



213 rue Lafayette
75480 Paris cedex 10



45 avenue des Etats-Unis
78035 Versailles cedex

L'EXPLOITATION DU DOCUMENT NUMÉRIQUE DANS UN SYSTÈME D'INFORMATION DOCUMENTAIRE

Mémoire
en vue de l'obtention du DESS Information scientifique et technique

.....

Diane STITELET – LE HENAFF

Sous la direction de : Dominique CAVET
Directeur du secteur documentation de la délégation de l'information et de la communication de l'IRD

Date de la soutenance
20-21 septembre 2004

A n n é e 2003 – 2004

AVANT-PROPOS

Remerciements

Je remercie sincèrement tous les collègues de l'IRD qui m'ont très bien accueillie.

Je remercie Lucile Gauer, chef de projet SI et Caroline Doucouré, chef de projet documentation, qui m'ont intégrée au sein de l'équipe projet de l'IRD pour le déploiement du système d'information documentaire Ex Libris de Cadic.

Pour son soutien, ses conseils, sa disponibilité et sa rigueur, je remercie plus particulièrement mon directeur de stage, Monsieur Dominique Cavet.

Résumé :

Dans un contexte de nouvelles technologies appliquées à l'information, cette étude s'intéresse à l'exploitation du document numérique au sein d'un système d'information.

Celui ci permet de modéliser les procédures documentaires d'une organisation à travers différentes étapes : du stockage à la diffusion, en passant par l'indexation et la recherche des documents numériques.

L'évaluation de la recherche d'information dans le système dépend des techniques utilisées dans l'indexation et la recherche des documents. Cette étude fait le point sur les différentes techniques utilisées à ce jour.

L'expérience de déploiement d'un système d'information documentaire à l'IRD permet de présenter des cas pratiques, notamment dans le traitement des aspects juridiques liés à l'exploitation du document numérique.

Mots clés :

Document numérique, système d'information documentaire, système de recherche d'information, moteur de recherche, indexation, droit d'auteur.

Abstract :

Considering the impact of new technology on information, this study focuses on the exploitation of the digital document within a library information system.

This system makes it possible to construct a model of the documentary procedures of an organization throughout their various stages from storage to distribution, including the intermediate steps of indexing and digital document research.

Assessing the search for information within the system depends upon the techniques employed for the indexing and the finding of documents. This study focuses on these different, currently used techniques.

Experimenting with the launch of a library information system at IRD also allows for the presentation of practical cases, particularly in the treatment of legal aspects related to the utilization of the digital document.

Keywords :

Digital document, library information system, search engine, indexing, copyright.

SOMMAIRE

1. INTRODUCTION	6
2. PRESENTATION DE L'IRD	7
3. LE DOCUMENT NUMERIQUE DANS UN SYSTEME D'INFORMATION DOCUMENTAIRE	8
3.1. LES TECHNIQUES DE NUMERISATION	8
3.2. LE DOCUMENT PRODUIT NUMERIQUEMENT	9
3.3. AVANTAGES ET LIMITES DE LA STRUCTURATION	10
3.4. LE SYSTEME D'INFORMATION DOCUMENTAIRE	12
3.5. LE WORKFLOW	17
3.6. LES DROITS D'ACCES	19
3.7. LE PORTAIL DOCUMENTAIRE	20
4. LE SYSTEME DE RECHERCHE D'INFORMATION (SRI)	22
4.1. L'INDEXATION	23
4.2. LES TECHNIQUES D'IDENTIFICATION DES ENTITES D'INDEXATION	24
4.3. LES TECHNIQUES DE CONSTRUCTION DES INDEX	26
4.4. APPLICATION A L'INDEXATION AUTOMATIQUE	27
4.5. LA RECHERCHE D'INFORMATION	28
4.6. LES MODELES DE RECHERCHE D'INFORMATION	29
4.7. L'EVALUATION D'UN SYSTEME DE RECHERCHE D'INFORMATION	31

5. LES DROITS D'EXPLOITATION DU DOCUMENT NUMERIQUE	32
5.1. LA PROPRIETE INTELLECTUELLE	32
5.2. REGIME JURIDIQUE DES AGENTS DE L'ADMINISTRATION	33
5.3. ETABLISSEMENT DE CONTRATS AVEC LES EDITEURS : NOTION DE CONTRAT D'EDITION	34
5.4. TYPES DE DOCUMENTS POUVANT CONSTITUER UN FONDS DOCUMENTAIRE	35
6. CONCLUSION	39
BIBLIOGRAPHIE	40
ANNEXES	41
ANNEXE 1 : CONTRAT DE REPRODUCTION ET D'EXPLOITATION DES OUVRAGES PUBLIES PAR KARTHALA DONT AU MOINS UN DES AUTEURS EST UN CHERCHEUR DE L'IRD	41
ANNEXE 2 : CONTRAT DE REPRODUCTION ET D'EXPLOITATION DES ARTICLES PUBLIES PAR LE MNHN DONT AU MOINS UN DES AUTEURS EST UN CHERCHEUR DE L'IRD	46
ANNEXE 3 : LES POLITIQUES EDITORIALES EN MATIERE D'ARCHIVAGE	50

2. PRESENTATION DE L'IRD

L'IRD (Institut de Recherche pour le Développement)¹ est un établissement public scientifique et technologique placé sous la double tutelle du Ministère de la recherche et du Ministère de la coopération.



Figure 1. : Les chiffres de l'IRD de l'année 2003²

Les recherches scientifiques de l'IRD sont centrées sur les relations entre l'homme et son environnement dans les régions tropicales et méditerranéennes, dans la perspective d'un développement durable de ces régions.

L'IRD mène des recherches en Afrique, en Asie, dans l'Océan Indien, en Amérique latine et dans le Pacifique.

Ses implantations sont réparties dans le monde :

- 23 implantations dans des pays de la zone intertropicale,
- 5 implantations en métropole (Paris, Bondy, Orléans, Montpellier et Brest),
- 5 implantations dans les DOM-TOM.

La documentation est un secteur qui dépend de la délégation à l'information et à la communication.

¹ Site web : www.ird.fr

² Source : rapport annuel 2003 de l'IRD

3. LE DOCUMENT NUMERIQUE DANS UN SYSTEME D'INFORMATION DOCUMENTAIRE

Le document véhicule de l'information sous différentes formes : texte, image, vidéo, données factuelles et expérimentales...Nous nous intéressons dans cette étude au document numérique textuel.

Le document est numérique car :

- il a été numérisé,
- il a été nativement produit sur support numérique.

3.1. Les techniques de numérisation

Il existe deux modes de numérisation apportant chacun des spécificités que nous rappelons.

Le mode image :

Le mode image produit un fac-similé numérique du document papier grâce au scanner. Ce mode permet de reproduire et d'étudier les illustrations, les photographies, l'écriture et la typographie.

Dans le domaine scientifique, le rendu des figures et des photographies est à prendre en compte. Cependant, le mode image ne permet pas de rechercher des caractères (par extension des unités documentaires) à l'intérieur d'un document textuel.

Le mode texte :

Le mode texte est obtenu par la technique de la reconnaissance optique de caractères (OCR³). A partir d'un document numérisé en mode image, le logiciel utilisé convertit la configuration des points en signes typographiques dont il déduit la valeur dans une table de codification comme l'ISO 8859-1 (ISO Latin-1) ou l'ISO 10646 (UNICODE).

Le mode texte permet une recherche plein texte dans le document. A noter toutefois que les outils de reconnaissance optique ne sont pas encore totalement fiables. Les fabricants de logiciels de reconnaissance optique de caractères améliorent régulièrement leurs produits et ont recours à des dictionnaires et outils d'analyse grammaticale.

L'utilisation du mode texte seul ne permet pas de conserver la mise en page.

³ Optical Character Recognition

Ainsi les deux modes sont généralement utilisés conjointement dans la numérisation de documents textuels.

On note que la numérisation ne produit aucune structure au document. Dans une problématique d'exploitation du document, la structuration peut avoir une importance.

3.2. Le document produit numériquement

Le document produit numériquement est celui réalisé par son auteur (chercheur, enseignant...) qui utilisera les outils de traitement de texte usuels (LaTeX, Open Office Texte ou Microsoft Office Word).

Si le format LaTeX permet d'associer un style au contenu du document, il n'en est pas de même pour les autres traitements de texte.

3.2.1. Les macros

Un document texte ".doc" ou ".sxd", utilisant une macro, sera structuré car l'auteur du document utilisera les styles pour caractériser les parties de son texte.

Ex :

Ceci est un titre1

Ceci est un titre2

Voici du texte

Ceci est un nouveau titre2

Encore du texte

Cette structuration n'est pas forte car elle n'intègre que les styles mais elle donne une structuration *a minima* d'un document numérique.

3.2.2. Le format XML

XML est un langage de structuration qui permet de décrire la structure du document et de repérer les différentes parties de ce document grâce à des méta-données. On parle de structuration forte.

Ex :

