HAL
open science

Numérisation de la littérature grise : mise au point d'une chaîne documentaire

Marie-France Claerebout

To cite this version:

Marie-France Claerebout. Numérisation de la littérature grise : mise au point d'une chaîne documentaire. Sciences de l'information et de la communication. 2000. dumas-01952193

HAL Id: dumas-01952193

https://dumas.ccsd.cnrs.fr/dumas-01952193v1

Submitted on 12 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Marie-France CLAEREBOUT

DESS Stratégies de l'Information et de la Documentation

Année 1999-2000

Numérisation de la littérature grise : mise au point d'une chaîne documentaire

Rapport de stage

effectué

au Service Commun de la Documentation, Université de Lille1

sous la direction de

Monsieur Dominique COTTE, enseignant à l'IDIST

Monsieur Jean-Bernard MARINO, directeur du SCD-BU de Lille 1

Lille 3 UNIVERSITE CHARLES-DE-GAULLE

Idist Information, documentation, information, Scientifique et technique

TABLE DES MATIERES

PARTIE I - NUMERISER LA LITTERATURE GRISE : LES ENJEUX 4

1. LE CONTEXTE LOCAL : L'UNIVERSITE DE LILLE 1 4
   1.1. le Service Commun de la Documentation 4
   1.2. les autres services de l'USTL 7
2. LE CONTEXTE ECONOMIQUE ET CULTUREL 8
   2.1. la défense de la langue française 8
   2.2. l'IST numérisée chez les chercheurs 12
   2.3. la propriété des documents numériques 13

PARTIE II - LE PROJET : FONCTIONNALITES ATTENDUES 15

3. ETAT DES LIEUX 15
   3.1. quelles collections, pour quel public ? 15
4. L'INTERFACE UTILISATEUR 17
   4.1. fonctions de consultation 17
   4.2. le site web 19
   4.3. fonctions d'administration 24
5. LES DONNEES 25
   5.1. règles de gestion 25
   5.2. entités 25

PARTIE III - PREALABLES SUR LES DOCUMENTS NUMERIQUES . 26

6. LES FORMATS DE DOCUMENTS NUMERIQUES 26
   6.1. le mode image 27
   6.2. le mode texte 29
   6.3. mode structuré : le balisage de textes 30
7. LE REFERENCEMENT DES DOCUMENTS NUMERIQUES 34
   7.1. web et métadonnées 34
   7.2. bibliothèques et notices enrichies 35

PARTIE IV - LA CHAINE DE TRAITEMENT : SOLUTIONS TECHNIQUES ENVISAGEABLES 36

8. CONTRAINTES DE MISE EN OEUVRE 36
   8.1. l'environnement technique 36
   8.2. contraintes éditoriales 38
9. LA COLLECTE DES DOCUMENTS 41
   9.1. format d'acquisition des documents 41
   9.2. choisir un mode de numérisation 42
10. STOCKAGE 43
   10.1. mode image 43
   10.2. mode image+texte 43
   10.3. mode texte structuré 44
11. REFERENCEMENT 44
   11.1. les informations secondaires 44
   11.2. des codifications normalisées 45
12. LIER INFORMATIONS SECONDAIRES ET DOCUMENT 47
   12.1. notice MARC pointant sur le document numérique 47
   12.2. notice SGML pointant sur le document numérique 48
   12.3. notice bibliographique intégrée dans le document numérique 49

12.4. documents format image encapsulés dans le document numérique *50*
13. LA RESTITUTION SUR LE WEB 52
13.1. rappel des objectifs *52*
13.2. les différents formats de restitution *52*
13.3. du serveur au client : les protocoles *53*

PARTIE V - PRECONISATIONS, PERSPECTIVES 54

14. RECOMMANDATIONS 54
14.1. promotion du site *54*
14.2. financement du projet *55*
15. OPTIONS POUR LA CHAINE DOCUMENTAIRE 56
15.1. stockage et restitution des documents primaires *56*
15.2. stockage des informations secondaires *56*
15.3. première solution envisagée : TEI *57*
15.4. deuxième solution envisagée : MARC SGML *59*
15.5. moyens à mettre en oeuvre *59*
16. LE PROJET : EVOLUTIONS ENVISAGEABLES 61
16.1. le travail en partenariat *61*
16.2. à Lille1, des documents primaires au format XML *62*
16.3. revues universitaires électroniques *62*
17. BIBLIOTHEQUES ET EDITION SCIENTIFIQUE 63
17.1. les nouveaux acteurs de l'édition scientifique *63*
17.2. et les bibliothèques dans tout cela ? *66*

PARTIE VI - CONCLUSION 69

PARTIE VII - BIBLIOGRAPHIE 70

PARTIE VIII - ANNEXES 72

Introduction

Le Service Commun de la Documentation de L'Université de Lille 1 possède un fonds de littérature grise peu exploité. Il s'agit de documents scientifiques produits en dehors des circuits commerciaux de l'édition et de la diffusion : cours, communications à des congrès ou des séminaires, rapports ou pré-publications, positions de thèses de Lille1 (date, jury, etc).

Il a été prévu de numériser ceux de ces documents qui sont rédigés en français, afin de leur offrir une visibilité universelle sur le web.

Les offrir à la consultation, in extenso et gratuitement, aura pour effets de

- valoriser les travaux des laboratoires de recherche concernés,
- permettre que la portée de ces travaux dépasse leur discipline de rattachement,
- contribuer à la vitrine de Lille1 en promouvant ses travaux de recherche ou d'enseignement,
- contribuer aux relations internationales en valorisant les travaux d'autres pays francophones,
- défendre la présence du français sur le web et la francophonie en général,
- valoriser la technicité des bibliothèques.

Un tel service pourra de plus être proposé à tous les enseignants/chercheurs francophones, dès la conception de nouveaux textes.

C'est dans ce cadre que j'ai réalisé mon stage de DESS, avec pour mission de concevoir une chaîne documentaire pour ces documents de littérature grise, qu'ils soient déjà présents en bibliothèque ou collectés ultérieurement.

Une première étape a été de m'interroger sur le contexte dans lequel s'inscrit ce projet, sur le domaine de l'étude, les organisations impliquées, les objectifs visés ; c'est l'objet de la partie1.

La partie 2 précise les besoins exprimés et les fonctionnalités attendues, avant de détailler l'ergonomie de l'interface qui donnera accès aux documents.

La mise en œuvre de bibliothèques numériques lève de nombreuses questions méthodologiques, techniques ou juridiques qui donnent parfois lieu à des débats

passionnés. Il m'a semblé utile, en prélable à l'étude proprement dite, de faire en partie 3 la synthèse des connaissances voire des principaux points de vue actuels.

Un projet de numérisation demande une évaluation très poussée de toutes les contraintes et des obstacles qui lui sont inhérents. Je recense ces contraintes en partie 4, avant d'aborder le cœur du travail effectué, l'élaboration d'une chaîne documentaire : comme pour les ouvrages imprimés, toute la chaine doit être pensée, de l'acquisition à la mise sur les « rayonnages » en passant par le catalogage. Pour chacune de ces étapes, sont envisagées diverses solutions, éclairées par les avantages et inconvénients de chacune.

La partie 5 apporte quelques recommandations pour la mise en œuvre du projet sous les incontournables aspects de promotion, organisation .. et financement. Elle se conclut par une réflexion élargie sur l'édition scientifique numérique et le rôle que peuvent y jouer les bibliothèques.

Partie I - Numériser la littérature grise : les enjeux

1. LE CONTEXTE LOCAL : L'UNIVERSITE DE LILLE 1

L'Université des Sciences et Technologies de Lille (USTL ou 'Lille1'), était initialement une université à dominante scientifique et technique. Y ont été incluses par la suite de nouvelles disciplines : la géographie, les sciences économiques et sociales.

1.1. le Service Commun de la Documentation

Au sein de Lille1, le Service Commun de la Documentation (SCD-BU) a pour mission essentielle d'acquérir, gérer et communiquer les documents de toutes sortes qui appartiennent à l'Université ou sont à sa disposition. Le SCD-BU est placé sous l'autorité du président de l'Université, son directeur est nommé par le ministre chargé des universités.

Toute bibliothèque d'UFR ou de composante fonctionnant dans l'université est associée à ce service. Le noyau du SCD est la Bibliothèque Universitaire (BU) , pluridisciplinaire et ouverte à tous publics.

1.1.1. Le fonds de la BU[1]

- Les collections de la BU couvrent essentiellement toutes les disciplines enseignées au sein de l'université, à savoir :

➢ *Sciences*
  - Mathématiques
  - Astronomie
  - Physique
  - Chimie
  - Sciences de la Terre
  - Paléontologie - paléozoologie
  - Sciences de la vie - biologie
  - Botanique
  - Zoologie
  - Médecine.

➢ *Techniques*
  - Agriculture-Agronomie
  - Automatique
  - Électronique-Électrotechnique
  - Génie chimique-Biotechnologies-Technologies connexes
  - Ingénierie
  - Informatique
  - Matériaux
  - Physique appliquée
  - Techniques du bâtiment
  - Techniques et fabrications industrielles-Textile
  - Technologie des télécommunications.

➢ *Sciences économiques et sciences humaines*
  - Sciences sociales : sociologie, anthropologie, science politique, économie, droit des affaires, administration, problèmes et services sociaux, éducation, commerce et transports
  - Langage et langues : dictionnaires, méthodes d'apprentissage
  - Gestion de l'entreprise, mercatique
  - Arts, urbanisme, architecture, sports
  - Littérature, expression écrite

[1] d'après le site internet de la bibliothèque universitaire de Lille1 http://www.univ-lille1.fr/bustl/accueil

- Géographie, histoire, archéologie.

L'essentiel du fonds de la BU est constitué de 120 000 livres, 60 000 thèses françaises ou étrangères, 2000 titres de périodiques dont 600 abonnements en cours, ainsi que de fonds spécifiques : collection Asie-Pacifique, dépôt de la Société géologique du Nord et de la Maison des sciences de l'homme.

La BU possède par ailleurs 3 000 documents environ provenant de l'enseignement supérieur ou de la recherche et n'ayant fait l'objet d'aucune publication commerciale : cours, communications à des congrès ou des séminaires, rapports de recherche, pré-publications. Rédigés en langue française ou dans diverses langues étrangères (anglais et allemand essentiellement), ces textes sont stockés en magasin et signalés au catalogue.

Joints aux 60 000 thèses pré-citées, ils constituent le fonds de « littérature grise », qui s'enrichit régulièrement.

1.1.2. Les services offerts

La BU est fréquentée en priorité par les étudiants, chercheurs ou enseignants de l'Université de Lille1, mais est ouverte à tout public. Outre sa fonction traditionnelle de consultation sur place et de prêt, elle propose divers services comme

- La médiathèque, le centre de ressources pédagogiques
- La recherche informatisée
- La consultation à distance du catalogue via Internet
- L'abonnement aux revues électroniques
- La consultation de sujets d'examen numérisés

Notons que le SCD-BU ne propose encore que peu de revues électroniques en ligne car ce service est onéreux. Il a donc invité un groupe de travail d'enseignants-chercheurs de toutes disciplines à évaluer les besoins dans les laboratoires. L'évaluation doit déboucher sur un plan d'acquisitions avec les autres SCD-BU de France (mutualisation des abonnements dans le cadre du groupement d'achat Couperin).

La numérisation des sujets d'examen a pour sa part été mise en place en 1998. L'ensemble des sujets proposés à Lille1 au cours des 3 dernières années est maintenant consultable soit en local sur des terminaux de la BU, soit par le biais de l'intranet du campus.

1.2. les autres services de l'USTL

1.2.1. les relations internationales

L'Université des Sciences et Technologies de Lille est engagée dans la coopération avec les universitaires d'autres pays francophones, d'Afrique et d'Asie particulièrement.

Rencontré durant ce stage, le vice-président chargé des Relations Internationales se montre intéressé par le projet de numérisation de littérature grise :

« De nombreux chercheurs/enseignants africains ont fait leurs études en France. Rentrés chez eux, ils se trouvent peu à peu isolés de la communauté scientifique. »

« A Franceville (Gabon), les chercheurs publient partiellement dans des revues internationales, mais essentiellement dans des revues locales à faible notoriété. Il serait bon de leur offrir un tremplin, tout en précisant qu'une diffusion de notre part ne remet pas en cause leur propre circuit de diffusion. »

1.2.2. les relations avec les entreprises

L'Université des Sciences et Technologies de Lille s'implique également dans la vie économique de la région, par le biais de son service de Relations avec les Entreprises.

Engagées dans la veille scientifique, certaines de ces dernières ont manifesté récemmment leur souhait d'accéder par Internet aux travaux émanant de laboratoires universitaires de Lille1. Le projet de diffusion de la littérature grise numérisée pourrait partiellement répondre à cette attente.

