implémentés.....	130
Figure 4-24 : les cas d'utilisation de RECITAL mis en oeuvre.....	130
Figure 4-25 : diagramme d'activités d'un fonctionnement de RECITAL.....	132
Figure 4-26 : la structure hiérarchique des acteurs.....	133
Figure 4-27 : l'arborescence de répertoires.....	134

Figure 4-28 : les identifiants réseau des acteurs (avec descriptions).	134
Figure 4-29 : un exemple de fichier permissions.txt.	134
Figure 4-30 : synthèse des permissions de consultation pour les répertoires.	135
Figure 4-31 : exemple de balises d'un "document".	136
Figure 4-32 : diagramme d'activités de RECITAL adapté au scénario 1.	137
Figure 4-33 : un exemple de fichier de paramètres pour RS1 (param.txt).	137
Figure 4-34 : un exemple de QUESTION à RS1 (fichier question.txt).	137
Figure 4-35 : un exemple d'en-tête pour le tableau contenant la REPONSE de RS1.	138
Figure 4-36 : diagramme d'états de RS1.	138
Figure 4-37 : diagramme d'états du scénario 1 (les 3 fenêtres d'exécution).	139
Figure 4-38 : diagramme des interactions pour le scénario 1.	139
Figure 4-39 : diagramme séquentiel d'une démonstration (extrait du scénario 1).	140
Figure 4-40 : diagramme d'états de la Fenêtre_1 (exécution de RS1).	140
Figure 4-41 : diagramme d'états de la Fenêtre_2 (édition de la QUESTION).	141
Figure 4-42 : diagramme d'états de la Fenêtre_3 (consultation de la REPONSE avec le butineur).	141
Figure 4-43 : la REPONSE de RS1 à <i>acteur_02</i> .	142
Figure 4-44 : la REPONSE de RS1 à <i>acteur_05</i> .	143
Figure 4-45 : diagramme d'activités de RECITAL adapté au scénario 2.	145
Figure 4-46 : un exemple de fichier de paramètres pour RS2 (param.txt).	145
Figure 4-47 : exemples de QUESTION à RS2 (fichiers question.txt).	146
Figure 4-48 : diagramme d'états de RS2.	146
Figure 4-49 : activités de "construction de la REPONSE" pour RS2.	146
Figure 4-50 : diagramme d'états du scénario 2 (les 3 fenêtres d'exécution).	147
Figure 4-51 : diagramme des interactions pour le scénario 2.	147
Figure 4-52 : diagramme séquentiel d'une démonstration (extrait du scénario 2).	148
Figure 4-53 : diagramme d'états de la Fenêtre_1 (exécution de RS2).	148
Figure 4-54 : diagramme d'états de la Fenêtre_2 (édition de la QUESTION).	149
Figure 4-55 : diagramme d'états de la Fenêtre_3 (consultation de la REPONSE avec le butineur).	149
Figure 4-56 : le contenu du "document" avant préparation par RS2.	150
Figure 4-57 : le "document" avant préparation par RS2 (consulté avec le butineur).	151
Figure 4-58 : le contenu de la REPONSE pour <i>acteur_07</i> (le fichier reponse.html).	152
Figure 4-59 : le document "les affaires en cours" vu par <i>acteur_07</i> .	152
Figure 4-60 : le document "les affaires en cours" vu par <i>acteur_12</i> .	153
Figure 4-61 : diagramme d'activités de RECITAL adapté au scénario 4.	155
Figure 4-62 : un exemple de fichier de paramètres pour RS4 (param.txt).	156
Figure 4-63 : trois exemples de QUESTION à RS4 (fichiers <i>question.txt</i>).	156
Figure 4-64 : diagramme d'états de RS4.	157
Figure 4-65 : activités de "construction de la REPONSE" pour RS4.	157
Figure 4-66 : diagramme d'états du scénario 4 (les 3 fenêtres d'exécution).	158
Figure 4-67 : diagramme des interactions pour le scénario 4.	158
Figure 4-68 : diagramme séquentiel d'une démonstration (extrait du scénario 4).	159
Figure 4-69 : diagramme d'états de la Fenêtre_1 (exécution de RS4).	159
Figure 4-70 : diagramme d'états de la Fenêtre_2 (édition de la QUESTION).	160
Figure 4-71 : diagramme d'états de la Fenêtre_3 (consultation de la REPONSE avec le butineur).	160
Figure 4-72 : la première REPONSE de RS4.	161
Figure 4-73 : le fichier <i>reponse.html</i> proposant de choisir la version du document.	162
Figure 4-74 : la REPONSE de RS4 proposant de choisir la version du document.	162

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 multidimensionnels, ...) 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 oeuvre 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.