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
<head>
  <meta name="description" content="exemple de document structure">
  <meta name="Author" content="Diane Stitelet - Le Henaff">
  <meta name="Creation_date" content="juillet 2004">
  <title>Ceci est le titre</title>
</head>
<body>
  <part>
    <title>Ceci est un titre1</title>
    <chapter>
      <title>Ceci est un titre2</title>
      <p>Voici du texte (normal)</p>
    </chapter>
    <chapter>
      <title>Ceci est un nouveau titre2</title>
      <p>Encore du texte</p>
    </chapter>
  </part>
</body>
```

3.3. Avantages et limites de la structuration

L'exploitation du document passe par son traitement dans une application informatique. Pour pouvoir indexer et retrouver facilement un document dans un corpus, il convient d'en extraire des éléments descriptifs.

Un document XML détient ces éléments directement dans les méta-données. L'auteur du document pourrait alors encoder directement dans son document des éléments de catalogage et d'indexation. Y. Prié⁴ dans [PRI 00] préconise l'annotation des documents par l'auteur et les différents acteurs (re-lecteurs, comité éditorial...). Cette annotation serait inscrite directement dans le document lui-même.

Si un tel scénario peut séduire, il paraît impensable que dans un avenir proche, les auteurs de documents, qu'ils soient chercheurs ou enseignants, structurent leur document en XML. En effet, la simple utilisation des styles n'est toujours pas généralisée et nous ne voyons pas comment les habitudes pourraient changer en si peu de temps. Cependant, les logiciels de traitement de texte commencent à intégrer un formatage XML natif et pourraient accompagner un changement plus rapide des habitudes.

⁴ Yannick PRIÉ préfère le terme d'annotation à celui d'indexation car l'annotation décrit le document en vue de n'importe quelle tâche d'exploitation.

On peut donc penser qu'à terme, le document pourrait être déposé directement dans sa version XML native. Si cette option semble intéressante, le problème risque d'être lié à l'hétérogénéité des schémas XML qui ne faciliteront pas le traitement du document.

En effet, dans le domaine de la documentation, on dénombre déjà de nombreuses DTD⁵ :

- ✓ ISO 12083 : <http://www.xmlxperts.com/12083xml.htm>
- ✓ ISO 12083 pour les monographies et les imprimés : <http://www.xmlxperts.com/xmlbookdtd.htm>
- ✓ ISO 12083 pour les périodiques : <http://www.xmlxperts.com/xmlserialdtd.htm>
- ✓ ISO 12083 pour les articles de périodiques : <http://www.xmlxperts.com/xmlarticledtd.htm>
- ✓ ISO 12083 pour les formules mathématiques : <http://www.xmlxperts.com/xmlmathdtd.htm>
- ✓ TEI (Text Encoding Initiative) : structuration de textes électroniques, notamment en sciences humaines et linguistique : <http://www.tei-c.org>
- ✓ EAD (Encoded Archival Description) : traitement des fonds d'archives et des collections spécialisées organisées en fonds : <http://lcweb.loc.gov/ead/>
- ✓ EBIND pour lier les métadonnées au document numérique : <http://sunsite.berkeley.edu/Ebind>
- ✓ MASTER pour les manuscrits médiévaux : <http://www.cta.dmu.ac.uk/projects/master/>
- ✓ MathML : <http://www.w3.org/TR/REC-MathML/>
- ✓ ONIX : <http://www.editeur.org/onix.html>
- ✓ XMLMARC : <http://laneweb.stanford.edu:2380/wiki/medlane/xmlmarc>
- ✓ MARCXML : <http://www.loc.gov/standards/marcxml/>
- ✓ BiblioML (bibliographique et autorités) : <http://www.biblioml.org/fr/index.html>

Cet aspect du document structuré est important lorsque l'on souhaite mettre en place une procédure de dépôt du document à travers le système d'information documentaire.

⁵ Définition de Type de Document

3.4. Le système d'information documentaire

3.4.1. Définition

D'un point de vue global, un système d'information est une modélisation de l'activité d'une organisation permettant de gérer les processus (production, vente..).

Le système d'information offre les outils technologiques à la mise en œuvre de procédures au sein d'une organisation. Le terme "documentaire" associé au système d'information permet de délimiter les flux aux documents mais la prise en compte des différents acteurs mobilisés et des processus, reste intrinsèquement liée au système d'information.

Y. Prié propose qu'un système d'information documentaire ait pour objectif de gérer un ensemble de documents en vue de leur utilisation, au cours d'une collaboration entre l'homme et le système [PRI 00]. Cette approche est intéressante car elle rappelle que le système d'information est une application informatique qui :

- modélise des procédures formalisées par des acteurs,
- offre les outils répondant au besoin des utilisateurs.

Le système d'information documentaire est-il une archive ?

Une archive, telle qu'elle a été définie par les archivistes, renvoie à la notion de conservation.

Gabriel Gallezot, Christian Rossi, Ghislaine Chartron et Jean-Max Noyer ajoutent la notion de préservation sur le long terme associée à une politique institutionnelle [GAL 03].

La politique en information scientifique et technique visant à valoriser la production interne d'un institut de recherche conduirait à mettre un place un projet de GED⁶ plutôt qu'une archive traditionnelle. C'est le choix que l'IRD a fait en déployant le système d'information documentaire Ex-Libris qui intègre un module de GED.

Le système d'information documentaire est-il une base de données ?

D'après le Dictionnaire encyclopédique de l'information et de la documentation⁷, *l'expression "base de données" est utilisée en informatique pour décrire les systèmes de gestion de données et de leurs relations.*

Le terme de "systèmes de gestion de données" prête à confusion car il évoque la possibilité de travailler sur les flux. Or une base de données est souvent perçue en documentation comme un entrepôt de données archivées, ce qui lui ôte toute signification de flux.

⁶ Gestion électronique des documents.

⁷ Op. Cit. Serge Cacaly (dir.) – Paris : Nathan, 2001. 634p. – ISBN : 2-09-191252-2

3.4.2. Le déploiement d'un système d'information documentaire

La mise en place d'un système d'information relève d'une politique claire et motivée nécessitant des compétences spécifiques ainsi que des moyens techniques et financiers.

Une équipe projet pourra ainsi réunir des expertises documentaires avec des compétences en informatique et en ingénierie des systèmes d'information. Les moyens techniques relèvent surtout de l'infrastructure informatique de l'organisation. Pour les institutions publiques, le coût élevé d'un système d'information nécessitera le lancement d'un appel d'offre public.

Le déploiement d'un système d'information suit en général une méthodologie de gestion de projet [BUS 03] [MAR 02]. Celle qui est décrite ci-dessous est proposée par la société Cadic pour le déploiement de leur système d'information documentaire Ex Libris.

A noter qu'une étape préalable doit permettre de formaliser les procédures des différentes tâches documentaires ainsi que des objectifs visés par le changement de système. L'étude des besoins, l'étude de l'existant, la rédaction du cahier des charges, le lancement de l'appel d'offre puis le choix du prestataire ont été réalisées antérieurement au déploiement. Nous ne décrivons donc pas ces étapes dans cette étude.

1ère phase du projet : lancement du projet documentaire

Cette phase consiste essentiellement en une rencontre des deux parties (le prestataire et l'équipe projet interne de l'organisation, à savoir la maîtrise d'oeuvre, l'assistance et la maîtrise d'ouvrage) afin d'établir des règles de communication et un planning prévisionnel.

2^{ème} phase du projet : installation de l'infrastructure

L'IRD comporte plusieurs implantations en métropole et dans les pays en voie de développement. Le choix politique de l'IRD a été de fédérer toutes les tâches documentaires des différents centres de documentation autour d'un nouveau système. Ainsi, il est nécessaire, en vue de cette ré-informatisation complète, de sélectionner certains centres "pilotes" de l'IRD qui permettront de tester les serveurs, les synchronisations et les procédures du prestataire.

Un serveur central est installé en métropole ainsi que des serveurs relais dans les différents centres pilotes de l'IRD.

3^{ème} phase du projet : récupération des données

En vue de la migration des données, le chef de projet documentaire devra fournir la liste des tables de l'ancien système : les relations entre elles, leur nature, le nombre d'enregistrements par table ainsi qu'une description fonctionnelle des attributs.

La phase de récupération des données consiste à récupérer les données des champs de la base source vers les champs de la base cible.

Aucun traitement n'est effectué sur les données elles mêmes. Si un travail est nécessaire, il doit être fait préalablement car le prestataire n'offre que des outils de modification des données, de façon globale, à posteriori.

La récupération des données s'effectue en plusieurs étapes :

- export des données dans un format compatible avec le nouveau système,
- élaboration d'un schéma de récupération des données,
- migration des données,
- test de validation.

A noter que certains systèmes du marché ne proposent pas de base de données relationnelles. Une base relationnelle permet une structure dynamique qui offre des outils statistiques et une plus grande puissance dans la recherche d'information.

En effet, une recherche sur l'éditeur "Elsevier" génèrera une interrogation de l'ensemble des notices de la table unique alors que dans un modèle relationnel, "Elsevier" sera recherché dans la table spécifique "éditeur" contenant beaucoup moins de données. Les notices associées à "Elsevier" seront alors listées grâce à la relation entre les tables "éditeur" et "notice".