1.2.3. le Campus Virtuel

L'USTL se tourne résolument vers le numérique, avec son projet de Campus Virtuel qui proposera aux étudiants distants l'accès en ligne à des modules de formation personnalisée.

Modélisé et conçu par le laboratoire TRIGONE[1] de Lille 1 en association avec la société ARCHIMED, le Campus Virtuel sera mis en œuvre par le Service pour l'Enseignement sur Mesure Médiatisé[1] de l'université.

---

[1] http://www-trigone.univ-lille1.fr/. Spécialisé dans la recherche en formation permanente des adultes, le laboratoire Trigone est hébergé par le CUEEP.

Ce projet relève d'une même démarche que celui du SCD : le developpement d'une « université en ligne », proposant à distance services d'enseignement et services documentaires.

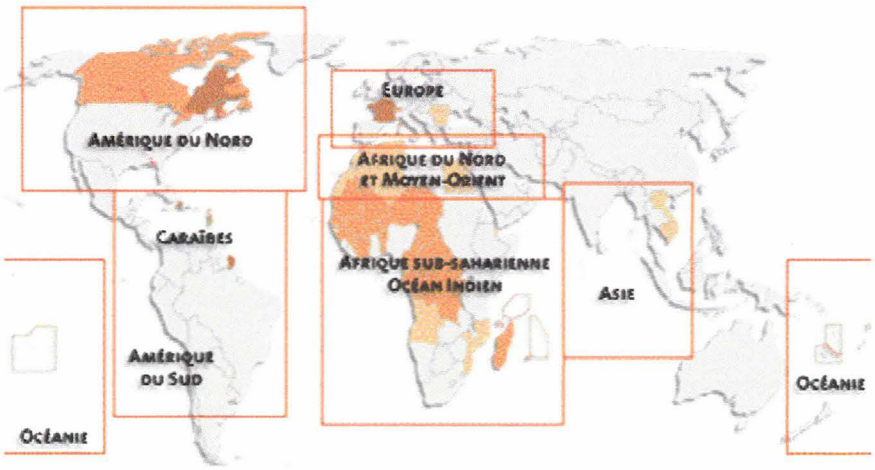
2. LE CONTEXTE ECONOMIQUE ET CULTUREL

2.1. la défense de la langue française

2.1.1. la communauté francophone

En se centrant sur la diffusion de documents rédigés en langue française, le SCD s'inscrit dans un mouvement de défense du français comme langue d'étude et de recherche.

Force est de constater que l'hégémonie de l'anglais est plus marquée que jamais avec l'explosion du web. Or plus de 100 millions de personnes, réparties sur les 5 continents, parlent le français :



Pays ou région où le français est langue officielle et maternelle
Pays ou région où le français est langue officielle ou administrative
Pays où le français est langue d'enseignement privilégiée

Situation de la Francophonie dans le monde[2]

---

[1] www.semm-lille1.fr. Créé en 1998, le SEMM assure le développement et la mise en œuvre de l'enseignement sur mesure à l'université.

[2] Carte issue du site http://www.france.diplomatie.fr/francophonie/francais/carte/carte.html

Le groupe francophone a acquis une dimension politico-diplomatique en novembre 1997, avec la nomination de Monsieur Boutros Boutros-Ghali au poste de premier Secrétaire Général de la Francophonie.

Au niveau international, diverses organisations oeuvrent pour la défense de la langue française et la coopération entre pays francophones :

- *l'Agence de la Francophonie (http://agence.francophonie.org/) :*

Opérateur principal de l'Organisation internationale de la Francophonie, l'Agence intergouvernementale de la Francophonie a été créée par la convention de Niamey (Niger) le 20 mars 1970 sous l'impulsion de trois chefs d'État africains : Léopold Sédar Senghor du Sénégal, Habib Bourguiba de Tunisie et Hamani Diori du Niger.

Elle regroupe aujourd'hui 49 États et gouvernements qui, unis par les liens que crée le partage de la langue française, souhaitent, par des actions de coopération multilatérale, utiliser ces liens au service de la paix, du dialogue des cultures et du développement. L'Agence de la Francophonie est l'unique opérateur intergouvernemental de l'Organisation internationale de la Francophonie.

- *Le réseau ANAIS (http://www.anais.org/)*

Le réseau ANAIS a été créé à la suite de la conférence de Genève d'octobre 1996. Il vise à faciliter l'appropriation des technologies de l'information et de la communication par les africains. Les membres du réseau ANAIS sont des entités juridiquement indépendantes qui, en Afrique ou en Europe, développent des projets qui leur sont spécifiques dans un esprit de partenariat.

Bamako (Mali) a accueilli en février 2000 une rencontre internationale organisée par le réseau ANAIS. Bilan de cette rencontre, qui a permis de faire le point sur les usages des TIC pour le développement local, au Sud et au Nord : http://www.bamako2000.org/rencontre/bilan.html

2.1.2. Universités et francophonie

Pour l'enseignement supérieur et la recherche, l'opérateur direct et reconnu des Sommets des chefs d'État et de Gouvernement des pays ayant le français en

partage est l'Agence Universitaire de la Francophonie(www.aupelf-uref.org)[1], créée en 1961 à Montréal

L'Agence Universitaire de la Francophonie a pour mandat officiel de contribuer à la construction francophone en consolidant un espace scientifique de langue française animé par ses principaux acteurs, les établissements, les enseignants, les chercheurs et les étudiants.

Elle rassemble aujourd'hui 375 établissements d'enseignement supérieur, grandes écoles, conférences internationales de doyens et chefs d'établissements universitaires, ainsi que 334 départements d'études françaises et francophones.

Parmi ses principaux programmes , citons :

- l'UREF, projet de réseau des universités ayant le français en partage,
- l'Université Virtuelle Francophone (http://www.aupelf-uref.org/uvf) qui, en s'appuyant sur les Nouvelles Technologies de l'Information et de la Communication, développe le travail en réseau et la mise en commun des ressources universitaires en français, dans une optique de solidarité et de co-développement,

➢ *La science en français*

L'Agence universitaire de la francophonie mène un plan d'action visant à doter la Francophonie de capacités propres de référencement et de promotion de sa littérature scientifique. Elle précise sa démarche (http://www.aupelf-uref.org/cadres/fr_prog.htm) :

« Le maintien de multilinguisme dans les sciences est un point qui préoccupe la Francophonie. Le français s'appauvrira, sera marginalisé s'il ne reste pas au coeur de la production scientifique, de l'innovation.

La Francophonie sera vivante si elle dispose d'une recherche et d'une science de qualité; si les résultats de la recherche sont aussi publiés en français et

---

[1] AUPELF = Association des universités partiellement ou entièrement de langue française

UREF = Université des réseaux d'expression française

valorisés; si le français se maintient comme une grande langue de la science et de la technologie contemporaines.»

Trois grands programmes ont été identifiés :

- l'Institut francophone de référencement de l'information scientifique et technique (IRIS),
- l'édition de revues de synthèse faisant état en français des grandes avançées de la science.
- le soutien aux publications primaires éditées en français,

2.1.3. Francophonie et Internet

Internet est un puissant atout pour la communauté francophone, c'est le constat qui a été exprimé lors du colloque « INITIATIVES99 »,organisé par l'AUF à Edmundston[1] :

« La Francophonie est encore très minoritaire sur les Inforoutes. Sa diversité géographique et culturelle, sa volonté de créer un maillage linguistique et culturel sont des atouts importants dans la réussite de ses ambitions humanistes. Les inforoutes permettent aujourd'hui à la Francophonie de s'exprimer au sein d'un continent virtuel. L'enjeu est de taille, les défis à relever. »

C'est pourquoi l'Agence universitaire de la Francophonie met en oeuvre des actions visant à multiplier, au Sud comme au Nord, les contenus en français sur Internet. Elle s'appuie sur ses sites SYFED-REFER[2] qui, répartis dans tout le monde francophone, ont pour buts de rompre l'isolement des enseignants et des chercheurs et de renforcer le potentiel pédagogique dans le milieu universitaire francophone[3].

---

1 plus de détails sont disponibles sur http://www.aupelf-uref.org/initiatives/colloque/colloque.htm

2 SYFED est le SYstème Francophone d'Edition et de Diffusion. Il dispose d'un infoport qui permet la connexion à REFER (Réseau Electronique Francophone pour l'Enseignement et la Recherche), et donc à Internet.

3 Voir aussi http://www.rond-point.org, portail vers des sites de cours universitaires francophones, créés partout dans le monde.

C'est dans un même esprit qu'à été créé, en 1998, le Fonds francophone des inforoutes issu du plan d'action élaboré par le VIIe Sommet de la Francophonie (http://www.francophonie.org/fonds/fonds.htm).

Les crédits de ce Fonds proviennent de différents Etats et Gouvernements ayant le français en partage (Cameroun, Canada, Canada-Québec, Canada-Nouveau Brunswick, Communauté Française de Belgique, Côte d'Ivoire, France, Gabon, Liban, Monaco, Suisse, Sénégal).

Sa mission porte sur les points suivants :

- Démocratiser l'accès aux inforoutes ;
- Développer l'aire d'éducation, de formation et de recherche ;
- Renforcer l'aire de création et de circulation des contenus ;
- Promouvoir une aire de développement économique et social ;
- Etablir une vigie francophone ;
- Sensibiliser les jeunes, les producteurs et les investisseurs ;
- Assurer une présence concertée des Francophones dans les instances internationales chargées du développement des inforoutes.

2.1.4. position de la France

Les universités d'été de la communication, qui se tiennent annuellement à Hourtin, sont l'occasion pour le premier ministre de préciser les actions gouvernementales favorisant le positionnement de la France dans la société de l'information

Lors du dernier congrès de la Fédération internationale des professeurs de français[1], le chef du gouvernement s'est plus précisément exprimé en faveur du français comme arme de contre-pouvoir, de résistance à l'uniformité du monde. Il a, à cette occasion, promis son soutien à toute initiative qui contribuerait « à nourrir la Toile en français »

Ces prises de position sont relayées par le Ministère de la culture et de la francophonie.

2.2. l'IST numérisée chez les chercheurs

Dans le cadre du projet Couperin, les conservateurs de la BU ont mené en mai-juin 2000 une enquête sur les usages des chercheurs de Lille1 : sources

[1] Voir en page 3 du Monde daté du 23 juillet 2000

documentaires utilisées, usage du numérique. A ce jour, nous ne disposons pas du résultat du dépouillement.

2.3. la propriété des documents numériques

Pour s'imposer définitivement, l'édition électronique sur Internet devra apporter des réponses à de nombreuses questions d'ordre économique, juridique, intellectuel et technique. La libre circulation des documents et leur facilité de reproduction pose le problème des droits d'auteur.

2.3.1. Les concepts

En droit français, la propriété intellectuelle comprend deux domaines principaux :

- la propriété industrielle, qui porte principalement sur les inventions, les marques, les dessins et modèles industriels et les appellations d'origine;
- le droit d'auteur, qui porte principalement sur les œuvres littéraires, musicales, artistiques, photographiques et audiovisuelles.

Voici, extraits du code de la propriété intellectuelle du droit français, quelques éléments sur le droit d'auteur, utiles au projet.[1]

➢ *Nature du droit d'auteur (extrait de l'article L. 111-1)*

L'auteur d'une oeuvre de l'esprit jouit sur cette oeuvre, du seul fait de sa création, d'un droit de propriété incorporelle exclusif et opposable à tous. Ce droit comporte des attributs d'ordre intellectuel et moral ainsi que des attributs d'ordre patrimonial.

➢ *Droits moraux (extrait de l'article L. 121-1)*

L'auteur jouit du droit au respect de son nom, de sa qualité et de son oeuvre. Ce droit est attaché à sa personne. Il est perpétuel, inaliénable et imprescriptible.

➢ *Droits patrimoniaux (extrait de l'article L. 122-1)*

Le droit d'exploitation appartenant à l'auteur comprend le droit de représentation et le droit de reproduction.

➢ *Cession des droits (extraits des article L. 131-3 et L. 121-4)*

La transmission des droits de l'auteur est subordonnée à la condition que chacun des droits cédés fasse l'objet d'une mention distincte dans l'acte de cession et que le domaine d'exploitation des droits cédés soit délimité quant à son étendue et à sa destination, quant au lieu et quant à la durée.

---

[1] tirés du site http://www.celog.fr/cpi/sommaires/livre_1.htm , qui propose une version commentée du code de la propriété intellectuelle

Nonobstant la cession de son droit d'exploitation, l'auteur, même postérieurement à la publication de son oeuvre, jouit d'un droit de repentir ou de retrait vis-à-vis du cessionnaire.