La connaissance de la structure de la base est fondamentale, c'est cette structure qui va créer les contraintes éventuelles dans la récupération des données.

Si l'institution utilise un thésaurus ou des listes d'autorité, ceux-ci seront récupérés en format de fichier compatible avec les normes AFNOR (Z47100 pour les thésaurus).

4^{ème} phase du projet : paramétrage des vues professionnelles

Un certain nombre de vues professionnelles peuvent être livrées. Elles correspondent aux besoins de l'institution en terme de workflow⁸ ou/et de gestion documentaire.

Les vues professionnelles peuvent correspondre :

- au dépôt de document par le chercheur,
- à la création d'une notice par les différentes documentalistes ou/et le chercheur,
- à l'indexation,
- à la révision et validation,
- à l'exemplarisation,
- à l'acquisition de documents par le ou les centres de documentation.

Ces phases doivent être validées à chaque étape à la suite d'une série de tests. Des administrateurs (documentalistes) seront formés afin de pouvoir apporter par la suite des modifications à ces paramétrages.

5^{ème} phase du projet : paramétrage du portail documentaire et réalisation des vues publiques

Des vues publiques sont réalisées et livrées par le prestataire en réponse au cahier des charges. Ces vues seront accessibles à partir d'un portail documentaire. Cette phase va permettre de rédiger les spécifications du portail documentaire en y associant les vues publiques et éventuellement d'autres services documentaires.

L'ensemble des services documentaires ne pourra pas être proposé aux utilisateurs. Les spécifications devront ainsi prendre en compte :

- la gestion des droits d'accès,
- les profils utilisateurs,
- la connexion avec des bases externes,
- les services documentaires

Par exemple, le dépôt de document scientifique dans le fonds documentaire de l'IRD n'est proposé qu'aux agents affiliés à l'institution (chercheurs, documentalistes...). Ainsi, ce service n'apparaîtra sur le portail documentaire qu'une fois l'utilisateur authentifié.

⁸ Le Workflow est un système d'ordonnancement des flux de travaux dans une organisation. Cf page 17.

6^{ème} phase du projet : formations

Une série de formations professionnelles sont proposées par le prestataire :

- aux administrateurs qui seront susceptibles de modifier le paramétrage,
- aux différents acteurs du système d'information (documentalistes).

7^{ème} phase du projet : déploiement et passage en production

Les étapes antérieures ont permis de tester les équipements et les procédures et d'évaluer les vues professionnelles et publiques livrées.

Le passage en production implique, à ce stade, l'ensemble des acteurs impliqués dans le projet. Il s'agit alors de généraliser l'installation et l'utilisation du système sur l'ensemble des sites.

L'équipe projet effectue ensuite une vérification de service régulier (VSR). Si celle-ci est satisfaisante pour l'ensemble des sites, 35% du montant total restant sera alors versé au prestataire. Cette dernière phase met un terme à l'engagement contractuel.

3.5. Le workflow

3.5.1. Définition

L'APROGED⁹ définit le workflow comme "un système d'ordonnancement des flux de travaux dans une organisation. A la base du workflow, il y a une modélisation des fonctions et procédures de l'organisation."

Le système d'information tel que nous l'avons défini précédemment est une modélisation de l'activité d'une organisation permettant de gérer les flux. Le workflow est un donc un élément important d'un système d'information.

3.5.2. Formalisation des procédures

L'IRD a formalisé les procédures relatives à la gestion des documents IRD qui constituent le fonds documentaire dans son cahier des charges¹⁰, en vue de la ré-informatisation du fonds. Ces procédures de fonctionnement sont décrites ci dessous :

Dépôt du document :

- 1) Le chercheur IRD fournit un exemplaire — papier ou électronique — de son document à son centre de rattachement. Les documents en format papier sont adressés dès leur réception par les centres de rattachement au centre de Bondy.
- 2) Le chercheur IRD saisit directement un formulaire d'envoi et joint un fichier électronique. Ce formulaire est adressé au centre de Bondy.

Création d'une notice bibliographique :

Avant toute création de notice, une recherche préalable est effectuée pour éviter les doublons. Si la référence du document existe déjà, il suffira de créer une fiche exemplaire.

Le centre de document qui réceptionne le document crée une notice minimale (pré-catalogage) et transmet le document physique au centre de documentation de Bondy qui contrôle la notice, recherche d'éventuels doublons et procède aux corrections si nécessaire, affecte un numéro d'inventaire et une cote.

Les documents ainsi réceptionnés sont ensuite adressés tous les deux mois au prestataire qui procède au catalogage et à l'indexation documentaire. Le prestataire retourne les documents physiques ainsi qu'un listing des notices, sous un délai d'environ 1 mois. Les notices sont contrôlées et corrigées manuellement. C'est le prestataire qui effectue alors les corrections et adresse une disquette contenant les notices complètes. Le délai total de prise en compte d'un nouveau document est

⁹ Association des professionnels de la GED : <http://www.aproged.org/>

¹⁰ Appel d'offre : Système d'information documentaire, CCTP n°03-02 du 3 avril 2003

de l'ordre de deux mois, corrections comprises. Les documents ont à ce stade de la procédure le statut "provisoire".

Contrôle de la qualité des notices réceptionnées :

Le contrôle de la qualité des notices est effectué par sondage et par lot. Le contrôle porte sur 5 à 20 % des documents fournis, échantillon représentatif des domaines traités dans le lot. C'est le centre de documentation de Bondy qui effectue ce contrôle.

Si l'échantillon contrôlé ne comporte aucune erreur ou des erreurs jugées minimales, le lot est accepté : dans ce cas, les notices du lot prennent le statut "validé".

Dans le cas contraire, les notices du lot reprennent le statut "provisoire". Le prestataire doit alors reprendre toutes les notices du lot. Les nouvelles notices modifiées ou recrées annulent celles qui sont provisoires.

La formalisation des procédures concernant le traitement du document fait apparaître les besoins en terme de workflow. On se rend compte que le dépôt du document n'est pas simplement l'action d'importer un fichier qui sera stocké dans une base de données. Le document va subir un traitement impliquant différents acteurs (documentalistes, prestataires...). La procédure doit être non seulement formalisée mais doit aussi pouvoir s'inscrire dans l'ensemble des procédures documentaires internes.

La société Cadic a élaboré, pour l'IRD, un workflow sur la base des procédures décrites dans l'appel d'offre.

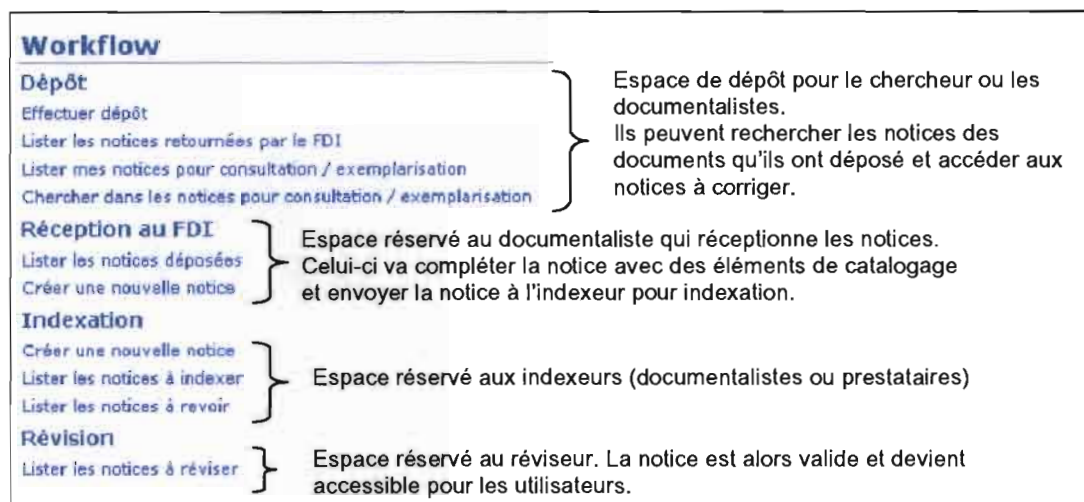


Schéma 1. : Workflow élaboré par la société Cadic pour l'IRD

3.6. Les droits d'accès

La gestion des droits d'accès se situe à plusieurs niveaux.

Au niveau du document :

Certains documents du fonds documentaire peuvent être accessibles uniquement sur l'Intranet.

Pour ce faire, il s'agit d'associer au document une référence permettant d'identifier "Intranet" (interne) ou "Internet" (public). Cette référence sera transmise avec la requête dans l'URL.

Au niveau des fonctionnalités proposées à travers le portail documentaire :

Le portail documentaire a vocation à offrir l'ensemble des services documentaires d'une institution.

Cependant, certains services aux usagers ne concernent que les utilisateurs affiliés à l'institution (ex : bases de données, revues en ligne...).

De même, le portail documentaire pourrait également être le point d'accès aux tâches documentaires des professionnels à travers des vues professionnelles.

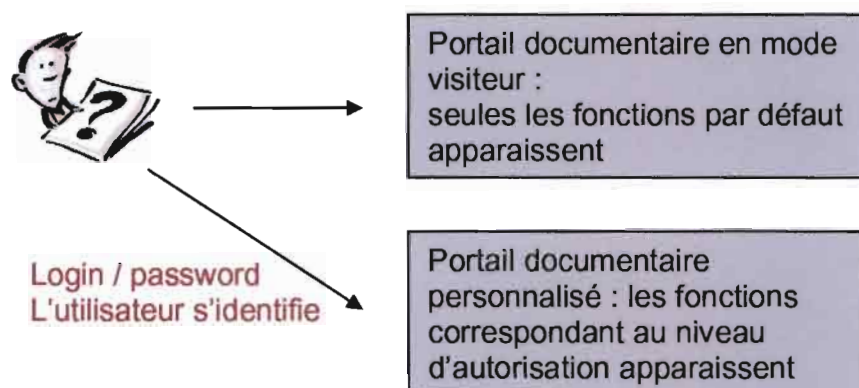


Schéma 2. : L'utilisateur se connecte au portail documentaire

La gestion des droits d'accès aux fonctions d'un portail documentaire est possible en associant, à un serveur de contrôle des identifiants, une table dans laquelle sont inscrits les membres affiliés à l'institution.

Cette table des utilisateurs est mise à jour par une synchronisation avec un annuaire LDAP¹¹.

¹¹ LDAP est le protocole d'annuaire sur TCP/IP. Les annuaires permettent de partager des bases d'informations sur le réseau interne ou externe. Ces bases peuvent contenir toute sorte d'information que ce soit des coordonnées de personnes ou des données systèmes. Pour les spécificité techniques, voir le site : <http://www-sop.inria.fr/semir/personnel/Laurent.Mirtain/ldap-livre.html>

3.7. Le portail documentaire

Le portail documentaire est l'interface permettant d'interagir avec les utilisateurs.

Ceux-ci pourront accéder aux services de la documentation mais également interroger le fonds documentaire.

3.7.1. Interrogation du fonds documentaire

Le schéma 3 présente la modélisation du cheminement d'une requête effectuée par un utilisateur. Celle-ci va être traitée par le système d'information.

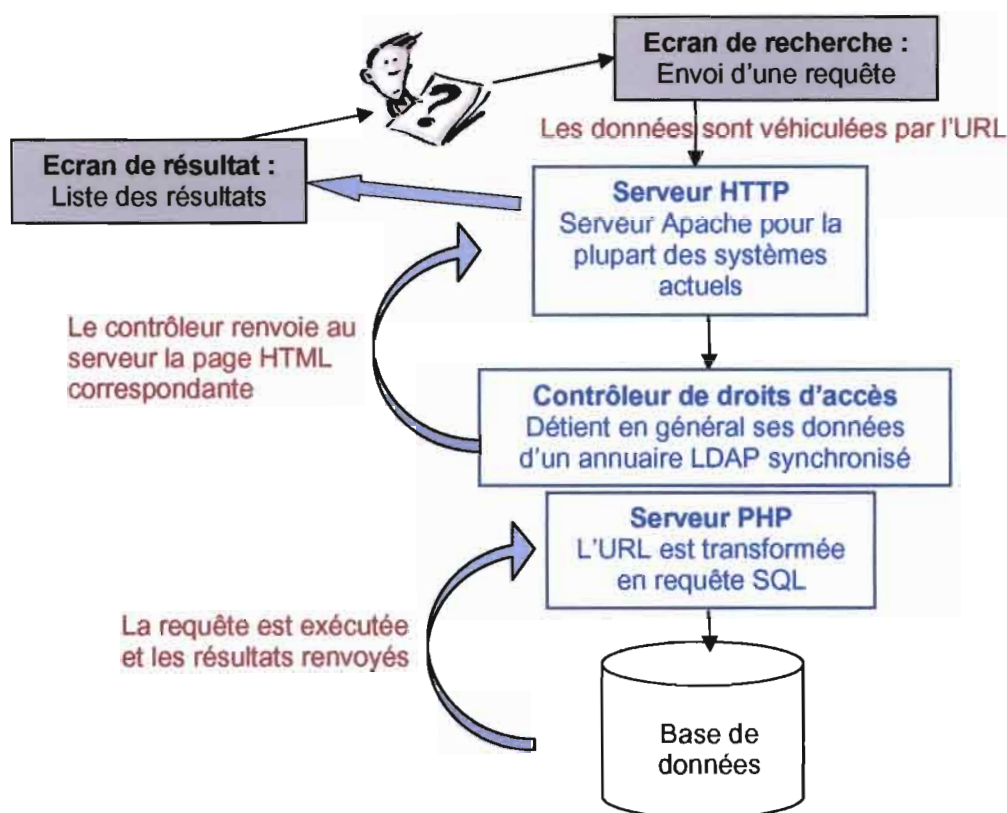


Schéma 3. : La requête au sein du système d'information documentaire

3.7.2. Visibilité des documents

L'exploitation des documents ne doit pas être la seule préoccupation des documentalistes. Leur objectif est la valorisation du fonds.

La valorisation du fonds est obtenue grâce à :

- une forte visibilité des documents,
- la qualité de l'indexation,
- la performance du moteur de recherche.

Auparavant les documents des archives ou des bases de données rentraient dans la catégorie du "Web invisible".

Pour améliorer la visibilité des documents, il existe un protocole d'échange de méta-données qui permet une interopérabilité entre les systèmes compatibles.