➢ *Contrat d'édition (extraits des article L. 132-1 et L. 132-8)*

Le contrat d'édition est le contrat par lequel l'auteur d'une oeuvre 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'oeuvre, à charge pour elle d'en assurer la publication et la diffusion.

L'auteur doit garantir à l'éditeur l'exercice paisible et, sauf convention contraire, exclusif du droit.

2.3.2. Propriété intellectuelle et mondialisation

Le respect des droits à travers le monde ne peut passer que par une coopération entre pays. C'est à cette fin qu'a été créée l'OMPI, Organisation Mondiale de la Propriété Intellectuelle (http://www.ompi.org/fre/fdgtext.htm).

L'OMPI est une organisation intergouvernementale dont le siège est à Genève, en Suisse. C'est l'une des 16 institutions spécialisées du système des Nations Unies.

L'OMPI est chargée de promouvoir la protection de la propriété intellectuelle à travers le monde par la coopération des États et d'assurer l'administration de divers traités multilatéraux touchant aux aspects juridiques et administratifs de la propriété intellectuelle. Le nombre des États membres de l'OMPI était de 161 en février 2000.

2.3.3. Droit d'auteur et documents numériques

En juin 2000, le Conseil Européen a trouvé un accord concernant la directive sur l'harmonisation du droit d'auteur dans la société de l'information.[1]

La directive modifie et complète le cadre d'orientation européen sur le droit d'auteur et les droits voisins pour répondre aux nouveaux défis de la technologie et de la société de l'information, au bénéfice des titulaires du droit et des utilisateurs. Elle couvre en particulier les droits de reproduction, de communication au public, de diffusion, la protection juridique des dispositifs anti-copie et les systèmes de gestion de droits.

---

[1] http://europa.eu.int/comm/internal_market/fr/intprop/intprop/news/601.htm

Il demeure que le respect de ces principes reste bien difficile à contrôler et ne relève encore que d'une certaine déontologie des internautes ...

Partie II - Le projet : fonctionnalités attendues

La partie II décrit l'interface utilisateur, les traitements, les données concernées, tels que perçus à cette étape du projet .

3. ETAT DES LIEUX

3.1. quelles collections, pour quel public ?

3.1.1. les collections concernées

Ne sont pris en compte, dans le cadre du projet de numérisation, que les documents de littérature grise rédigés en langue française et, dans un premier temps, de taille 'raisonnable' : nous convenons un peu arbitrairement de nous limiter aux textes assimilables à des brochures (soit de moins de 50 pages).

Nous baserons l'étude sur une cinquantaine de documents répondant à ces critères et dont les auteurs autorisent la diffusion. Disponibles sous forme imprimée seulement, ces textes proviennent d'écoles ou d'universités françaises, belges, suisses, tunisiennes, québecoises. Ils couvrent l'essentiel des types de documents de littérature grise envisageables, puisqu'ils se répartissent en :

- prépublications,
- rapports de recherche,
- communications à des colloques,
- cours de DEA

Pour les thèses soutenues à Lille1, la solution retenue est de proposer en ligne les positions de thèses enrichies des tables des matières. Précisons qu'une « position de thèse » consiste en un feuillet A4 où sont indiqués titre, doctorant, date, jury, résumé, ...

3.1.2. Pour quel public ?

Ce service sera ouvert à tous publics de tous pays. Toutefois il est plutôt destiné, par nature, au public universitaire francophone (étudiants de l'enseignement supérieur, enseignants, chercheurs)

Sa finalité n'est pas la communication entre chercheurs, qui impliquerait de fournir des documents « outils de travail » manipulables à l'écran mais n'entre pas dans les missions de la bibliothèque.

3.1.3. la démarche engagée

➢ *Numériser le fonds existant*

Une première étape est de procéder à la rétroconversion des documents que possède la BU : partir des documents imprimés pour en constituer l'image numérique. Ceci ne peut se faire qu'avec l'accord des auteurs ; une bibliothécaire se charge donc depuis quelques mois de les contacter afin d'obtenir leur autorisation de numériser et diffuser leurs productions.

➢ *Enrichir la collection*

La BU a, par ailleurs, entrepris d'informer de son projet un grand nombre d'enseignants/chercheurs de différents pays francophones et de toutes disciplines. En retour, le fonds s'enrichit peu à peu de documents propres à être diffusés, reçus le plus souvent sur papier mais parfois au format électronique. A ce jour, il s'agit essentiellement de supports de cours de 3ème cycle ou de notes de recherche.

Au sein de Lille1 :

- les directeurs de recherche ont été contactés par messagerie électronique. Le manque d'articles rédigés en français explique peut-être le faible taux de réponses,

- les doctorants viennent personnellement à la BU déposer leur projet de thèse, ce qui permet un contact direct plus efficace que le courrier,
- les polycopiés de cours et TD seront inventoriés dans le cadre d'un projet mené conjointement par le SEMM et la BU.

A l'extérieur :

- des enseignants/chercheurs ont été contactés en France, au Québec, en Belgique et en Suisse. Les pays qui ont le français en partage se montrent les plus sensibilisés,
- le contact avec les pays africains et asiatiques se fera prochainement par l'intermédiaire du service des Relations Internationales de Lille1.

4. L'INTERFACE UTILISATEUR

La consultation de la base se fera via Internet, nous ne prévoyons pas dans l'immédiat d'accès local.

Il faut aussi prévoir les diverses fonctions de gestion de la base.

4.1. fonctions de consultation

Nous considérons que, pour accéder à la base, les utilisateurs disposent d'un navigateur web standard (type Explorer ou Netscape) ou d'un accès Telnet. Il conviendra de s'assurer que cette dernière solution s'avère vraiment utile pour les pays du Sud .

4.1.1. Consulter la base des notices

4.1.1.1. <u>un accès style 'portail'</u>

Il faudra que l'utilisateur puisse accéder directement à la liste de tous les documents d'un type sélectionné (rapports de recherche, cours, ...), classés par discipline.

4.1.1.2. <u>un accès sur critères de recherche</u>

Prévoir divers menus : mode de recherche simple ou expert

- Une **recherche simple** de type minitel. Lors d'une recherche de ce type, l'utilisateur se voit proposer un masque de saisie présentant les champs d'indexation existants.Il saisit alors dans certains champs les valeurs qui spécifient sa requête ou, le cas échéant, sélectionne une valeur dans une liste d'autorité. La requête générée lie les différentes conditions saisies par un seul et même opérateur logique (par défaut : et).
  Champs d'indexation à prévoir :
  - titre du document (accès sur les mots du titre)
  - mots-clés
  - type de document (rapport de recherche, cours, ...)
  - discipline
  - nom d'auteur
  - année de création
  - lieu de création
- Une **recherche de type expert**, avec l'utilisation d'opérateurs booléens (et, ou, sauf), numériques, de proximité (sur les champs de type texte), de troncatures gauches et/ou droites dont l'usage ne doit pas être implicite, de parenthèses sans limites de niveau, et cela entre des valeurs d'un même champ d'indexation ou dans des champs différents.
- Ultérieurement, pourra être envisagée une **recherche en texte intégral** sur tous les champs de la notice. Mais il faudra être vigilants sur la pertinence des résultats : le risque de bruit peut freiner l'utilisation de ce mode de recherche.

4.1.2. Visualiser un document primaire

Accès au document primaire depuis la notice.

Le temps d'affichage doit être raisonnable, même dans les pays où les réseaux sont lents. Les documents volumineux auront donc à être découpés le cas échéant.

Le système pourrait permettre d'offrir une fonction sommaire et toutes les fonctions de navigation au sein du document (page précédente, page suivante, ...)

Il faut proposer l'affichage du document dans divers formats afin d'assurer un service optimal à l'utilisateur équipé du navigateur le plus récent comme à celui qui ne dispose que d'un accès Telnet.

Rappelons toutefois que l'objet n'est pas ici de proposer un document manipulable à l'écran.

4.1.3. Imprimer le document primaire

L'utilisateur doit pouvoir aisément imprimer le document complet, même s'il n'en a visualisé qu'une partie à l'écran.

Il faut aussi qu'il puisse n'imprimer qu'un sous-ensemble de pages sélectionnées.

Dans tous les cas, il est impératif que le document soit restitué dans sa mise en page initiale.

4.1.4. Proposer un nouveau document à intégrer dans la base

l'auteur potentiel doit pouvoir entrer en contact avec l'administrateur de la base, il faut donc lui indiquer l'adresse électronique de la personne à contacter.

La numérisation et la diffusion d'un document par la BU ne pourront se faire qu'avec l'autorisation écrite de son ou de ses auteurs. Il serait bon qu'ils trouvent sur le site un formulaire d'autorisation à imprimer et à envoyer dûment signé à la BU.

En France, la loi relative à l'adoption du droit de la preuve et à la signature électronique a été adoptée en première lecture au Parlement le 13 mars 2000. Dès que la signature électronique sera techniquement opérationnelle, le site pourra proposer un formulaire à renvoyer sous forme électronique.

4.2. le site web

4.2.1. principes de navigation

Il est souhaitable que, depuis la page d'accueil, l'information recherchée soit accessible en 4 clics de souris au maximum.

4.2.1.1. Recherche dans la base

le lancement d'une recherche aboutira à un affichage en plusieurs étapes :

- d'abord une liste de notices abrégées, réduites à certains éléments (auteur, titre, discipline,...)
- la sélection de l'une d'entre elles pourra entraîner au choix l'affichage de la notice complète ou l'affichage du document primaire.

4.2.1.2. Liens à envisager

- de notice abrégée vers notice détaillée)
- de notice vers document numérique (mais pas l'inverse)
- entre productions d'un même auteur
- sur la base d'autres similitudes qui restent à préciser (mots-clés ou autres)

4.2.2. plan du site

Accueil

Interroger le catalogue

Liste de notices abrégées -> Document sélectionné

Liste de notices abrégées -> Notice complète -> Document sélectionné

Contacter la BU

Emettre des remarques sur le site

Proposer un nouveau document

Autorisation de numériser : formulaire à imprimer

Informations générales

Sur le projet : ses motivations, les évolutions prévues

Sur les aspects juridiques : rappel du Code de la Propriété Intellectuelle

4.2.3. maquette

Les fonctionnalités attendues pour le site sont illustrées par une maquette, visible dans un premier temps à l'adresse http://perso.wanadoo.fr/chamm/bu-maquette.

Les pages ont été testées sur IE5 et Netscape sur PC, écran 17" mais ne sont pas optimisées.

Les principales pages sont les suivantes :

Service Commun de Documentation de l'université de Lille1 - Littérature grise numérisée - Microsoft Internet Explorer with iHar

Fichier Edition Affichage Favoris Outils ?

Précédente Suivante Arrêter Actualiser Démarrage Rechercher Favoris Historique Courrier Imprimer Edition

Adresse C:\maquette\index.html OK

Littérature grise en ligne

- Consulter le catalogue *(prototype)*
  Le Service Commun de Documentation de l'Université des Sciences et Technologies de Lille vous donne accès en texte intégral à son fonds de littérature grise francophone.
- Vos droits d'internautes
  Quelques rappels concernant la propriété intellectuelle des documents mis en ligne.
- Pour tout savoir sur le projet
  La démarche, les métadonnées, les prochaines étapes et quelques liens utiles.
- Chercheurs et enseignants francophones
  Si vous souhaitez donner un rayonnement international à vos travaux tout en contribuant à la présence francophone sur Internet, ceci vous concerne.
- Ecrivez-nous
  Ce site vous a plu ? Vous espérez quelques améliorations ? Faites-le nous savoir.

Page mise à jour le 23 juin 2000

Poste de travail

Démarrer Navigation ... stage.doc - ... Service ... 19:33

Confier un document à la bibliothèque - Microsoft Internet Explorer with iHarvest One

Fichier Edition Affichage Favoris Outils ?

Précédente Suivante Arrêter Actualiser Démarrage Rechercher Favoris Historique Courrier Imprimer Edition

Adresse C:\Bu-Maquette\Pages\contribution.html OK

CONFIER UN DOCUMENT A LA BIBLIOTHEQUE

Nom _____ Prénom _____

Ecole/Université _____

Discipline _____

Vous souhaitez voir diffuser sur ce serveur

☐ des documents de recherche ☐ des documents d'enseignement

Nous vous communiquerons sous peu les modalités pratiques et juridiques de leur publication en ligne.

envoyer

Terminé Poste de travail

Démarrer 16:16

Recherche dans le catalogue - Microsoft Internet Explorer with iHarvest One

Fichier Edition Affichage Favoris Outils ?