Si de plus en plus de systèmes d'information documentaire ou d'archives scientifiques adoptent ce protocole OAI-PMH¹², c'est le référencement par les moteurs de recherche (Google et Yahoo) qui offrent la visibilité souhaitée.

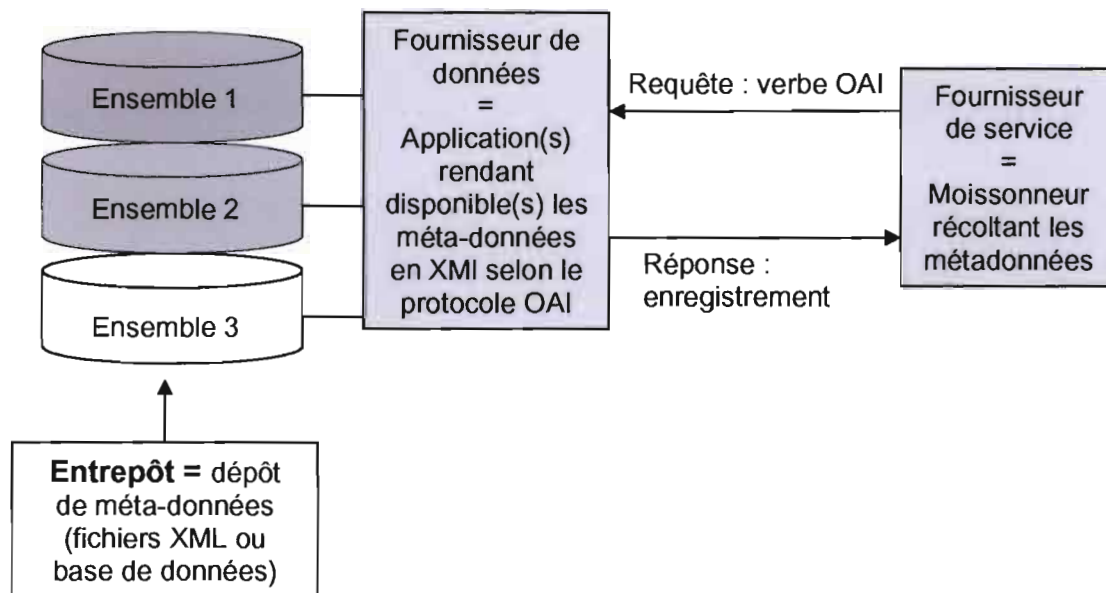


Schéma 4. : Schéma présentant l'échange de données suivant le protocole OAI.

La qualité de l'indexation et la performance du moteur de recherche sont traitées dans le chapitre suivant.

3.7.3. Services documentaires

Les systèmes d'information documentaire du marché proposent très souvent des services intégrés à la recherche documentaire :

- espace personnel dans lequel on retrouve l'historique de ses recherches ou ses alertes,
- une fonction d'export des références bibliographiques vers des logiciels de gestion des références ou un export XML des notices.

Ces services sont très appréciés par les utilisateurs et constituent un atout important dans leur fidélisation.

¹² Le protocole OAI-PMH signifie "Open Archive Initiative Protocol for Metadata Harvesting". Les spécifications sont disponibles sur le site <http://www.openarchives.org/>

4. LE SYSTEME DE RECHERCHE D'INFORMATION (SRI)

Stocker des documents dans un système de gestion électronique de documents (GED) n'est pas une finalité en soi. Les documents stockés sont conservés en vue d'une utilisation. Les documents doivent donc pouvoir être retrouvés le plus facilement et le plus efficacement possible.

Si le système d'information documentaire permet de modéliser les procédures de traitement de la chaîne documentaire, il doit également permettre d'interagir avec des utilisateurs en faisant correspondre les documents stockés au besoin d'information d'un utilisateur.

Si certains considèrent le système de recherche d'information uniquement comme un moteur de recherche et d'indexation, G. Salton et J. Mc Gill [SAL 83] définissent le système de recherche d'information de façon plus complète : *"Le SRI traite de la représentation, du stockage, de l'organisation et de l'accès à des unités documentaires."*

Le système de recherche d'information est donc complètement intégré au système d'information documentaire et se caractérise par :

- un module chargé du traitement de l'indexation et du stockage de l'information ;
- un portail ou une interface de recherche permettant de recevoir les requêtes des utilisateurs.

La connaissance des techniques d'indexation et de recherche est indispensable à la compréhension du fonctionnement des systèmes de recherche d'information du marché.

Les éditeurs ne décrivent pas ces techniques dans leurs plaquettes de présentation, ni dans les manuels d'utilisation ; c'est pourquoi, ce chapitre sera consacré à réaliser un état de l'art des techniques d'indexation et de recherche. Ainsi, les professionnels de l'information pourront mieux appréhender cet aspect technique et inclure des spécificités dans leur cahier des charges en vue de l'acquisition d'un système de recherche d'information.

Enfin, l'évaluation du système de recherche d'information nécessite une série de tests sur un corpus de documents connus et identifiés afin de connaître les différentes mesures d'efficacité. Ces mesures seront présentées en fin de chapitre.

4.1. L'indexation

Indexer un document revient à extraire des éléments censés représenter au mieux son contenu.

L'indexation peut revêtir deux sens qu'il convient d'explicitier :

- L'indexation, au sens documentaire est l'opération visant à la représentation du contenu d'un document par affectation manuelle ou automatique d'une série de mots clés.
- L'indexation, au sens informatique est l'opération de création d'index en vue de la recherche documentaire. Le contenu de l'index sera nommé alternativement entités d'indexation ou descripteurs. Les candidats seront plutôt nommés mots clés, termes ou concepts.

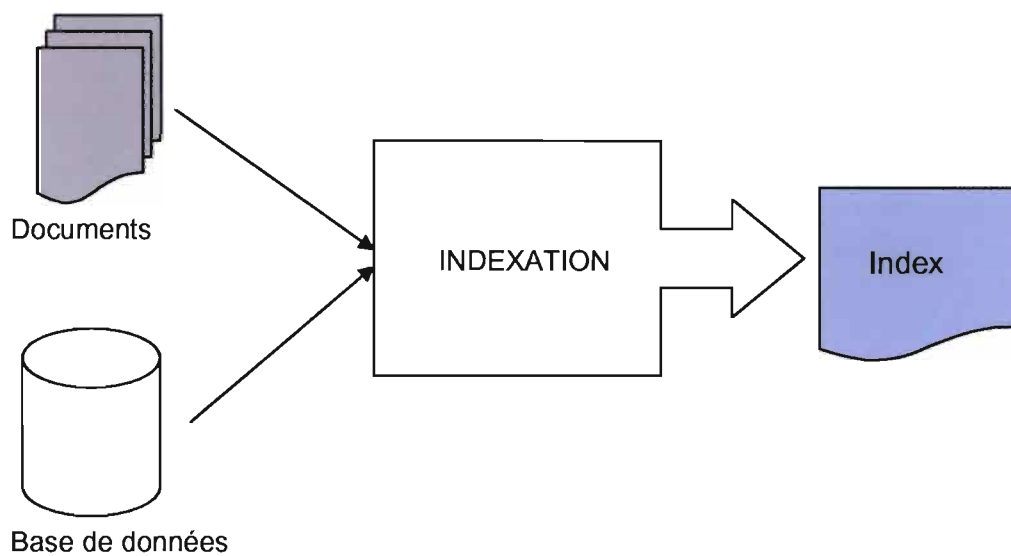


Schéma 5. : Processus de création d'un index

4.2. Les techniques d'identification des entités d'indexation

"Les processus d'indexation utilisés dans le domaine de la documentation automatique cherchent à trouver dans un texte, par une analyse morphologique, lexicale et syntaxique, les éléments de sens caractéristiques ; l'indexation de ces termes fournit une liste qui peut être considérée comme une représentation du texte, pertinente pour donner une idée de son contenu" [BAL 96].

Les techniques d'indexation étudiées dans ce paragraphe sont celles qui déterminent de façon automatique les entités d'indexation décrivant le contenu sémantique d'un document.

Si presque tous les systèmes actuels procèdent à une analyse morphologique, lexicale et syntaxique en vue de créer des index sur des corpus de documents, les techniques évoluent et la recherche dans l'analyse sémantique est très active.

4.2.1. Simplification du texte

La première étape segmente le texte en retirant les mots "vides" (articles, prépositions...) et ne conserve que des phrases et des mots.

Ensuite, certains mots sont supprimés car ils sont considérés comme trop fréquents (article, adverbe, pronom...) et donc non significatifs. Ces mots sont issus d'une liste appelée "liste des mots stop" [SAL 84].

L'index est ainsi constitué d'entités d'indexation (mots ou groupes de mots du texte) sans traitement particulier, ce qui implique deux inconvénients :

- la taille de l'index est importante (quasiment proportionnelle à la longueur du texte),
- les descripteurs n'ont pas de poids particulier signifiant leur degré d'importance dans la description sémantique du document.

Pour résoudre ces inconvénients, certains systèmes de recherche d'information intègrent une pondération des descripteurs basée sur la fréquence d'apparition du descripteur et sa position dans le document (titre, corps du texte...).

4.2.2. Traitements linguistiques

➤ Analyse morphologique

L'analyse morphologique consiste à réduire une chaîne de caractères en mots ou groupe de mots reconnus par leur existence dans un lexique (dictionnaire, thesaurus) puis à réduire ces termes sous leur forme canonique.

L'index est ainsi constitué de lemmes.

L'analyse morphologique a aussi pour objectif de reconnaître les expressions idiomatiques comme "chemin de fer" ou "avoir tort".

➤ Analyse syntaxique

L'analyse syntaxique est basée sur la règle qu'une phrase simple est composée d'un seul verbe conjugué.

Outre l'élimination de certains mots "pauvres" (car, aussi,...), cette règle permet de lever des ambiguïtés lors de la présence d'homographes (nom et verbe) dans le texte : "l'homme approcha une chaise de la table", table étant considéré comme un nom et pas le verbe conjugué de son infinitif "tablér".

➤ Analyse phonétique

Essentiellement utilisée pour la recherche sur des noms propres, l'analyse phonétique va s'efforcer de dissocier les mots en phonèmes¹³ en leur appliquant des algorithmes de recherche.

Ex : Mao Tse Toung peut s'écrire de plus de 2000 façons différentes.

➤ Analyse sémantique

L'analyse sémantique recherche des relations de dépendance entre les syntagmes nominaux contenues dans un autre syntagme nominal¹⁴

Ex.15 : "absorption d'une particule par le noyau" est décomposé en deux syntagmes "absorption de particule" et "absorption par le noyau"

Cette approche nécessite un apport en terme de règles (dictionnaire, thesaurus et ontologies) à intégrer au système. En l'état actuel des techniques, un système ne peut traiter qu'une discipline particulière mais en aucun cas un domaine pluridisciplinaire.

Par ailleurs, ces connaissances doivent être mises à jour avec l'arrivée de nouveaux concepts. Cette maintenance de la base de connaissance constitue un coût important pour l'institution.

¹³ Le phonème est une unité sonore non significative de langue qui n'est pas susceptible d'être dissociée en unités plus simples.

(Dictionnaire de linguistique et des sciences du langage, Paris, Larousse, 1994)

¹⁴ Un syntagme nominal se compose d'un nom, précédé d'un article, ou déterminant, et d'un nombre quelconque d'adjectifs.

(L'Instinct du Langage, Odile Jacob, 1999, page 95)

¹⁵ Op. cit. p132 dans [ROU 99]

4.3. Les techniques de construction des index

Pierre Le Loarer dans [LOA 94] présente les quatre types d'indexation qui conduisent à la construction d'index dans des systèmes de recherche d'information.

4.3.1. Indexation à plat

L'indexation dite à plat considère que les descripteurs ont tous le même statut vis à vis du texte à indexer et qu'ils entretiennent les mêmes rapports entre eux. La représentation du texte est donc une succession non ordonnée d'entités d'indexation.

4.3.2. Indexation pondérée

L'indexation pondérée permet de donner à chaque descripteur un niveau d'importance permettant de différencier le sujet principal et les thèmes secondaires d'un texte.

Ce poids est calculé à l'aide de fonctions mais il est généralement proportionnel à la fréquence d'apparition du descripteur dans le document. Plus le poids est important, plus le descripteur est jugé apte à appartenir à l'index du document [ROU 99].

Si w_{ij} = poids du terme T_j dans le document D_i

- tf_{ij} = fréquence du terme T_j dans le document D_i
- N = nombre de documents dans la collection
- n = nombre de documents où le terme T_j apparaît au moins une fois

Alors le poids peut être calculé comme suit :

$$w_{ij} = tf_{ij} \cdot \log_2 \frac{N}{n}$$

Ainsi, un descripteur à forte pondération est un terme fréquent dans un document (tf_{ij}) et absent dans les autres documents du corpus (N/n).

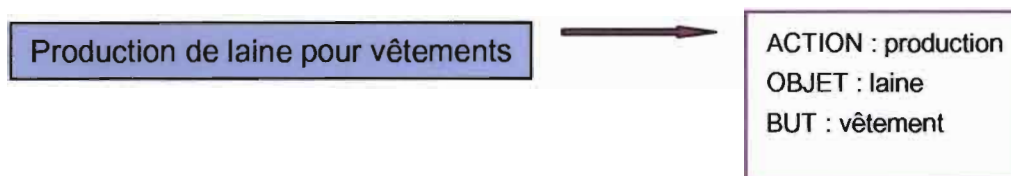
Il existe d'autres fonctions de pondération, notamment celle utilisée dans le système SMART de Salton [SAL 75] ou celle utilisée dans le logiciel SPIRIT. Cette étude n'a pas vocation à faire un état de l'art exhaustif de la recherche en indexation.

Si cette méthode semble plus judicieuse que l'indexation à plat, il convient de signaler que les termes étrangers, les noms propres ou les mots mal orthographiés prennent aussi de l'importance dans l'index.

4.3.3. Indexation à rôles

L'indexation à rôles établit des relations sémantiques entre les descripteurs. Elle consiste à affecter un rôle aux descripteurs pour définir les relations sémantiques qui les associent. Les rôles sont définis dans le contexte du corpus de documents et du domaine considéré.

Ex¹⁶ : En considérant les rôles : ACTION, OBJET, BUT



4.3.4. Indexation structurée

L'indexation structurée généralise l'indexation à rôles en y intégrant de nouvelles propriétés.

4.4. Application à l'indexation automatique

Dans un centre de documentation, l'indexation de documents est réalisée par des documentalistes. Or l'indexation manuelle a un coût, celui du temps de travail. Est-elle pour autant plus efficace et plus cohérente que l'indexation automatique ?

Deux personnes différentes ont moins de 20 % de chances de choisir spontanément le même terme pour décrire un objet, alors que l'indexation automatique apporte de la cohérence dans l'indexation [FUR 87].

La question de l'indexation automatique prend donc tout son sens et le déploiement d'un système d'information documentaire offre la structure permettant de mettre en place de nouvelles procédures intégrant l'automatisme.

Les techniques d'identification des entités d'indexation et les techniques de construction des index décrites dans les paragraphes précédents ont montré que les recherches dans ce domaine ont beaucoup évolué.

Cependant, le langage est complexe et ne peut pas être complètement modélisé. Il conviendra donc d'utiliser l'indexation automatique comme indexation préalable qui devra être revue et corrigée par des professionnels.

¹⁶ Exemple tiré de [ROU 99]

4.5. La recherche d'information

La recherche d'information s'appuie sur deux étapes intimement liées : l'indexation et la gestion de la requête.

D'un côté se trouve l'utilisateur avec un besoin d'information qu'il va modéliser à travers une requête et de l'autre se trouve le corpus de documents représenté par un ensemble de descripteurs.

Le problème sous-jacent à la recherche d'information textuelle est un problème linguistique. En effet, il s'agit de faire correspondre un besoin que l'utilisateur exprime avec ses propres termes et les textes réponses qui sont exprimés dans d'autres termes que ceux choisis par l'utilisateur dans sa requête.

Les techniques de recherche d'information utilisent la notion de comparaison entre la représentation d'un document et celle d'une requête.

Le processus de comparaison, que C. Roussey et al. nomment la fonction de comparaison dans [ROU 99], va permettre d'établir la correspondance entre les documents et la requête exprimant un besoin d'information.

La fonction de comparaison C va s'appliquer à deux espaces de représentation, la base des index du corpus documentaire et celui représentant la requête. Elle peut donner deux types de résultat :

- elle sépare le corpus en deux parties (approche booléenne) : les documents pertinents ou non ;
- elle ordonne le corpus, du document le plus pertinent au moins pertinent.

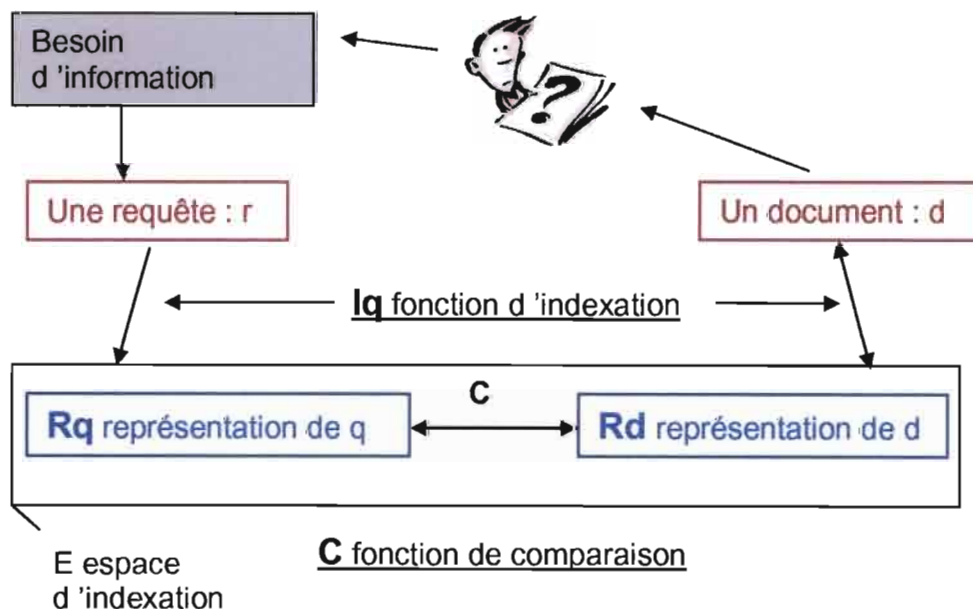


Schéma 6. : Schéma de la recherche documentaire

4.6. Les modèles de recherche d'information

4.6.1. Le modèle booléen

Le modèle booléen tire son nom du mathématicien britannique George Boole, à l'origine de la logique moderne.

Un document est sélectionné si -et seulement si- il satisfait à l'expression booléenne (VRAI ou FAUX). L'inconvénient majeur de ce modèle est l'absence d'ordonnement des documents résultats [SAL 88]. C'est pourquoi, Gérard Salton propose le principe du modèle booléen pondéré [SAL 84] qui affecte des poids aux descripteurs, tant au niveau de l'indexation des documents que celle de la requête¹⁷.

4.6.2. Les systèmes automatiques

Le modèle vectoriel :

Ce modèle, introduit par Gérard Salton [SAL 83] est utilisé pour calculer la proximité des requêtes et des documents. Documents et requêtes sont exprimés comme des vecteurs de termes

R : modélisation vectorielle de la requête
D_n : modélisation vectorielle du document n

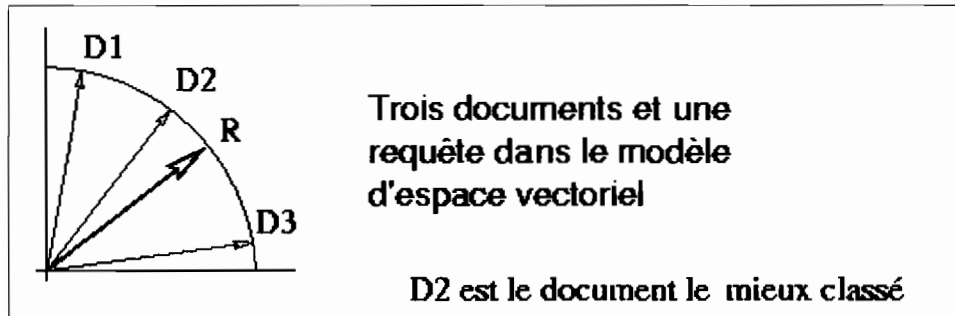


Figure 2. Représentation d'une requête dans un espace de documents [JACQUEMIN]

Le calcul de la distance entre les vecteurs document et requête, par la fonction cosinus, permet de déterminer le ou les documents qui correspondent le mieux à la requête.

¹⁷ Le calcul du poids des descripteurs a été traité dans le paragraphe 4.3.2 Indexation pondérée

Le degré de correspondance entre une requête et un document peut également être déterminé par une fonction de similarité.

Voici une des façons de calculer la similarité entre deux vecteurs :

$\text{Sim}(Q, D_i)$ = similitude entre la requête Q et le document D_i

- d_{ij} = poids du terme T_j dans le document D_i
- w_{qj} = poids du terme T_j dans la requête Q

$$\text{Sim}(Q, D_i) = \frac{\sum_j w_{qj} \cdot d_{ij}}{\sqrt{\sum_j (d_{ij})^2 \cdot \sum_j (w_{qj})^2}}$$

Schéma 7. : Fonction de similarité entre la requête Q et le document D_i

Le "Latent Semantic Indexing" :

Alors que le modèle précédant utilise les descripteurs pour représenter le contenu d'un document, la technique du "Latent Semantic Indexing (LSI)" se base sur un espace vectoriel différent.

Ce principe part de la constatation que deux mots de sens proche sont répartis de la même manière dans des documents.

En représentant les paragraphes comme des vecteurs, on construit une matrice de distribution des mots sur les paragraphes où la valeur est la fréquence du mot i dans le paragraphe j .

Ainsi la nouvelle représentation contient beaucoup moins de "bruit" que celle basée sur les mots clés au sein du document et du corpus de documents.

4.6.3. Le modèle probabiliste

Le modèle probabiliste considère la recherche d'information comme un espace d'évènements possibles. Un événement peut être le jugement de pertinence porté par l'utilisateur sur un document par rapport à une requête.

On construit ainsi un jeu de données initiales à partir de tests effectués sur un corpus de documents connus et identifiés.

Les documents sont ainsi représentés sous la forme d'un index pondéré où les poids des descripteurs correspondent à la probabilité que le descripteur soit pertinent pour le document.

4.7. L'évaluation d'un système de recherche d'information

Toutes les techniques d'identification des entités d'indexation présentées précédemment sont utilisées dans les systèmes de recherche d'information de façon complémentaire. La différence peut se faire sur l'analyse sémantique. L'efficacité d'un système de recherche d'information sera différente en fonction du degré de spécialisation du fonds documentaire et des listes d'autorités disponibles.

Dans les systèmes actuels, on retrouve plutôt le modèle vectoriel pondéré.

Pour évaluer un système de recherche d'information, on effectue des tests sur un corpus de documents connus :

Résultat du test	Documents pertinents	Documents non pertinents
Documents sélectionnés par le SRI	Documents pertinents retrouvés	Documents hors contexte
Documents non sélectionnés par le SRI	Documents oubliés	

Tableau 1. Evaluation d'un SRI : résultat d'un test

Grâce à cette répartition, on peut établir des mesures d'efficacité :

$$\text{Précision} = \frac{\text{Documents pertinents retrouvés}}{\text{Documents sélectionnés}}$$
 Parmi les documents récupérés par le SRI, la précision donne le taux de documents réellement pertinents. Plus la précision est bonne, plus le SRI est efficace.

$$\text{Bruit} = \frac{\text{Documents hors contexte}}{\text{Documents sélectionnés}}$$
 Parmi les documents récupérés par le SRI, le bruit donne le taux de documents non pertinents. Si le bruit est élevé, le SRI est particulièrement non performant.

Ex : La technique de construction des index, "l'indexation à plat", peut générer plus de bruit que l'indexation pondérée. Il s'agira alors de demander à connaître la technique utilisée avant d'acquiescer un système de recherche d'information.

$$\text{Rappel} = \frac{\text{Documents pertinents retrouvés}}{\text{Documents pertinents}}$$
 Les documents pertinents répondant à une requête sont connus. Le rappel mesure la capacité du SRI à retrouver ces documents dans le corpus. Plus le rappel est bon, plus le SRI est efficace.

$$\text{Silence} = \frac{\text{Documents oubliés}}{\text{Documents pertinents}}$$
 Les documents pertinents répondant à une requête sont connus. Le silence mesure l'incapacité du SRI à retrouver ces documents dans le corpus. Plus le silence est élevé, moins le SRI est performant.

5. LES DROITS D'EXPLOITATION DU DOCUMENT NUMERIQUE

Le système d'information documentaire permet de gérer et exploiter les documents numériques.

Le stockage, la reproduction sous forme numérique, et l'exploitation de ces documents nécessitent de déterminer si l'institution détient les droits correspondants.

Après avoir présenté les différentes notions de droit nécessaires à l'instruction de ce dossier, nous nous attacherons à traiter les différents types de documents constituant le fonds documentaire de l'IRD.

5.1. La propriété intellectuelle

Le droit d'auteur est un droit qui diverge de la notion anglo-saxonne de copyright.

Le droit d'auteur protège l'œuvre d'un auteur lorsque celle-ci est une création intellectuelle reflétant la personnalité de son auteur. Celui-ci "dispose, du seul fait de sa création, d'un droit de propriété incorporel exclusif et opposable à tous"¹⁸