Précédente Suivante Arrêter Actualiser Démarrage Rechercher Favoris Historique Courrier Imprimer Edition

Adresse C:\maquette\recherche.html OK

Si vous êtes arrivés directement sur cette page, pensez à prendre connaissance des aspects juridiques

Recherche avancée dans la base de littérature grise

spécifiez une valeur dans le ou les champs souhaités (troncature possible, par ex : Dupon)*

Auteur :

Titre :

Mots-clés :

Discipline :

Pays/Ecole : Tunisie/Ecole Nationale de Tunis

Année : 1999

Rechercher

- Prépublications *(18)*
- Rapports de recherche *(16)*
- Séminaires - Colloques *(7)*
- Positions de thèses *(32)*
- Archives *(4)*
- Travaux d'étudiants *(2)*
- Cours, travaux dirigés *(4)*

Terminé Poste de travail

Démarrer Navigation ... stage.doc - ... Recherc... 19:34

résultat de la recherche - Microsoft Internet Explorer with iHarvest One

Fichier Edition Affichage Favoris Outils ?

Précédente Suivante Arrêter Actualiser Démarrage Rechercher Favoris Historique Courrier Imprimer Edition

Adresse C:\maquette\resultat.html OK

En cliquant sur le titre d'un document, vous accédez à sa notice détaillée (format HTML).
Vous pouvez aussi :
- visualiser la notice détaillée au format **XML**, sous réserve de posséder le navigateur Internet Explorer 5 de Microsoft.
- visualiser au format **PDF** l'intégralité du document , sous réserve de posséder le lecteur Acrobat de Adobe.

Résultat de la recherche **Ecole/Univ = "Tunis" , Année = "1999" :** *3 documents trouvés*

- **Identification de contraintes dans une zone de contact : cas axisymétrique** / J.Ben Abdallah, M Bonnet. - Tunis : ENIT, 1999. - Mécanique.

  Notice détaillée [XML] ouvrir le document (1ère page 67 Ko, 25 pages au total)

- Titre 2 / Auteurs. - Tunis : labo, 1999. - Mathématiques.

  Notice détaillée [XML] ouvrir le document (1ère page 21 Ko, 7 pages au total)

- Titre 3 / Auteurs. - Tunis : labo, 1999. - Gestion.

  Notice détaillée [XML] ouvrir le document (1ère page 39 Ko, 13 pages au total)

Terminé Poste de travail

Démarrer Navigation ... stage.doc - ... résultat d... 19:35

Notice détaillée - Microsoft Internet Explorer with iHarvest One

Fichier Edition Affichage Favoris Outils ?

Précédente Suivante Arrêter Actualiser Démarrage Rechercher Favoris Historique Courrier Imprimer Edition

Adresse C:\maquette\F1999-55.html

Notice détaillée

Identification de contraintes dans une zone de contact : cas axisymétrique / J. Ben Abdallah, M. Bonnet. - Tunis : ENIT, 1999. - 23 p. : ill.; 30cm. - (Rapport de recherche /ENIT-LAMSIN;99-01)

Ecole nationale d'ingénieurs de Tunis, laboratoire de modélisation et de simulation numérique dans les sciences de l'ingénieur (ENIT-LAMSIN). - Février 1999, bibliogr.

Mécanique du contact électrostatique
Problèmes inverses
Identification de paramètres
Régularisation

ouvrir le document (1ère feuille 67 Ko, 25 feuilles au total)

Page mise à jour le 23 juin 2000

Terminé Poste de travail

Démarrer Navigation ... stage.doc - ... Notice d... 19:35

C:\Bu-Maquette\Documents\F1999-55_P00.pdf - Microsoft Internet Explorer with iHarvest One

Fichier Edition Affichage Favoris Outils ?

Précédente Suivante Arrêter Actualiser Démarrage Rechercher Favoris Historique Courrier Imprimer Edition

Adresse C:\Bu-Maquette\Documents\F1999-55_P00.pdf

Signets Vues miniatures Annotations Articles

- Titre
- Résumé
- 1 Introduction
- 2 Le problème direct
  - 2.1 Expression des dérivées premières d
  - 2.2 Calcul des déformations et du déplac
    - 2.2.1 déformation tangentielle u1.1
    - 2.2.2 Déformation tangentielle u2,2
    - 2.2.3 Déplacement vertical u3
  - 2.3 Discrétisation par des éléments de fr
- 3 Le problème inverse
  - 3.1 Régularisation par décomposition en
  - 3.2 Régularisation par pénalisation des o
- 4 Résultats
  - 4.1 Résultats obtenus par pénalisation d
  - 4.2 Résultats obtenus par décomposition
- 5 Conclusion et perspectives
- Annexe
- Références
- RAPPORTS DE RECHERCHE PARUS EN 19
- RAPPORTS DE RECHERCHE PARUS EN 19

ENIT LAMSIN

Imprimer le dossier

Quitter

Identification de contraintes dans une zone de contact : Cas axisymétrique

J. Ben Abdallah, M. Bonnet

Rapport de Recherche
N° : 99/01
Février 1999

77% 1 sur 1 210,3 x 297 mm

Terminé Poste de travail

Démarrer 16:18

4.3. fonctions d'administration

4.3.1. gestion bibliographique

Le personnel de la BU aura à gérer les notices bibliographiques qui décrivent les documents disponibles dans la base. Il s'agira en particulier de

4.3.1.1. insérer une notice

Prévoir de pouvoir partir d'une notice existante, prévoir un contrôle des doublons.

Les documents de recherche sont souvent réalisés en LaTex, il peut arriver qu'un titre contienne des formules ou signes mathématiques. La solution actuellement retenue à la BU dans le cas des thèses est de saisir phonétiquement les éventuels symboles contenus dans les titres.

(Le descriptif des champs composant une notice sera introduit dans le chapitre suivant)

4.3.1.2. modifier une notice

accès à la notice par la référence du document ou par le nom d'auteur

4.3.1.3. supprimer une notice

A tout moment, un auteur peut demander à ce que l'un des documents soit retiré de la base. Une solution est de supprimer la notice et donc l'accès au document. Il serait toutefois plus sûr de retirer au plus tôt le document primaire lui-même.

4.3.1.4. ajouter ou supprimer un document

A prévoir, soit directement, soit par le biais d'une demande à l'informatique.

4.3.2. autres outils d'administration

4.3.2.1. listes de contrôle

Il faut pouvoir imprimer des listes de documents par discipline, par lieu de création, par auteur ou par d'autres critères à préciser

4.3.2.2. suivi des connexions

Il est essentiel de disposer de statistiques sur les accès à la base : depuis quel pays, durée, pages visitées, ...

4.3.2.3. gérer la base des auteurs

Fonctionnalité non retenue pour l'instant. Il s'agirait de proposer un service supplémentaire aux auteurs, CV en ligne par exemple, à l'instar de ce que nous avons rencontré sur des sites similaires.

5. LES DONNEES

5.1. règles de gestion

- Chaque document est stocké un un seul exemplaire,
- par nature un document risque peu de figurer au catalogue d'une autre bibliothèque,
- on ne gère pas de versions successives d'un même document,
- un document peut avoir fait l'objet d'une publication non commerciale,
- un document est écrit par un ou plusieurs auteurs,
- un auteur est identifié par son nom et son prénom. Pour différencier les homonymes, il faudra disposer d'une information discriminante (date de naissance ? discipline ?)

5.2. entités

Pour décrire les informations à stocker, il est bon d'avoir une approche relationnelle, sans présumer pour autant de la solution technique qui sera retenue.

Il restera à élaborer un dictionnaire de données complet, indiquant le format de chaque champ et précisant s'il est facultatif ou obligatoire, répétable ou non. Les principales entités sont les suivantes :

5.2.1. Auteur

IdentAuteur : nom, prénom(, datnaiss ?)
Dernière fonction connue, à quelle date
Adresse professionnelle
Adresse électronique

5.2.2. Document source

IdentDoc : numéro d'inventaire
Origine éditoriale
    Lieu de création
    Référence d'une éventuelle publication
description matérielle (aspect extérieur : format, nb pages)
indications sur une biblio en fin de document
discipline
mots clés
résumé éventuel
table des matières éventuelle
typologie
Date de réception par la BU
Format d'acquisition par la BU (papier, RTF, PS, ...)

5.2.3. Auteur-document

IdentAuteur
IdentDoc
Date d'accord pour numérisation
Date de retrait (surtout si pré-publication) si on choisit de garder la notice

5.2.4. Document numérique

Date de numérisation ou de conversion
Auteur de numérisation (BU ou prestataire)
Format de numérisation
Suivi de vie

Partie III - Préalables sur les documents numériques

6. LES FORMATS DE DOCUMENTS NUMERIQUES

L'édition électronique repose sur la numérisation des textes et des documents, c'est-à-dire sur leur codage en une suite de 0 et de 1 manipulables par un ordinateur. Il y a deux grandes façons de procéder. La première consiste à coder le

document en mode image, on obtient ainsi une reproduction électronique du document qui en restitue fidèlement la mise en page. La deuxième technique, qui ne peut s'appliquer qu'aux textes, consiste à coder les caractères selon une norme qui attribue à chacun une valeur numérique déterminée.

La version numérisée du document doit être exploitable et pérenne dans la mesure du possible au même titre que le document imprimé. Le format constitue un élément clé pour la pérénité du document numérisé car il conditionne sa lisibilité à moyen et long terme.

6.1. le mode image

Nous évoquerons ici le mode bitmap où une matrice de points, colorés ou non, forme une image ou un caractère. Il s'oppose au mode vectoriel où les éléments constitutifs d'un dessin sont décrits mathématiquement.

6.1.1. numérisation par scanner

Appliquée à des documents de types images, texte imprimé, etc, la numérisation consiste, à l'aide du matériel approprié appelé "numériseur" (en anglais "*scanner*"), à obtenir un fichier informatique à partir d'un document analogique, par exemple une page d'un livre imprimé.

La technique de numérisation consiste à découper le document en petits carrés appelés pixels « picture elements » avec pour chaque carré des caractéristiques propres (luminosité, couleur, niveau de gris etc. etc). Plus le nombre de points augmente, plus la résolution du scanner est importante. La résolution du scanner est ainsi exprimée en dpi « Dots Per Inch » soit « pixels par pouce », le pouce valant 2,54 cm. Aujourd'hui la plupart des scanners du marché proposent une résolution allant de 75/75 à 600/600 dpi, voire 1200 ou 2400/2400 pour les modèles professionnels.

Le choix de la résolution dépend de la qualité du document initial (support, caractères, encre, images, couleurs etc.) mais aussi du type d'exploitation désiré :

- L'archivage nécessite une résolution haute, 600 dpi au minimum
- La lecture à l'écran peut se satisfaire d'une résolution moyenne d'environ 300 dpi
- A l'impression, une résolution de 300 dpi s'avère suffisante. Elle entraîne une légère perte de précision par rapport au document original, ce qui ne peut que décourager les reproductions pirates.

Il existe différents modèles de scanners. Les plus simples, semblables à des photocopieuses, permettent la numérisation de feuilles volantes. Numériser des ouvrages reliés nécessite du matériel plus perfectionné, fonctionnant selon le principe d'une caméra numérique.

6.1.2. des fichiers volumineux

Quel que soit le document de base, le résultat de cette numérisation est un fichier informatique de type "image", souvent volumineux. Le stockage d'un carré ou pixel de base varie selon les caractéristiques du document d'origine :

- s'il est en noir et blanc, chaque pixel sera codé sur un seul bit, soit un 8ème d'octet
- s'il contient une gradation de gris, le pixel sera souvent codé sur 8 bits, soit 1 octet, permettant ainsi 256 niveaux de gris
- si enfin il est en couleurs, le codage du pixel pourra nécessiter jusqu'à 32 bits, soit 4 octets, d'où un volume de stockage souvent très important

Diverses techniques de compression permettent de réduire le volume des fichiers de type image, parfois au prix d'une perte de précision. Pour les documents textuels, la plus utilisée est la compression CCITT groupe IV qui optimise le codage de suites de pixels identiques, et ne génère aucune perte d'informations.

6.1.3. les avantages et inconvénients du mode image

- *avantages*
  - c'est le mode le plus simple à réaliser
  - c'est le moins coûteux
  - il restitue entièrement l'apparence et le contenu du document original
  - C'est le seul format fiable par rapport au document original.
- *inconvénients*
  - il produit des fichiers informatiques lourds qui nécessitent beaucoup de place sur le support
  - sans indexation, ni mots-clés, le texte ne peut être que feuilleté. Ce mode interdit toute recherche sur le texte.