Ce droit comprend des attributs d'ordre moral et des attributs d'ordre patrimonial.

➤ Le droit moral¹⁹

De nature extra patrimoniale, le droit moral est un droit de la personnalité. Il est perpétuel, inaliénable et imprescriptible.

Il comprend :

- Le droit de divulgation : l'auteur est seul juge du moment et de l'opportunité de livrer sa création au public. Le droit d'auteur est dévolu à celui sous le nom duquel l'œuvre est divulguée.
- Le droit au respect du nom : il impose l'indication du nom de l'auteur et à sa discrétion la mention de ses qualités lors de la divulgation
- Le droit au respect de l'œuvre : il implique le respect de l'intégration de l'œuvre, dans sa forme, dans son esprit, et dans ses détails.
- Le droit de repentir, de retrait : il permet de revenir sur sa décision de divulgation (sous réserve d'indemnisation du diffuseur ou de l'éditeur).

¹⁸ Article L.111-1 du code de la propriété intellectuelle

¹⁹ Op. Cit. Stéphane Marguerin, cours de droit. DESS IST (2003).

➤ Le droit patrimonial

L'auteur est investi d'un monopole d'exploitation de l'œuvre qui comprend le droit de représentation et de reproduction.²⁰

Le copyright ne considère pas l'objet créé comme une œuvre de l'esprit mais comme un objet exploitable et commercialisable.

Par ailleurs, pour bénéficier d'une protection au titre du Copyright, l'auteur doit réaliser certaines déclarations auprès d'un organisme, ce qui n'est pas le cas en droit français.

5.2. Régime juridique des agents de l'administration

La majorité des articles scientifiques sont écrits par des chercheurs relevant du régime des agents de l'administration ou des contractuels exerçant dans un institut de recherche public.

Le statut du chercheur par rapport à l'administration est différent selon qu'il s'agit d'un contractuel ou d'un fonctionnaire :

- Le chercheur est contractuel : il relève alors du droit régissant les salariés du secteur privé. Les modalités d'exploitation de l'œuvre sont incluses dans le contrat. Le salarié peut convenir d'une rétribution mais le droit patrimonial est généralement accordé à l'employeur.
- Le chercheur est fonctionnaire : il dépendra du projet de loi relatif au droit d'auteur et aux droits voisins dans la société de l'information (transposition de la directive communautaire 2001/29/CE du 22 mai 2001 et des traités de l'Organisation Mondiale de la Propriété Intellectuelle). Ce projet de loi a été présenté à l'Assemblée Nationale le 12 novembre 2003 par le ministre de la Culture, Jean Jacques Aillagon. Il sera examiné par le Parlement d'ici la fin de l'année 2004.

Le projet de loi relatif au droit d'auteur et aux droits voisins dans la société de l'information :

Concernant le régime juridique des agents de l'administration, cette loi va permettre de clarifier une situation ambiguë.

Le code de la propriété intellectuelle investit l'auteur des droits d'auteur sur les œuvres qu'il a créées mais l'avis Ofratème émis par le Conseil d'Etat le 21 novembre 1972 a marqué sa préférence pour la cession des droits patrimoniaux à l'administration pour les œuvres créées par ses agents dans le cadre de leur mission.

²⁰ Articles L.122-2 et L.122-3 du code de la propriété intellectuelle

Cette loi va accorder plus de droits aux auteurs lorsqu'une exploitation commerciale est en jeu.

Cependant l'article 17 de ce projet de loi limite l'exercice du droit d'auteur de l'agent de manière à ne pas entraver le fonctionnement du service public. L'article 18 prévoit que l'administration bénéficie d'une cession légale des droits patrimoniaux lorsque l'œuvre est exploitée pour la réalisation d'une mission de service public ne donnant pas lieu à exploitation commerciale.

Conséquence :

Dans le cadre de la recherche scientifique, l'administration peut donc exploiter la production du chercheur si celui-ci n'a pas cédé son droit d'exploitation lors d'un contrat éditorial.

La production de littérature scientifique est considérée comme une mission de service public dans les établissements publics de recherche et nous considérerons par la suite, que l'institution peut stocker, reproduire sous forme numérique et diffuser la production de littérature scientifique de ses chercheurs via son site Web institutionnel à des fins de valorisation.

5.3. Etablissement de contrats avec les éditeurs : notion de contrat d'édition

Nous rappelons que l'article L 132-1 du code de la propriété intellectuelle définit le contrat d'édition comme "le contrat par lequel l'auteur d'une œuvre de l'esprit ou ses ayants droit cèdent à des conditions déterminées à une personne appelée éditeur le droit de fabriquer ou de faire fabriquer en nombre des exemplaires de l'œuvre, à charge pour elle d'en assurer la publication et la diffusion".

Les éléments caractéristiques mentionnés ci-dessus précisent que le contrat d'édition est une cession de tout ou partie du droit de reproduction. La notion de licence (ou autorisation) n'est pas reconnue²¹.

Il est cependant utile de rajouter que le contrat d'édition n'est valable qu'en droit français. Dans le cas des éditeurs étrangers, le contrat de référence est celui de l'éditeur.

²¹ André LUCA, Propriété littéraire et artistique. Editions du Juris Classeur. 1995. Fasc. 1320

5.4. Types de documents pouvant constituer un fonds documentaire

Afin de traiter les points de droit relatifs à la reproduction sous forme numérique et à la diffusion des documents du fonds documentaire, il est nécessaire de dégager une typologie des documents

Le fonds documentaire de l'IRD présente un excellent cas d'étude. Les documents numérisés ou numériques sont :

1. des documents produits par des chercheurs de l'IRD et de l'ORSTOM qui n'ont pas été publiés ;
2. des articles et ouvrages publiés par l'IRD ;
3. des ouvrages publiés par des éditeurs autres que l'IRD ;
4. des ouvrages co-édités par l'IRD et un autre éditeur ;
5. des articles publiés dans des revues scientifiques.

Cas 1 : documents produits par des chercheurs de l'IRD et l'ORSTOM et non publiés

Considérant qu'il n'y a pas d'enjeu financier, que la mise à disposition du document sur un serveur et son accessibilité via le Web contribue à la valorisation de l'œuvre et à la notoriété de son auteur, l'IRD exploite le document et permet aux auteurs d'exercer leur droit de retrait.

Lorsque le chercheur est fonctionnaire, le droit qui régit la propriété intellectuelle de ces documents est celui traité dans le chapitre 4.2.

Lorsque le document est une œuvre de collaboration²² : cette "œuvre de collaboration est la propriété commune des co-auteurs".

- Si les chercheurs sont affiliés à l'IRD, l'IRD exploite le document pour les raisons précédemment évoquées.
- Si certains chercheurs ne sont pas affiliés à l'IRD, l'exploitation du document ne devrait se faire qu'après le consentement de chacun d'entre eux. Mais considérant qu'il n'y a pas d'enjeu financier, que la mise à disposition du document sur un serveur et son accessibilité via le Web contribue à la valorisation de l'œuvre et à la notoriété de ces chercheurs, l'IRD exploite le document et permet aux auteurs d'exercer leur droit de retrait.

²² Article L.113-2 alinéa 1 du code de la propriété intellectuelle

Cas 2 : des articles et ouvrages publiés par l'IRD

L'IRD détenant le droit d'exploitation des documents, il n'y a pas de barrière juridique à la mise à disposition via le site institutionnel de ces documents.

Cas 3 : des ouvrages publiés par des éditeurs autres que l'IRD

Ces ouvrages ont été écrits par au moins un chercheur de l'IRD.

Nous rappelons qu'il ne s'agit pas de reproduire et exploiter des ouvrages n'ayant pas de liens avec l'IRD. Le fonds documentaire de l'IRD n'a pas la prétention d'être une archive thématique mais la mémoire de la production de l'institut.

L'éditeur est titulaire des droits de reproduction et d'exploitation des ouvrages. Une négociation sur la cession non exclusive de ces droits au profit de l'IRD est donc à envisager.

L'éditeur :

L'éditeur n'a, à priori, aucun intérêt à accepter cette cession car les ouvrages numérisés seront en accès libre via le site institutionnel de l'institut.

L'IRD :

L'IRD est un institut de recherche public. Ses missions visent au développement des pays en voie de développement. A ce titre, la diffusion de l'information scientifique et technique dans les pays en développement favorise le progrès scientifique dans des régions qui n'ont pas accès à la littérature scientifique.

Négociation :

Si les missions de l'IRD doivent être mises en avant, la négociation ne peut pas aboutir sans respecter les investissements financiers des éditeurs.

Ainsi, l'éditeur peut demander le retrait des ouvrages pour lesquels celui-ci a relevé une diminution de son chiffre d'affaires suite à leur mise en ligne sur le site Web de l'IRD. Cette modification devra être formalisée par voie d'avenant au contrat.

Un exemple de contrat est joint en [ANNEXE 1]. Ce contrat entre l'IRD et l'éditeur d'ouvrages scientifiques Karthala concerne la reproduction et la représentation sous forme numérique des ouvrages dont un au moins des auteurs est affilié à l'IRD (ou anciennement ORSTOM).

Cas 4 : des ouvrages co-édités par l'IRD et un autre éditeur

Dans ce cas, il s'agira de déterminer en rapport avec le contrat d'édition quel est l'éditeur titulaire des droits de reproduction et d'exploitation des ouvrages.

Il se peut que les éditeurs soient co-titulaires de ces droits et il convient alors de négocier avec chacun d'eux la cession non exclusive des droits de reproduction et d'exploitation des ouvrages. Cependant dans la majorité des cas, l'éditeur principal est seul titulaire des droits d'exploitation.

Cas 5 : des articles publiés dans des revues scientifiques

Lors de l'acceptation de son manuscrit, le chercheur signe un contrat avec l'éditeur.

- Si l'éditeur est français, c'est le droit français qui s'applique et nous parlerons de droit d'auteur.
- Si l'éditeur est étranger, c'est le droit du pays de résidence de l'éditeur qui s'applique et nous parlerons de copyright (pour la plupart des pays de résidence des éditeurs scientifiques).

La diversité des éditeurs est souvent trop grande pour envisager d'obtenir l'accord de mise en ligne de l'ensemble des articles numériques ou numérisés.

Dans un premier temps, il s'agit de catégoriser les articles suivant les éditeurs. Ces données ne sont pas évidentes à obtenir car :

- le système de gestion n'offre pas toujours d'outils de bibliothéconomie,
- la mention de l'éditeur n'apparaît pas toujours dans les notices bibliographiques des articles du fonds,
- l'éditeur a changé ou les droits d'exploitation ont été cédés.

Dans un deuxième temps, il s'agit de s'intéresser aux politiques éditoriales en matière d'archivage.

Le "Joint Information Systems Committee(JISC)²³" a financé, en Août 2002, un projet visant à étudier les licences éditoriales des éditeurs représentatifs à des fins d'archivage des articles publiés.

Ainsi, le projet Romeo²⁴ recense les éditeurs et les classes par couleurs, en fonction de leur politique éditoriale en matière d'archivage.

Même si les informations fournies par Roméo ne concernent que des éditeurs étrangers et de référence, cette classification est néanmoins très appréciable.

Couleur ROMEO	Politique d'archivage
blanc	Archiver n'est pas autorisé
jaune	Archiver seulement la prépublication
bleu	Archiver seulement l'article publié
vert	Archiver la pré et post publication

Tableau 2 : classement par couleur des politiques d'archivage des éditeurs scientifiques²⁵

Une étude des données de la base Romeo se trouve en [ANNEXE 3].

Si le fonds documentaire est constitué uniquement d'articles publiés, nous ne nous intéresserons pas aux éditeurs catégorisés en jaune.

Si le fonds documentaire est constitué de prépublications, nous devons prendre en compte les éditeurs catégorisés en jaune.

Pour les éditeurs, qui n'autorisent pas le niveau d'archivage souhaité de façon contractuelle avec les auteurs, il est nécessaire d'envisager une négociation aboutissant à l'établissement d'un contrat spécifique.

En [ANNEXE 2], figure un exemple de contrat concernant les articles publiés par le MNHN et dont au moins un des auteurs est un chercheur de l'IRD.

L'ampleur de la négociation (nombre d'éditeurs et durée) est à évaluer en fonction des contraintes de temps et des moyens humains dont dispose l'institution.

²³ Site du JISC : <http://www.jisc.ac.uk/>

²⁴ Projet Romeo : <http://www.sherpa.ac.uk/romeo.php>

²⁵ Ibid.

6. CONCLUSION

Les systèmes d'information prennent une place de plus en plus importante dans les organisations.

En documentation, on remplace les archives numériques par des bases de données relationnelles depuis plusieurs années.

Au delà des termes utilisés, se trouvent l'appropriation des nouvelles technologies de l'information par les professionnels de la documentation.

Cette étude a voulu montrer que le système d'information documentaire est un moyen de modéliser les processus d'un organisme et de placer le document au centre des objectifs documentaires en rationalisant les tâches et procédures documentaires qui lui sont associées. Les acteurs (documentalistes et utilisateurs) jouent alors un rôle identifié dans l'exploitation du document numérique.

Le déploiement d'un système d'information nécessite la mise en œuvre de moyens humains et de compétences pluridisciplinaires : documentation, ingénierie documentaire, informatique et droit. Les conséquences organisationnelles sur les institutions sont importantes en termes de communication, formation, redéfinition des tâches et formalisation de procédures.

Ce mémoire aborde des aspects organisationnels et techniques, notamment sur les systèmes de recherche d'information, ainsi que des aspects juridiques liés à l'exploitation du document numérique.

Le stage à l'IRD m'a permis de découvrir les systèmes d'information et ses enjeux sur les institutions. Ils resteront, dorénavant, au centre de mon activité professionnelle.

BIBLIOGRAPHIE

La bibliographie apparaît par ordre de citation dans le texte.

[PRI 00] Prié Y. ; (2000) Sur la piste de l'indexation conceptuelle de documents : une approche par l'annotation. *Document numérique* : 4(1-2), pp. 11-35

[GAL 03] Gabriel Gallezot, Christian Rossi, Ghislaine Chartron et Jean-Max Noyer. Conception d'une Archive ouverte en SIC : le sens de la technique. Hypertextes hypermedias : créer du sens à l'ère du numérique, H2PTM'03. 24 septembre 2003.

[BUS 03] Buscal, C., Conduire un projet de GED. *Archimag*, 2003(162): p. 40-42.

[MAR 02] Hugues Marchat. Kit de conduite de projet. 2002, 4^{ème} édition, Editions d'organisation.

[SAL 83] Salton G., McGill J.M., (1983). Information retrieval (IR) is concerned with representation, storage, organization and accessing of information items. *In* : Introduction to modern information retrieval. New York : McGraw-Hill Book Company.

[BAL 96] Balpe J.P. et al., Techniques avancées pour l'hypertexte, Hermès, Paris, 1996.

[SAL 84] Salton G., McGill J.M., (1984). Introduction to modern information retrieval. New York : McGraw-Hill Book Company.

[ROU 99] Roussey C., Calabretto S., Pinon J.M. ; (1999) Etat de l'art en indexation et recherche d'information. *Document numérique* : 3(3-4), p. 121-149

[LOA 94] Le Loarer P., Indexation automatique, recherche d'information et évaluation », *In* Le Traitement électronique du document Institut national de recherche en informatique et en automatique. Cours INRIA, 3-7 octobre 1994, Aix-en-Provence. Paris : ADBS Éditions, 1994. – (Sciences de l'information : études et techniques). – 287 p. ; 24 cm. ISBN 2-901046-76-2

[SAL 75] Salton G., Wong A., Yang C. S.: A Vector Space Model for Automatic Indexing. *Communication of the ACM*. 1975. 18(11): p. 613-620

[FUR 87] Furnas G.W., Landauer J.K., Gomez L.M., Dumais S.T., The vocabulary problem in human system communication. *Communication of the ACM*, novembre 1987. 30(11) : p.964-971

[SAL 88] Salton G. A simple blue print for automatic Boolean query processing. *Information Process Management*. 1988. 24(3): p. 269-280.

[SAL 83] Salton G. et al. Extended Boolean information retrieval. *Communication of the ACM*. Novembre 1983. 26(11) : p1022-1036

[JACQUEMIN] Christian Jacquemin (Université Paris 11 et LIMSI-CNRS). Visualisation et recherche d'information. Cours du DEA I3, Université Paris 11. Disponible sur : <http://www.limsi.fr/Individu/jacquemi/>

ANNEXES

ANNEXE 1 : Contrat de reproduction et d'exploitation des ouvrages publiés par Karthala dont au moins un des auteurs est un chercheur de l'IRD

ANNEXE 2 : Contrat de reproduction et d'exploitation des articles publiés par le MNHN dont au moins un des auteurs est un chercheur de l'IRD

ANNEXE 3 : Les politiques éditoriales en matière d'archivage

ANNEXE 1

CONTRAT

Entre :

L'Institut de Recherche pour le Développement, établissement public de recherche scientifique et technologique, ci-après désigné IRD, dont le siège est à Paris, 213 rue la Fayette 75480 Paris 10, représenté par son Directeur général, Monsieur Serge Calabre, lequel a délégué sa signature à Madame Marie-Noëlle Favier.

D'une part,

ET

Les Editions Karthala, SARL au capital de 293065,20 euros, ci-après dénommée Karthala, immatriculée au RCS de Paris sous le numéro.319012233 dont l'adresse est au 22-24 boulevard Arago, 75013 PARIS, représentée par son gérant Monsieur Robert Ageneau.

d'autre part,

L'IRD et Karthala étant ci-après désignés ensemble "les Parties"

IL EST PREALABLEMENT EXPOSE QUE

Une des missions de l'IRD est d'assurer l'information scientifique et technique notamment dans les pays en développement.

L'IRD gère sur son site de Bondy, un fonds documentaire constitué par l'ensemble de la production scientifique des personnels de l'IRD (ex-ORSTOM).

Afin de diffuser cette information scientifique au plus grand nombre, l'IRD numérise une partie de ces fonds.

DEFINITIONS DES TERMES

Reproduction sous forme numérique : le droit de reproduction s'exerce à l'occasion :

- du transfert du document à partir d'un support donné (par exemple papier) vers un autre support (par exemple numérique),
- du stockage nécessaire à la visualisation des données,
- du déchargement du document consulté (par exemple sur papier, sur disquette).

Représentation sous forme numérique : communication au public sur des postes de lecture (par exemple ordinateurs)

Ouvrages : ouvrages dont la liste détaillée est jointe en annexe

CECI ETANT EXPOSE, IL A ETE CONVENU CE QUI SUIT

ARTICLE 1 : OBJET

Par le présent contrat, Karthala autorise l'IRD à numériser et à diffuser sur Internet (via le site institutionnel de l'IRD) les Ouvrages.

ARTICLE 2 : DROITS ET OBLIGATIONS DES PARTIES

2-1 - DROITS ET OBLIGATIONS DE KARTHALA

KARTHALA cède à titre non exclusif et gratuit à l'IRD le droit de reproduire et de représenter sous forme numérique les Ouvrages.

KARTHALA s'engage détenir les droits d'exploitation sur les Ouvrages.

La présente cession est consentie à l'IRD pour la reproduction et la représentation sous forme numérique des Ouvrages sur Internet, dans la langue employée pour la publication initiale et en tous pays, pour la durée du droit d'exploitation sur les Ouvrages.

KARTHALA cède, en outre, à l'IRD le droit d'adapter les Ouvrages à fin de répondre aux impératifs techniques liés à sa consultation, sans toutefois que cette adaptation ne puisse porter atteinte à la forme et au contenu des Ouvrages dans leur forme initiale.