6.1.4. Les formats de fichiers images les plus courants

Les plus souvent utilisés dans les bibliothèques virtuelles sont

- le format TIFF (Tagged Image File Format)

- le format Adobe® PDF (Portable Document Format)
  Bien que format propriétaire, PDF est un standard de fait, car produit dominant sur le marché. Il permet de conserver l'apparence de la mise en page originale. [1]

6.2. le mode texte

6.2.1. la Reconnaissance Optique de Caractères (OCR)

Le fichier image issu de la numérisation peut être transformé en un document texte via un logiciel d'OCR: (« Optical Character Recognisation » ou « Reconnaissance Optique de Caractère »). L'OCR permet ainsi à cette image d'être exploitable dans n'importe quel logiciel de traitement de texte, et d'être archivée à moindres frais (les documents textes sont moins gourmands en place que les documents images sur le disque dur)

Le document texte se présente comme une suite de caractères codés suivant une norme qui attribue à chacun une valeur numérique déterminée. La plus ancienne et la plus universelle de ces normes est le code A.S.C.I.I. (American Standard for Communication International Interchange). Malheureusement, elle ne permet de coder que les caractères de l'alphabet des langues anglo-saxonnes. Des normes plus récentes, comme Unicode ou l'ISO10646, codent les informations textuelles sur plusieurs octets et non un seul octet comme en ASCII. Ce nouveau mode de codage permet de représenter toutes les écritures du monde.

6.2.2. Avantages et inconvénients du mode texte

➢ *avantages*

- permet une recherche en plein texte et une navigation au sein du document. Par exemple "naviguer" d'une table des matières vers un chapitre, ou de document à document.
- permet un déplacement rapide à l'intérieur d'un document
- permet une interrogation en langage naturel, facilitant l'accès simple et direct à l'information à tout utilisateur (expert ou occasionnel)
- permet d'associer la question à des critères qui caractérisent les documents (date, auteur, thème...)
- permet de recoder en SGML, par exemple le titre, le titre de paragraphe

[1] pour en savoir plus, se reporter à l'observatoire des NTIC publié par les étudiants du DESS SID de Lille3

- satisfaisant pour la recherche

➢ *Inconvénients*

- l'OCR modifie la présentation du document original
- la recherche en plein texte peut créer du "bruit".
- l'OCR ne peut reconnaître actuellement, les caractères non latins (dont les caractères gothiques), les signes diacritiques, les lettres manuscrites.

Le mode texte rend possible l'automatisation de travaux fastidieux comme la recherche de tous les contextes d'un mot ou d'une chaîne de caractères. Il présente cependant l'inconvénient de faire disparaître la structure formelle du texte. C'est pour pallier cette perte que se développent aujourd'hui des langages de balisages capables de renseigner la machine sur la structure logique d'un document (titres, paragraphes, etc.).

6.3. mode structuré : le balisage de textes

Le mode structuré consiste à encapsuler un document plein texte dans une série de balises qui vont lui restituer, à l'exploitation (lecture mais aussi manipulations de corpus important) une partie importante des informations et des contenus qui lui ont été ôtés lors du passage au plein texte.

Des solutions techniques en vue d'organiser l'information ont été recherchées dès les années 60 avec le langage GML (Generalized Markup Langage), balisage généralisé, qui a donné naissance à SGML (Standard Generalized Markup Langage), norme existant depuis 1986, mais dont le domaine d'application est resté limité à la documentation technique et à l'informatique éditoriale.

On peut en fait baliser un document en plein texte selon deux grands types d'outils : les langages de description et les langages de structuration.

6.3.1. HTML, langage de description

Le plus connu des langages de description est sans aucun doute le HTML (*HyperText Mark-up Language*), mise en œuvre de SGML utilisée pour coder les pages disponibles sur les serveurs web. A l'origine, les balises HTML appliquées à un texte étaient destinées à révéler sa structure, sans prêter attention à sa signification ou à sa présentation. Par exemple, un paragraphe est signalé non pas par un saut de ligne ou un retrait, mais parce qu'il est encadré par le couple de

balise . L'agent utilisateur (le plus souvent un navigateur comme Explorer ou Netscape) choisissait alors la présentation de ce paragraphe.

Les tailles d'affichage des caractères étaient exprimées non pas "absolument" mais "relativement" au navigateur, paramétré par son utilisateur. De même, on signalait qu'un paragraphe était un titre en l'encadrant par la balise <h> </h>, le navigateur prenant en charge la manière de mettre en évidence graphiquement ce titre lors de son affichage, en général par incrémentation de la taille du caractère (+ 2 points) et on disposait de 6 niveaux de titre.

Le langage HTML a considérablement évolué depuis, mais ces concepts fondamentaux sont toujours en action et il est tout à fait envisageable de fournir une version "structurée" d'un document en y recourant, d'autant plus qu'il est possible de réinsérer dans un texte, à un endroit précis, ses illustrations par l'intermédiaire d'une balise qui va commander l'affichage d'une image, ou créer un lien vers elle

Si le mode "plein texte" a été réalisé selon les recommandations faites plus haut, on peut le coder en HTML via un filtre de conversion, et en minimisant absolument l'intervention d'un opérateur humain, en respectant ces même recommandations. Comment sait-on qu'un texte ASCCI "propre" a été décrit correctement en HTML ? Par le principe de l'inversion : en retirant toutes les balises HTML, sans rien faire d'autre, on doit retomber exactement sur le texte ASCII tel qu'il était auparavant.

6.3.2. les langages de structuration

Le principe des langages de structuration est un peu identique à celui des langages de description mais, cette fois, certaines balises vont de plus " décrire " les unités qui structurent le document.

Prenons l'exemple d'une pièce de théâtre en alexandrins.

Elle est constituée d'actes, qui se découpent en scènes, elles-mêmes décomposables en tirades, formées de vers qu'on peut séparer en deux hémistiches.

On va donc distribuer dans le texte des balises <hémistiche>, <vers>, <tirade>, <scène>, <acte> pour le structurer.

Ces balises permettront à la fois d'associer une "présentation" graphique unifiée et cohérente lors de l'exploitation, mais aussi de manipuler le texte (ou un corpus) en utilisant les balises comme champs interrogeables.

6.3.3. présentation de XML

6.3.3.1. historique

Le besoin d'un méta-langage adapté au Web est à l'origine de la création de XML (*eXtensible Mark-up Language*)., sous-ensemble de la norme SGML. Son but est de permettre au SGML générique d'être transmis, reçu et traité sur le web comme l'est le HTML aujourd'hui. XML a été conçu pour être facile à mettre en œuvre et pour être interopérable avec SGML et HTML.

Disponible depuis 1996, XML a fait en 1998 l'objet d'une recommandation du consortium W3C[1]. Il semble promis à un brillant avenir puisque qu'il est maintenant au cœur de la stratégie de Microsoft

Méta langage de structuration des données, XML représente une série de règles pour définir des balises ayant du sens afin de structurer un document en différentes sections bien identifiées. Un document XML contient des balises décrivant la structure et la signification de son contenu mais pas son aspect, contrairement à HTML.

6.3.3.2. Intérêt du XML

- format de données non propriétaire,
- lisibilité par les humains comme par les machines,
- possibilité de n'afficher que certaines données stockées,
- pérennité de l'information,
- données auto-documentées (appréciable pour les informaticiens-archéologues),
- indépendance de l'aspect physique du document vis-à-vis de sa structure logique, le document devient manipulable selon le profil de l'utilisateur et selon la plateforme utilisée.
- codage des caractères sur 2 octets : XML implique l'usage d'Unicode

6.3.3.3. Notion de DTD

XML permet à chaque profession de créer son propre langage de description spécifique, c'est-à-dire de créer les balises dont la profession a besoin dans son

1 recommandation XML 1.0 du 10/02/98

domaine Ces balises sont décrites dans un document indépendant appelé Document Type Definition (DTD).

La DTD illustre la structuration des données, en fournissant la liste des éléments, attributs, notations et entités que contient le document, ainsi que les règles des relations qui les régissent. Un document XML est dit « valide » s'il est conforme à ce qui est défini dans sa DTD associée.

Un document XML peut ne pas contenir de déclaration de type de document (DTD), nous parlons dans ce cas de document « bien formé . L'avantage principal de la notion de document bien formé est la possibilité d'extraire des fragments de document et de les traiter ou de les échanger sans leur attribuer de DTD.

C'est là l'une des différences notables entre XML et SGML qui, lui, impose l'existence d'une DTD.

6.3.3.4. XML et internet

En XML, il est indispensable d'indiquer au navigateur comment gérer et afficher les documents. Deux solutions existent : feuilles de style CSS, langage XSL

➢ *CSS (Cascading Style Sheet)*

CSS est un langage simple, conçu en 1996 pour HTML. Il sert à contrôler l'application de style à du texte (choix de police, mise en gras...).

On appelle 'feuille de style' le document contenant ces paramètres d'affichage. Si le texte ne fait référence à aucune feuille de style c'est par défaut celle du navigateur qui est utilisée.

Avantage de ce principe : à des besoins variés peuvent correspondre différentes feuilles de style, selon que les documents doivent être consultés sur le web, imprimés ...On ne touche pas au contenu du document.

➢ *XSL (Extensible Style Language)*

Plus complexe que CSS mais mieux adapté aux navigateurs web, XSL se compose de deux langages :

- **XSLT**, un puissant langage de transformation pour sélectionner et/ou réorganiser l'information en fonction des besoins. Cette transformation s'appuie sur la représentation hiérarchique des documents XML, les noms des éléments, les noms et valeurs des attributs ainsi que sur le contenu lui même.
- **XSL FO**, un langage autorisant la création d'objets de formatage et de description de leurs propriétés. Ce langage permet par exemple de créer des

objets de type paragraphe, séquence, tableau, cellule, image etc. et de leur attribuer des propriétés d'affichage telles que l'espace avant et après, la fonte, les couleurs, etc. Ce langage est totalement indépendant du support de sortie. Un tel mécanisme permet de conserver toute la richesse sémantique du codage initial dans un seul fichier, dont on peut extraire, selon les besoins, de nombreux documents différents.

7. LE REFERENCEMENT DES DOCUMENTS NUMERIQUES

Outils traditionnels d'accès à l'information, les notices bibliographiques ont-elles encore des raisons d'être pour les documents numériques ?

Les avis sont partagés sur ce sujet, source de débats passionnés[1]. Deux tendances s'opposent :

- une approche 'moderne', venue du web : les métadonnées
- une approche plus classique, venue des bibliothèques : les notices enrichies

7.1. web et métadonnées

La notion de métadonnées (données sur les données) a été introduite pour référencer les pages web, afin d'améliorer l'efficacité des moteurs de recherche.

Une première mise en œuvre en est le jeu de balises META parfois intégrées en tête de pages HTML. Elles sont toutefois peu utilisées par les concepteurs de pages qui n'en voient pas la nécessité : bien des moteurs les ignorent.

Pour que sites et moteurs se comprennent, un standard de métadonnées s'est rapidement avéré nécessaire. Ainsi est née la norme RDF.

7.1.1. XML pour les métadonnées : la norme RDF[2]

Le format RDF, ou Resource Description Framework, est une application XML recommandée par le W3C pour le codage, l'échange et la réutilisation de métadonnées structurées.

---

[1] pour s'en convaincre : http://ifla.inist.fr/IV/ifla64/007-126f.htm
[2] http://www.w3.org/TR/REC-rdf-syntax/

RDF définit une infrastructure commune pour la représentation des métadonnées, infrastructure particulièrement adaptée à la description de l'architecture des sites et des pages web.

Un élément RDF est lié à une ressource et est de la forme (ressource, propriété, valeur).

7.1.2. le Dublin Core[1]

La plus connue des applications de la norme RDF est le Dublin Core.

Le Dublin Core est un ensemble de 15 champs (titre, créateur, sujet, ...)., prévu surtout pour la description de pages Internet à l'intention des moteurs de recherche

Une de ses motivations premières est que les auteurs puissent fournir leurs propres descriptions

7.2. bibliothèques et notices enrichies

7.2.1. MARC, expert es notices bibliographiques

Le format utilisé pour les catalogues informatisés de bibliothèques est le format Marc (plus précisément l'une de ses variantes). MARC permet une description très fine et exhaustive des documents.

Initialement pensé pour des ouvrages imprimés, Marc s'est progressivement adapté à la description de 'non-livres' dont récemment les documents électroniques en ligne. En témoigne la création du champ 856 qui définit le lien entre la notice descriptive et le document électronique disponible sur le réseau.

7.2.2. Quelles solutions pour les bibliothèques ?

Avant de répondre à cette question, il convient de s'interroger sur la nature des documents à référencer et sur la finalité recherchée.

Ainsi les bibliothèques sont de plus en plus nombreuses à proposer à leur public le référencement de pages web, non figées. Ces pages, documents vivants par excellence, ne justifient pas une description bibliographique exhaustive, risquant

[1] http://www-rocq.inria.fr/~vercoust/METADATA/DC-fr.1.1.html

d'être d'autant plus vite obsolète qu'elle est pointue. Quelques métadonnées de base comme celles du Dublin Core peuvent suffire.

L'approche pourra être diamétralement opposée s'il s'agit par exemple d'archives dont la version numérisée est appelée à remplacer la version imprimée.

Dans le cas d'articles scientifiques numériques, nous avons rencontré les 2 solutions : Dublin Core, notice classique. Notons que le choix d'une option semblait dépendre des sensibilités personnelles des différents acteurs plus que des buts poursuivis.

Partie IV - La chaîne de traitement : solutions techniques envisageables

8. CONTRAINTES DE MISE EN OEUVRE

8.1. l'environnement technique

8.1.1. matériels et réseaux

Le Centre de Ressources Informatiques (CRI[1]) de Lille1 assure, entre autres missions, l'hébergement et la maintenance du serveur web de l'université, la gestion des centres de documentation et bibliothèques de composantes.

Dans le domaine des bases de données documentaires, les développements réalisés et maintenus par le centre, sont des applications de consultation et de gestion de fonds documentaires et de catalogues.

Les bases de données informatisées sont hébergées sur une machine DEC ALPHA 4000 du centre.

[1] http://ustl.univ-lille1.fr/cri/

8.1.2. outils logiciels

8.1.2.1. logiciels de numérisation

A prévoir si on scanne avec OCR. Acrobat Capture par ex. Inutile dans un 1er temps car sous-traitance.

8.1.2.2. création de notices XML

Plusieurs méthodes peuvent être utilisées pour créer un document XML :

- Rien n'interdit de saisir texte et balises sous Notepad ou tout autre éditeur de texte, mais il existe heureusement des solutions moins fastidieuses.

- On peut aussi utiliser l'un des éditeurs XML disponibles sur le marché. Ces logiciels ont l'avantage d'offrir une interface graphique conviviale. Certains assurent la validation (parsing) du document balisé, par rapprochement avec un fichier DTD de référence.
  C'est vers ce type d'outil que nous nous sommes tournés :
  - utilisation de 'XMLspy', en version d'évaluation, pour création de TEI Header,
  - utilisation de 'XML notepad', éditeur basique mais gratuit de Microsof, pour création de notice Marc XML.

- Dans certains cas, un convertisseur (plus couramment 'moulinette') peut s'avérer utile : il permet de convertir au format XML un document existant, sans avoir à le réécrire. Il existe ainsi des convertisseurs de MARC vers XML.

8.1.2.3. outils de stockage et de consultation

Les bases de données documentaires sont gérées par le CRI sous le logiciel TEXTO de la société Cincom.

Les programmes de consultation, écrits avec le logiciel TEXTO-WEB, exploitent une interface client/serveur de type WEB et permettent d' interroger les fonds documentaires de la BU.

Il est prévu de remplacer prochainement TEXTO ; le remplaçant pressenti est CinDoc, diffusé lui aussi par Cincom.

CinDoc offre des outils intégrés pour acquérir, archiver et rechercher toute forme d'information. Il intègre un module web, permet l'interrogation en texte intégral et

offre la précision de la recherche structurée de données. Il devrait pouvoir gérer des notices au format XML.

8.2. contraintes éditoriales

8.2.1. Validation des sources

La possibilité offerte à chacun de publier des documents accessibles à tous remet en cause la fonction d'autorité intellectuelle des revues ou des maisons d'édition. L'instabilité des différents sites du réseau et la malléabilité des textes numérisés posent la question de fiabilité des publications électroniques.

8.2.1.1. Notion de comité éditorial ou de comité de lecture :

- dans le cas d'une revue, le document est soumis à un comité de validation avant publication
- dans le cas des thèses ou habilitations, la soutenance fait foi
- ***quid pour la littérature grise en général ?***

Ce sujet a déjà fait couler beaucoup d'encre ...

Une idée est de trouver une communauté de chercheurs avec qui travailler. C'est la solution retenue par divers sites mono-disciplinaires (médecine, ...) voire plus généraux (comme le site 'Université en ligne' auquel participent Paris 2 et Lille3).

Une autre approche plus répandue est de sélectionner des auteurs qui offrent une garantie suffisante. Ainsi le site Rondpoint.org qui propose des cours en ligne à destination des pays du Sud, et auquel participe l'EUDIL.

8.2.2. Aspects juridiques

Tout projet de numérisation soulève des questions juridiques, qu'il s'agisse de droit d'auteur, de relations avec les prestataires ou avec les éditeurs.

En jouant indirectement un rôle d'éditeur, la BU endosse vis à vis des auteurs des devoirs de garantie d'intégrité du document et de respect du droit d'auteur.

Il lui incombe aussi d'alerter clairement les enseignants/chercheurs sur leurs droits et devoirs : rien ne leur interdit de publier le même article par divers canaux non commerciaux. Le cas des pré-publications est plus sensible : si une revue retient l'article qui avait été soumis pour publication, l'auteur se doit de supprimer toute autre diffusion publique (sauf accord explicite de l'éditeur commercial).

Dans le cadre de ce projet, chaque personne qui confie un document à la BU signe en parallèle une autorisation de numériser et diffuser publiquement. Cette autorisation est-elle adaptée à tous les cas de figure ?

D'après le code de la propriété intellectuelle, l'auteur est seul propriétaire de son œuvre.

- dans le cas de travaux d'étudiants, la seule signature de l'étudiant suffit,
- nous avons rencontré le cas d'un rapport de recherche réalisé à la demande du ministère. Qu'en est-il dans ce cas ? (la question reste posée)

8.2.2.1. Protéger les auteurs

Le devoir de protection des auteurs existe même si nous diffusons à titre gratuit.

Il incombe de rappeler aux internautes leurs limites : il faudra intégrer dans le site un avertissement contre la copie et la diffusion collective, en rappelant le code de la propriété intellectuelle.

Un rappel exhaustif des textes à respecter voire des sanctions encourues en cas de non respect du droit d'auteur informe les 'rares' personnes qui font exception au « nul n'est censé ignorer la loi » et peut faire réfléchir les contrevenants potentiels.

Certains sites gratuits enregistrent l'identité des visiteurs qui consultent leur base de textes intégraux ; une solution qui peut s'avérer dissuasive (mais avec quelle valeur légale ?)

8.2.2.2. Authentifier les documents

Une préoccupation majeure est d'assurer l'intégrité du fond et de la forme des documents échangés.

L'éthique du bibliothécaire vue par G Teasdale :

« J'ai l'intuition qu'aucun bibliothécaire ne mettrait à la disposition de ses lecteurs, ceux du présent ou ceux du futur, un livre imprimé dont il saurait que le contenu aurait été, d'une façon ou d'une autre, altéré, modifié, censuré, amputé ; et qu'aucun bibliothécaire ne se livrerait à à ce genre de manipulation. Pourquoi cela serait-il différent pour des fonds électroniques ? La responsabilité des professionnels des bibliothéques est engagée sur ce point. C'est sur cette croyance que les livres qu'ils trouvent en bibliothèque sont intacts que les lecteurs fondent leur confiance sur les ressources dont ils disposent dans ces établissements. »

Comment la BU peut-elle se prémunir contre des copieurs mal intentionnés ?

D'abord en estampillant les documents primaires avant numérisation (l'apposition du numéro d'inventaire joue ce rôle)

Ensuite en veillant à ce que le document numérisé porte en filigrane une mention de copyright (point à préciser au prestataire). Cette mention est invisible à l'écran mais apparaît systématiquement à l'impression. Elle a été mise en œuvre pour le premier test fait par le prestataire.

8.2.2.3. Référencer les auteurs

A l'instar de ce qui est proposé sur d'autres sites, la BU pourrait proposer aux auteurs d'enregistrer et diffuser quelques infos 'publicitaires' à leur sujet.

Dans ce cas, penser à respecter la loi "Informatique et Liberté" : tout traitement automatisé d'informations "nominatives » doit faire l'objet d'une déclaration à la CNIL (Commission Nationale Informatique et Liberté).

Par ailleurs, la CNIL recommande que l'accord des personnes soit recueilli préalablement à toute diffusion sur Internet de données les concernant. On peut aussi les informer de leur droit de s'opposer, partiellement ou totalement, à cette

diffusion sur Internet mais que leur accord sera réputé tacitement acquis en l'absence de réponse de leur part au delà d'un certain délai (1 mois, par exemple).

Les personnes concernées doivent en outre être informées de l'existence et des modalités d'exercice du droit d'accès aux informations qui les concernent et du droit de les faire modifier (changement de nom, d'adresse, de fonctions, etc), rectifier (en cas d'erreur) ou supprimer.

Voir ce qu'implique le simple fait de citer les adresses électroniques.

9. LA COLLECTE DES DOCUMENTS

Le fonds de littérature grise de la BU est appelé à s'enrichir régulièrement. Nous traitons ici de la collecte des nouveaux documents.

9.1. format d'acquisition des documents

9.1.1. Acquisition sous forme électronique

Les nouveaux documents qui nous parviennent ont le plus souvent été produits sous environnement informatique.

Une solution, pour qu'ils soient directement exploitables, serait de les homogénéiser dès leur conception, en fournissant aux auteurs des recommandations sur le traitement de texte et les styles à utiliser. Cette démarche commence à se mettre en place au sein d'établissements qui développent leurs propres serveurs, de thèses en particulier. Elle n'est pas envisageable dans notre cas, où les documents sont de provenances et de natures diverses.

S'ils ont été conçus avec les outils de Microsoft, ils peuvent être récupérés au format RTF (texte). Dans le cas contraire, il faut demander à ce qu'ils soient fournis au format d'impression Postscript, ce qui permettra de les imprimer en local et de les convertir au format PDF texte+image grâce au logiciel Acrobat Distiller. En cas de problème, demander la version imprimée.

Une option de sécurité est de demander systématiquement la version imprimée pour contrôle.

9.1.2. Liens vers des ressources en ligne

Parmi les chercheurs contactés, certains proposent d'établir un lien avec leurs articles disponibles sur Internet. Il s'agirait alors d'inclure au serveur une fonctionnalité de portail, ce n'est pas l'objet de ce projet.

Rien n'interdit par contre d'intégrer dans la base de la BU une copie de ces articles, offrant par-là une meilleure visibilité à leurs auteurs.

9.1.3. Acquisition sous forme imprimée seulement

Si aucune version numérique n'est disponible, les documents seront collectés dans leur version imprimée.

Afin de faciliter la numérisation, il est souhaitable qu'ils soient tous au format A4 (ou inférieur) et non reliés. Dans le cas de documents reliés, un démontage est possible pour se ramener au feuille à feuille, dans la mesure où ces versions imprimées ne seront pas mises à la disposition du public.

9.2. choisir un mode de numérisation

La numérisation de ces lots de feuillets A4 nécessite l'usage d'un scanner à plat.

9.2.1. numériser en interne

Numériser en interne peut être envisagé, la manipulation du scanner à plat ne nécessitant pas de compétences particulières. Mais un tel investissement en matériel, logiciel et temps est-il justifié ? Difficile de le savoir encore alors que le volume de feuillets et la fréquence de numérisation ne sont pas connus.

Dans l'immédiat, la BU a choisi de confier cette étape à un prestataire.

9.2.2. sous-traiter la numérisation

Avant d'être confiés au prestataire, les documents imprimés doivent être préparés : sur chacun d'eux doit figurer un numéro d'identification qui constituera le nom du fichier résultat.

L'identifiant retenu est le numéro d'inventaire qui figure systématiquement en première page de chaque document. En retour de numérisation, le prestataire livre les CD-Rom comprenant les fichiers résultats. Il s'engage sur la qualité du travail fourni.

Un premier test de numérisation a été réalisé dans le cadre de ce projet. Il a révélé la nécessité de connaître un minimum d'aspects techniques pour un dialogue efficace avec le sous-traitant.

10. STOCKAGE

Précisons d'abord que c'est la version imprimée du document qui fera office d'archive. Ceci n'exclut pas la nécessité de stocker le document numérisé dans un format pérenne et évolutif.

10.1. mode image

C'est le mode obtenu par défaut en sortie de scanner, enregistré le plus souvent au format TIFF ou PDF image.

➢ *avantages*

- il permet, de façon simple et relativement peu onéreuse, de restituer entièrement l'apparence et le contenu du document original.

➢ *inconvénients*

- sauf si son contenu permet une compression efficace, le fichier résultat est volumineux.
- Il est totalement impossible d'interpréter le contenu.

10.2. mode image+texte

C'est le mode proposé par le logiciel Adobe PDF.