Tout autre type d'adaptation doit faire l'objet d'un accord préalable et écrit entre les Parties.

2-2 - DROITS ET OBLIGATIONS DE L'IRD

L'IRD s'engage à utiliser tout procédé technique et informatique garantissant la conformité de la version électronique des Ouvrages à celle publiée sur support papier.

L'IRD s'engage à respecter un délai de deux ans entre la première parution des Ouvrages et leur mise en ligne.

ARTICLE 3 : DUREE

Le présent contrat est conclu pour une durée de trois ans à compter de la signature du contrat par la dernière des Parties.

Le présent contrat est renouvelable par tacite reconduction, par période de trois ans

ARTICLE 4 : MODIFICATION, DENONCIATION DU CONTRAT

Le présent contrat peut être modifié par les Parties, par voie d'avenant.

Au cas où l'une des Parties souhaite dénoncer le présent contrat, elle doit le faire par lettre recommandée avec avis de réception en respectant un préavis de trois mois.

ARTICLE 5 : CONTESTATIONS ET LITIGES

Les Parties conviennent, de rechercher à l'amiable le règlement du différend qui pourrait survenir tant dans l'interprétation que dans l'exécution du contrat.

En cas d'échec de la conciliation, le litige est porté devant le tribunal compétent.

Fait à Paris

le 28 avril 2004

en deux exemplaires originaux constitués du présent contrat et de son annexe.

Pour KARTHALA

Pour l'IRD

ANNEXE 2

CONTRAT

ENTRE :

L'Institut de Recherche pour le Développement, établissement public à caractère scientifique et technologique, ci-après désigné l'IRD, dont le siège est à Paris, 213 rue la Fayette 75480 Paris 10, représenté par son Directeur général, Monsieur Serge Calabre, lequel a délégué sa signature à Madame Marie-Noëlle Favier.

d'une part,

ET

Le Muséum National d'Histoire Naturelle, établissement public à caractère scientifique, culturel et professionnel, ci-après dénommé le MNHN, dont le siège est à Paris, 57 rue Cuvier 75231 Paris cedex 5, représenté par son Directeur général, Monsieur Bertrand-Pierre Galey.

d'autre part,

IL EST PRÉALABLEMENT EXPOSÉ QUE

Une des missions de l'IRD est d'assurer l'information scientifique et technique notamment dans les pays en développement.

L'IRD gère sur son site de Bondy, un fonds documentaire constitué par l'ensemble de la production scientifique des personnels de l'IRD (ex-ORSTOM).

Afin de diffuser cette information scientifique au plus grand nombre, l'IRD numérise une partie de ses fonds.

Le MNHN a publié dans ses périodiques un certain nombre d'articles écrits par des enseignants-chercheurs de l'IRD. Le MNHN est cessionnaire du droit d'exploiter les droits d'auteur sur un certain nombre des articles écrits par des chercheurs de l'IRD et publiés dans ses périodiques. Il a, en outre, le droit de céder à des tiers les droits qui lui ont été cédés par les auteurs.

DEFINITIONS DES TERMES

Reproduction sous forme numérique : le droit de reproduction s'exerce à l'occasion :

- du transfert du document à partir d'un support donné (par exemple papier) vers un autre support (par exemple numérique),
- du stockage nécessaire à la visualisation des données,
- du déchargement du document consulté (par exemple sur papier, sur disquette).

Représentation sous forme numérique : communication au public sur des postes de lecture (par exemple ordinateurs)

Articles : articles écrits ou co-écrits par un chercheur de l'IRD et publiés par le Muséum.

CECI ETANT EXPOSE, IL A ETE CONVENU CE QUI SUIT

ARTICLE 1 : OBJET

Par le présent contrat, le MNHN autorise l'IRD à numériser et à diffuser sur Internet (via le site institutionnel de l'IRD) les articles visés au préambule.

ARTICLE 2 : DROITS ET OBLIGATIONS DES PARTIES

2-1 - DROITS ET OBLIGATIONS DU MNHN

2-1-1 - Pour les articles disponibles sous forme numérique

Le MNHN cède à titre non exclusif et gratuit à l'IRD le droit de représenter les articles sur son site internet, dans la langue employée pour la publication initiale et en tous pays, pour la durée du droit d'auteur sur les articles.

2-1-2 - Pour les articles disponibles sous forme papier

Le MNHN cède à titre non exclusif et gratuit à l'IRD le droit de reproduire et représenter sous forme numérique des articles sur son site internet, dans la langue employée pour la publication initiale et en tous pays, pour la durée du droit d'exploitation sur les articles.

Le MNHN cède, en outre, à l'IRD le droit d'adapter les articles à fin de répondre aux impératifs techniques liés à leur consultation, sans toutefois que cette adaptation puisse porter atteinte à la forme et au contenu des articles dans leur forme initiale.

Tout autre type d'adaptation doit faire l'objet d'un accord préalable et écrit entre les Parties.

2-2 - DROITS ET OBLIGATIONS DE L'IRD

L'IRD s'engage à utiliser tout procédé technique et informatique garantissant la conformité de la version électronique des articles à celle publiée sur support papier.

L'IRD s'engage à transmettre au MNHN un exemplaire de tous les articles numérisés.

De plus, l'IRD s'engage à faire un lien hypertexte de la notice des articles vers le site de la revue qui a publié lesdits articles.

ARTICLE 3 : DUREE

Le présent contrat est conclu pour une durée de trois ans à compter de la signature du contrat par la dernière des parties.

Le présent contrat est renouvelable par tacite reconduction, par période de trois ans

ARTICLE 4 : MODIFICATION, DENONCIATION DU CONTRAT

Le présent contrat peut être modifié par les parties, par voie d'avenant.

Au cas où l'une des parties souhaite dénoncer le présent contrat, elle doit le faire par lettre recommandée avec avis de réception en respectant un préavis de trois mois.

ARTICLE 5 : CONTESTATIONS ET LITIGES

Les parties conviennent, de rechercher à l'amiable le règlement du différend qui pourrait survenir tant dans l'interprétation que dans l'exécution du contrat.

En cas d'échec de la conciliation, le litige est porté devant le tribunal compétent.

Fait à Paris,
Le 01 juillet 2004,
en deux exemplaires originaux.

Pour le MNHN

Pour l'IRD

ANNEXE 3

Les politiques éditoriales en matière d'archivage

Les différentes politiques éditoriales se déclinent en 4 niveaux d'autorisation :

1. L'éditeur retient tous les droits sur l'exploitation et la reproduction du manuscrit aussi bien que le document publié.
2. L'éditeur laisse les droits à l'auteur d'un ouvrage ou d'un article de déposer son manuscrit sur un site institutionnel sans restriction d'accès.
3. L'éditeur laisse les droits à l'auteur d'un ouvrage ou d'un article de déposer le document publié sur un site institutionnel sans restriction d'accès.
4. L'éditeur laisse les droits à l'auteur d'un ouvrage ou d'un article de déposer le manuscrit et le document publié sur un site institutionnel sans restriction d'accès.

A noter qu'il existe des variantes et restrictions notamment au niveau de la frontière entre le manuscrit et le document publié.

Parfois l'autorisation ne porte pas sur le document publié mais sur le manuscrit accompagné des corrections du comité éditorial. Dans ce cas, l'éditeur recevra la couleur "vert pâle". Un exemple signifiant est celui d'Elsevier qui en "vert pâle", autorise l'archivage du manuscrit ainsi que le document corrigé.

Ces modèles peuvent se modéliser en 4 couleurs. Cette modélisation est reprise dans la littérature avec des couleurs différentes (blanc, vert pâle, vert foncé et vert OU blanc, jaune, bleu et vert).

Pour des raisons de facilité de représentation, nous utiliserons par la suite la deuxième modélisation.

Répartition des modèles sur un échantillon de 95 éditeurs :

Source : Projet Romeo sur Sherpa

Couleur (Romeo)	Politique d'archivage	Nb d'éditeurs	% d'éditeurs
<u>VERT</u>	Permet l'archivage du manuscrit et du document publié	41	43 %
BLEU	Permet l'archivage du document publié	13	14 %
JAUNE	Permet l'archivage du manuscrit	6	6 %
BLANC	Aucune forme d'archivage autorisée	35	37 %

63% des éditeurs autorisent quelque forme que ce soit d'archivage.

Remarque : certains éditeurs représentent 80% des revues. Il est donc intéressant de proposer des données sur des revues et non plus sur des éditeurs.

Répartition des modèles sur un échantillon de 10673 revues :

Source : Stevan Harnad

Couleur (Romeo)	Politique d'archivage	Nb de revues	% des revues
<u>VERT</u>	Permet l'archivage du manuscrit et du document publié	3855	36 %
BLEU	Permet l'archivage du document publié	1772	17 %
JAUNE	Permet l'archivage du manuscrit	3253	30 %
BLANC	Aucune forme d'archivage autorisée	1793	17 %

83% des revues publient des articles qui, sous une certaine forme, peuvent être archivés.

Les revues en libre accès :

Les revues en libre accès sont comptabilisées comme autorisant l'archivage du manuscrit et du document publié. Elles sont donc incluses dans les données "vertes".

Voici une étude intégrant l'ensemble des revues en libre accès répertoriées et présentant ces revues "en or".

Source : Stevan Harnad

Couleur (Romeo)	Politique d'archivage	Nb de revues	% des revues
OR	Revue en ligne accès	1177	10 %
VERT	Permet l'archivage du manuscrit et du document publié	3628	31 %
BLEU	Permet l'archivage du document publié	1772	15 %
JAUNE	Permet l'archivage du manuscrit	3253	28 %
BLANC	Aucune forme d'archivage autorisée	1793	16 %

56 % des revues publient des articles qui peuvent être archivés (article publié).

Sources :

Site du projet Romeo sur Sherpa : <http://www.sherpa.ac.uk/romeo.php>

La base Sherpa s'intéresse aux éditeurs

Site du projet Romeo sur Eprints : <http://romeo.eprints.org/>

La base Eprints s'intéresse aux revues

Etude de Stevan Harnad :

<http://www.ecs.soton.ac.uk/~harnad/Temp/Romeo/romeosum.html>

Données statistiques et évolution basés sur la base Eprints