➢ *avantages*

- il permet à la fois de restituer l'apparence du document original et d'exploiter tout ou partie du contenu (table des matières, résumé, ...)

➢ *inconvénients*

- le fichier résultat est volumineux. Pour qu'il soit exploitable et pérenne, il semble qu'il faille le stocker sous forme non compressée.

10.3. mode texte structuré

texte en SGML ou XML

➢ *avantages*

- mode de stockage le moins volumineux
- permet une recherche en plein texte et une navigation au sein du document

➢ *Inconvénients*

- long à mettre en œuvre surtout si une phase d'OCR est nécessaire,
- risque de perte de la présentation du document original. En XML, elle peut toutefois être stockée dans la structure du document et reconstituée via l'écriture d'un programme XSL de présentation.
- La recherche en plein texte est souvent peu pertinente (bruit)

11. REFERENCEMENT

Avant ou après l'éventuelle numérisation, il convient de procéder au traitement intellectuel : catalogage et indexation

11.1. les informations secondaires

La BU proposera un accès centralisé à toute la base de littérature grise. Quel que soit leur type, il est donc essentiel que tous les documents accessibles soient décrits de façon homogène.

Il faut prévoir de stocker toutes des données qui répondront à l'une ou l'autre de ces attentes :

- Description bibliographique (pour recherche ou affichage dans l'interface utilisateur surtout),
- Description du contenu : mots-clés, table des matières, voire résumé
- Description physique, informations sur le cycle de vie du document numérisé
- Données susceptibles de servir dans une version ultérieure (sur les auteurs par exemple)

11.2. des codifications normalisées

L'enregistrement de ces informations secondaires devra respecter les normes de catalogage en vigueur dans les bibliothèques : règles de transcription des données, d'ordre des éléments, de ponctuation (le cas échéant). Normes françaises : AFNOR, normes internationales : IFLA

11.2.1. identification locale

quand le document d'origine est sous forme papier, il possède un code inventaire local, apposé par la BU sur sa 1ère page :

- dans le cas de brochures, le code inventaire est de la forme <lettre><année>-<num chrono>, avec lettre = A si sciences, F si techniques, G si sciences éco ou humaines

- dans le cas de périodiques, le code inventaire est de la forme <lettre>P<num périodique>-<année>-<num chrono>, avec lettre de A à Z selon la discipline, conformément aux 'instructions de 1962'.

11.2.2. Autres identifiants pour ressources numériques

Dans un souci d'ouverture, il serait opportun de souscrire à un mode d'identification normalisé pour les documents numériques. Mais lequel ?

➢ *L'ISRN*

L'équivalent de l'ISBN pour les docs numériques est l'ISRN. Il est lié, par définition, à un dépôt légal et ne nous concerne pas.

➢ *Le DOI*[1]

Créé à l'initiative des éditeurs et diffuseurs, le DOI ou Digital Object Identifier est un identifiant de ressource numérique. Il semble avoir été conçu dans une approche commerciale avant tout, sa structure permettant de contrôler l'accès des utilisateurs aux documents électroniques par l'intermédiaire d'un réseau de serveurs de résolution DOI, contrôlés par les éditeurs qui participent au système. Nous ne le retiendrons donc pas

11.2.3. classification par discipline

La classification couramment utilisée en bibliothèque est la classification décimale Dewey (ou CDD). Conçue en 1876 par Melvil Dewey, bibliothécaire

[1] http://www.doi.org/about_the_doi.html/

américain, elle répartit l'ensemble des connaissances en dix classes numérotées de 000 à 900, elles-mêmes divisées en dix sous-classes, etc.

- 000 Généralités
- 100 Philosophie
- 200 Religion
- 300 Sciences sociales
- 400 Langues
- 500 Sciences pures
- 600 Techniques
- 700 Arts
- 800 Littérature
- 900 Géographie, histoire

L'ensemble est constamment remis à jour : la BU utilise actuellement la 21e édition de 1996 (plus de 4000 pages sur quatre volumes).

Cette nomenclature, qui a pour vocation d'être généraliste, n'est pas toujours la mieux adaptée à l'usage des chercheurs. Les spécialistes en sciences économiques, par exemple, préfèrent l'usage de la codification JEL (Journal of Economic Literature).

La CDD reste malgré tout la plus adéquate pour la base hétérogène de littérature grise. Elle permet aussi de garder une cohérence avec les divers catalogues de bibliothèques auxquels la base pourrait ultérieurement être rattachée.

Dans le cas de ce projet, ne pas hésiter à associer plusieurs indices Dewey à un document qui apparaît comme pluridisciplinaire. Les documents ne seront consultables que sous leur forme numérique, la notion de cote de rangement souvent liée à la Dewey n'a ici pas de raison d'être.

11.2.4. classification par type de document

La codification Dewey permet d'indiquer le type de document (rapport de recherche, note de cours, ...), mais de façon non standardisée et donc peu satisfaisante.

Pour coder le type d'un document, la solution envisagée est d'utiliser les subdivisions communes de la Classification Décimale Universelle(CDU) qui obéit toujours à la même nomenclature.

11.2.5. indexation :

Tous les critères de recherche prévus dans l'interface utilisateur (cf maquette en partie II) devront faire l'objet d'un index.

12. LIER INFORMATIONS SECONDAIRES ET DOCUMENT

diverses approches sont possibles pour lier le document à sa description.

L'important est de choisir une solution basée sur un format international et donc facile à faire migrer vers de futurs standards, le cas échéant.

12.1. notice MARC pointant sur le document numérique

On met 'à plat', dans une notice au format US-MARC, l'ensemble des données décrivant le document d'origine, le document numérisé et les auteurs. Des champs US-MARC sont prévus pour la majorité de ces données.

12.1.1. les champs supplémentaires

Il serait intéressant d'intégrer des champs supplémentaires utiles à une recherche en ligne

- *l'adresse e-mail de l'auteur*

- *le résumé*

- *la table des matières*

  la majorité des documents est structurée. La solution préconisée est d'intégrer la table des matières à la notice.[1]

- *la bibliographie*

  la majeure partie des documents se termine par une bibliographie. Une notice MARC indique l'existence ou non d'une bibliographie, on pourrait envisager d'y intégrer à terme une bibliographie détaillée, utile à la veille scientifique.

- *lien vers le doc numérique*

  prévu en champ 856 d'US-MARC

---

[1] Argumentaire dans ce sens: voir « les tables de matières dans les catalogues en ligne » www.abf.asso.fr/html/b178_3.htm

12.2. notice SGML pointant sur le document numérique

12.2.1. DTD MARC

La Bibliothèque du Congrès a, en 1998, fourni une DTD MARC conforme aux recommandations SGML.

12.2.1.1. principe

Reprend l'intégralité des champs US-MARC, hormis les champs locaux

structure d'une notice en SGML/MARC

annexer les spécifications de la DTD MARC : 7 pages recto-verso, trouvées sur le site de la bibl du Congrés.

convertir une notice MARC en SGML/MARC

la Bibl du Congrès propose les scripts PERL ad hoc.

12.2.1.2. peut-on envisager cette solution ?

CRI : oui, elle nécessitera toutefois l'écriture d'une table de conversion pour être exploitable par Zebra.

12.2.2. DTD maison

12.2.2.1. principe

Il est toujours possible de concevoir une 'DTD maison', collant précisément aux besoisn de la BU et aux contraintes de l'environnement technique.

12.2.2.2. peut-on envisager cette solution ?

oui, mais elle nécessite des compétences XML et peut-être un investissement important en temps : définition d'une DTD, etc ...

voir si c'est justifié et raisonnable

12.3. notice bibliographique intégrée dans le document numérique

Cas de figure rencontré dans les documents numérisés au format texte

12.3.1. Dublin Core

Dublin Core = en-tête de 15 champs. Prévu pour les pages Web. Adapté ou non aux catalogues de bibliothèques numériques ? Les avis sont partagés (voir plus haut)

Pour le projet de littérature grise, nous le jugeons inadapté.

12.3.2. le TEI Header

Cas le plus fréquemment rencontré dans les bibliothèques numériques, en particulier pour le projet international Cyberthèses.

12.3.2.1. principe

le document, numérisé en mode texte structuré, inclut les informations secondaires en en-tête (alias TEI Header).

Constitué de quatre parties essentielles, cet en-tête permet de décrire en profondeur :

- le fichier informatique lui-même, contenant le texte et l'en-tête (avec la possibilité d'indiquer les responsabilités attachées à ce fichier, sa taille, les modalités de sa constitution et de sa diffusion, la ou les sources utilisées,...),
- les règles de codage et les choix éditoriaux appliqués au contenu (niveau de balisage, balises utilisées, corrections, etc.),
- les caractéristiques du texte codé : mots-clés, nature du texte (oral vs écrit, original vs traduction, genre littéraire), contexte, lieux et personnes impliqués, ...,
- l'historique du fichier, permettant de suivre les mises à jour, ajouts et corrections successives apportées au fichier.

Le corps du document est entièrement balisé (balises de présentation, balises de structure, balises sémantiques).

TEI privilégie la structure intellectuelle du document. Pour décrire un document, l'unité de base est le caractère. La notion de page physique n'existe pas.

structure d'un document structuré en TEI

<en-tête> infos secondaires (arborescence elle-même balisée) </en-tête>

<corps du document> texte balisé voire liens externes balisés </corps du document>

convertir une notice MARC en en-tête TEI

pas de lien bi-univoque : MARC est plus riche pour décrire le document imprimé, TEI est plus riche pour décrire le document numérique

12.3.2.2. Peut-on envisager cette solution ?

CRI : oui, sous réserve de ne pas utiliser les attributs au sein des balises (Zebra ne peut les exploiter)

Avantages :

- TEI est conforme aux directives XML, format d'échange reconnu
- il est possible de l'envisager avec des documents image : dans ce cas le corps du document se limite à un lien vers le fichier image
- Acquérir une 1ère compétence XML

inconvénients

- ne reprend pas tous les champs d'une notice MARC (cf site de la Bibliothèque du Congrès)

12.4. documents format image encapsulés dans le document numérique

Cas de figure rencontré pour les sous-dossiers d'archives. DTD les plus connues : EAD, EBIND

12.4.1.1. Principe

ne sont numérisés au format texte, et donc utilisables par les moteurs de recherche, que l'en-tête (notice) et les titres de dossiers. Chaque titre donne accès au dossier stocké au format image.

EAD et EBIND privilégient la structure physique du document initial. Pour décrire un document, l'unité de base est la page.

structure d'un document structuré en EAD ou EBIND

<en-tête> infos secondaires (elles-même balisées) </en-tête>

<titre du chapitre 1> lien vers l'image du chapitre 1 </titre du chapitre 1>

<titre du chapitre 2> lien vers l'image du chapitre 2 </titre du chapitre 2>

<titre du chapitre i> lien vers l'image du chapitre i </titre du chapitre i>

convertir une notice MARC en en-tête EAD ou EBIND

cela reste à étudier

12.4.1.2. Peut-on envisager cette solution ?

A priori peu adaptée dans l'immédiat. CRI non consulté sur les aspects techniques.

Avantages :

- très utile pour de gros documents que l'on pourrait découper en chapitres de taille raisonnable. Cette solution permet de mettre en valeur la structure des documents et facilite la navigation dans les textes.
- les bibliothécaires français se tournent vers EAD (cf réunion organisée par le ministère de la culture mi-sept), envisageant de l'enrichir si nécessaire (cf prochaine réunion du groupe XML de l'ABF fin sept). Pour la BU, participer aux groupes de travail = se faire reconnaître de ses pairs.

inconvénients

- dans le cas de la littérature grise, beaucoup de documents sont structurés mais peu d'entre eux ont une table des matières. Il faut donc prévoir une étape de saisie.

13. LA RESTITUTION SUR LE WEB

13.1. rappel des objectifs

Les principaux objectifs sont de

- proposer, à partir d'un catalogue, des documents d'une discipline donnée, d'un type donné

- restituer à distance et rapidement la forme imprimée, dans sa présentation initiale.
  Un autre objectif, fréquent en GED, pourrait être de proposer à l'étudiant ou au chercheur la possibilité d'annoter les documents. Précisons que ce n'est pas l'objet de ce projet. Il entre dans les missions de la bibliothèque, non de fournir des documents de travail, mais de garantir le respect et l'intégrité des œuvres mises à la disposition du public.

Il faut proposer des formats standards pour la lecture, même si les lecteurs disposent d'environnements hétérogènes.

Pour les pays distants, envisager de stocker une copie de la base sur des serveurs miroirs, préciser la périodicité de mise à jour

13.2. les différents formats de restitution

Prévoir, pour l'affichage du document primaire, les formats

- PDF compacté
- PostScript (directement pour impression)
- HTML (sélectionnable si document entièrement au format texte)
- XML (sélectionnable si document entièrement au format texte). Actuellement, seuls les navigateurs récents lisent le XML, pour les autres c'est le code source qui s'affiche, avec toutes ses balises

Prendre garde au fait que les formats images sont plus lourds à charger que les formats textes.

13.3. du serveur au client : les protocoles

13.3.1. HTTP

L'accès par le Web s'effectue par le protocole HTTP (HyperText Transfer Protocol), basé sur une architecture client-serveur : les applications nécessaires au traitement des informations renvoyées par le serveur sont intégrées au navigateur du poste client, ce qui implique l'usage d'un micro-ordinateur.

13.3.2. Telnet

Un autre moyen est d'accéder à partir d'un logiciel client Telnet (type Minitel). L'interface est alors plus pauvre : longueur des lignes limitée à l'écran, modes de recherche plus restreints.

Telnet permet la connexion à distance sur un ordinateur relié au réseau. Certains catalogues de bibliothèques et bases documentaires sont accessibles via ce service, même si ce type de connexion disparaît peu à peu au profit d'un accès par l'intermédiaire d'un serveur.

A envisager si nécessaire pour l'Afrique.

13.3.3. Z39.50

Le protocole Z39.50 est une norme américaine conçue pour permettre l'échange de notices et l'interrogation simultanée de catalogues.

Les avantages de ce protocole sont :

- présentation des services et affichage des enregistrements dans un processus transparent ,
- exploitation de la richesse des bases structurés dans un format bibliographique donné (MARC par exemple)
- une même requête peut être lancée sur plusieurs bases sans que Z 39-50 connaisse le langage de commande spécifique à chaque catalogue

- les réponses à une requête sont dans un format standard, permettant ainsi leur réutilisation comme base pour d'autres services.

Partie V - Préconisations, perspectives

14. RECOMMANDATIONS

14.1. promotion du site

14.1.1. Le faire connaître

Donner un nom au projet (acte de naissance)

faire référencer le site auprès des principaux moteurs

le promouvoir, par les listes de diffusion professionnelles, auprès des principales instances de la francophonie, chez les chercheurs, ...

puis ajouter des fonctionnalités permettant de fidéliser les usagers

14.1.2. L'enrichir

Pour la crédibilité du site, il est important d'ajouter régulièrement des documents, même en petit nombre.

Assurer un suivi des connexions. Attention à ne pas se limiter à la page d'accueil car on prendrait alors en compte les visiteurs 'accidentels'

14.1.3. Acquérir de nouveaux documents

Il ne faut pas ignorer que, même si la littérature grise existe en véritables gisements, la collecter n'est pas chose aisée. Sensibiliser les enseignants et les chercheurs à la démarche est donc essentiel à la réussite du projet.

Sur le site, la possibilité d'interaction avec les enseignants-chercheurs doit être prise en compte : les inciter à une première prise de contact est une fonctionnalité à intégrer dans le prototype (de façon simplifiée)

On peut par ailleurs mener une politique de recherche ciblée : Par exemple, rechercher la littérature grise hors de nos frontières (sur sites SYFED) ou dans certaines disciplines en priorité.

Utiliser la messagerie pour recueillir le document primaire numérisé (préciser les formats acceptables).

14.1.4. Faire évoluer le site

S'appuyer sur le retour d'expérience : soumettre le prototype à un échantillon d'enseignants/chercheurs et d'étudiants et recueillir leurs commentaires

Faire une quantification de l'audience

Prévoir des solutions qui permettent d'envisager diverses évolutions possibles

envisager l'utilisation de la signature électronique pour automatiser l'envoi par messagerie de l'autorisation de numériser et de diffuser

14.2. financement du projet

Ne pas hésiter à déposer des demandes de financement auprès

- du ministère
- de la région (bien que l'université ne soit pas dans son giron, il lui arrive de subventionner partiellement des projets universitaires)
- des divers organismes qui lancent régulièrement des appels à projet (défense du français, diffusion du savoir-faire français, ...). Voir par exemple l'appel à propositions du fonds francophone des inforoutes : http://www.francophonie.org/fonds/appel/doss6.htm

15. OPTIONS POUR LA CHAINE DOCUMENTAIRE

15.1. stockage et restitution des documents primaires

Dans un souci de cohérence, tous les documents primaires seront stockés dans un même format.

Les stocker dans un format « texte structuré » est inenvisageable :

- de nombreuses sources n'existent que sous forme imprimée (coût rédhibitoire d'un OCR complet)
- les restituer au format image est jugé suffisant

Cependant, la solution retenue doit être évolutive :

- de plus en plus de documents source seront disponibles dans une version numérique
- à terme, le besoin sera peut-être d'en restituer une vue plus souple que l'image

Les documents primaires seront si possible stockés au format PDF, en mode 'image+texte caché' (reste à voir si la taille de tels fichiers reste raisonnable).

15.2. stockage des informations secondaires

L'important pour la BU :

- garder le concept de notice bibliographique, au contenu riche et structuré
- stocker dès que possible ces métadonnées au format XML

La grande question : l'état de l'art le permet-il aujourd'hui ?

15.3. première solution envisagée : TEI

créer des documents TEI dont l'en-tête serait la notice et dont le corps se limiterait dans un premier temps à un lien vers le doc primaire (PDF)

Voir en annexe un exemple d'en-tête détaillé

Inconvénients :

- pour passer de MARC au TEI header, beaucoup des infos de la notice n'existent pas sous forme d'éléments (ie de balises)
- on pourrait les réintégrer sous forme d'attributs mais le CRI ne dispose pas d'outil capable d'exploiter ces attributs

Avantages :

- La solution TEI est conforme aux recommandations XML
- Elle permettrait de faire cohabiter
  - des documents 'rétro-convertis' qui resteraient au format PDF
  - de nouvelles acquisitions dont la version numérique serait stockée intrinséquement dans le corps TEI

15.3.1. TEI : mise en oeuvre en 3 étapes

Dans l'immédiat, le CRI ne dispose pas de moteurs d'indexation et d'exploration capables d'exploiter les ttrubuts de l'en-tête TEI. Toutefois il est fort probable que des outils adaptés à XML se développeront.

Si l'otion TEI est retenue, une solution est de passer par 3 étapes :

Etape 1 = tenir les engagements pris auprès des auteurs et affiner les besoins (vérifier que les formats de restitution sont adaptés aux réseaux africains)

Etape 2 = amorcer le passage à XML, solution portable et universelle

Etape 3 = traiter les documents reçus au format numérique. Passer au 'tout XML' ?

voir le tableau page suivante

	Nature des sources	Stockage des infos Primaires	Stockage des infos secondaires	Interface Web	Acteurs	délais	commentaires
Etape 1	Imprimées	PDF format image avec liens depuis TdM	Base ACCESS (champs = ceux du TEI Header)	D'après maquette. Idem sujets d'examens pour la partie 'consultation'	BU + Mikros (CRI en supervision seulement)	Fin 2000 au plus tard	S'appuyer sur les sujets d'examen. Peut être confié à un étudiant en informatique
Etape 2	imprimées	PDF format image avec liens depuis TdM	en-tête d'un document XML (TEI Header a priori)	Pour la partie consultation, mise en œuvre d'outils de recherche sur pages XML	BU + Mikros + forte implication du CRI	??? Selon disponibilités (humaines et logicielles) du CRI	Pour les infos secondaires, le passage ACCESS -> XML peut être automatisé si nécessaire
Etape 3	Imprimées + numériques	- PDF si sources imprimées - PDF ou XML si sources numériques	en-tête d'un document XML (TEI Header a priori)	Inchangée si solution PDF maintenue, à revoir si on passe à une solution 'tout XML' (pour les docs ad hoc)	BU surtout, CRI si 'tout XML'	Selon options retenues, mais visons fin 2001 au plus tard.	

Planning de mise en œuvre progressive vers TEI

15.4. deuxième solution envisagée : MARC SGML

envisagée tardivement, mais la plus crédible ...

principe = stocker la notice dans un format XML directement issu du format USMARC

avantages :

- chaque champ MARC devient un élément (ie balise) XML
- moyennant l'écriture de tables de conversion, une telle structure est exploitable par les outils disponibles au CRI
- selon la bibliothèque du Congrès, il est possible d'intégrer la DTD Marc dans l'en-tête d'un document TEI (en remplacement donc des balises du TEI header)
- la notice au format XML est peut-être gérable par le logiciel qui remplacera TEXTO (à préciser dans le cahier des charges au CRI)

inconvénients :

- la DTD fournie par la Bibliothèque du Congrès n'intégre pas les champs locaux, pourtant indispensables. Il est donc nécessaire de créer une DTD maison, intégrant la DTD Marc officielle.

15.5. moyens à mettre en oeuvre

Il faut prendre en compte l'ensemble des coûts de mise en œuvre pour atteindre la cible, et de fonctionnement pendant 3 ans. On estimera un coût annuel en se basant sur une prévision optimiste du nb de documents/pages à traiter ..

15.5.1. Ressources logicielles

Pour transformer en PDF les documents électroniques reçus :

Adobe Acrobat 4.0 , licence monoposte = 2 300 F

Pour traiter les documents imprimés (récents), en se basant sur un volume de 5 000 pages :

Recours à Mikros :

- Solution Table des Matières seule : 4F la page, soit un total de 20 000 F

- Solution OCR sur tout le document (à ne pas exclure à terme, car ouvre la voie à la recherche en texte intégral) : je me base sur 10F la page (à vérifier auprès de Mikros), soit un total de 50 000 F

Si l'OCR est retenu, envisager l'acquisition d'un scanner à plat (15 000 F ?)+ logiciel style 'Acrobat Capture'(6 500 F). C'est beaucoup moins complexe à mettre en œuvre et moins lourd en charge de travail que la numérisation d'archives.

Pour indexer et rechercher dans des documents XML :

Passer à la version évoluée de Zebra = 100 000 F

15.5.2. Ressources matérielles

15.5.2.1. Un poste client

Un poste dédié aux tests Acrobat, XML, ... : 15 000 F

15.5.2.2. Le serveur de données

Quid du serveur actuel ?Faudra-t-il le 'gonfler' ?

15.5.3. Ressources humaines

prévision côté BU = l'équivalent de 1,5 à 2 personnes à temps plein pour l'ensemble du projet.

Prévoir coût de la maîtrise d'œuvre, assurée par le CRI.

16. LE PROJET : EVOLUTIONS ENVISAGEABLES

16.1. le travail en partenariat

Le projet de numérisation de littérature grise couvre des enjeux variés et ambitieux :

- de vitrine pour Lille1
- de valorisation de la littérature grise
- de valorisation de la technicité des bibliothèques (webDOC de l'ABES)
- de défense de la francophonie
- de relations Nord-Sud, voire de valorisation de travaux du Sud

Dès qu'une première base sera en ligne, le SCD aura à envisager un rapprochement avec des organismes poursuivant l'un ou l'autre de ces intérêts, tant dans un souci d'harmonisation que pour fédérer les coûts. La crédibilité et la réussite du projet en dépendront.

Diffuser la littérature grise sur internet est assimilable à la gestion des thèses électroniques, pour laquelle Stefan Gradmann prône la coopération entre bibliothèques universitaires (http://www.abes.fr/ara11.htm) :

« [...] La gestion des thèses électroniques n'est pas une question technique : les solutions à appliquer (ou à trouver) seront pour la plupart des solutions Internet standard, bien qu'il puisse y avoir une valeur ajoutée de la part des bibliothèques, surtout en ce qui concerne les aspects qualitatifs de la génération des métadonnées et de l'indexation. L'enjeu de fond ici est plutôt d'ordre culturel, voire politique, en ceci qu'il concerne directement le rôle futur des bibliothèques dans les nouveaux modes de gestion et de distribution des informations : si les bibliothèques universitaires ne parvenaient pas à acquérir un rôle majeur en ce domaine, qui est un des rares secteurs de production intellectuelle non encore totalement soumis à une logique commerciale, on aurait toute raison de se poser des questions sur l'avenir de ces mêmes bibliothèques. Ce constat, ainsi que le caractère très général des solutions techniques, devrait davantage renforcer les structures coopératives : plutôt que de réinventer éternellement la roue dans des domaines techniques, les bibliothèques ont tout intérêt à profiter des outils développés en coopération afin de pouvoir concentrer leur énergie dans les domaines de l'acquisition et de la gestion des thèses électroniques. »

16.2. à Lille1, des documents primaires au format XML

Il est prévisible que la diffusion et donc la connaissance de XML iront en s'amplifiant et que les chaînes de production documentaire des éditeurs, des industriels ainsi que le monde scientifique s'orienteront vers la documentation structurée en utilisant XML. L'efficacité de ces chaînes repose en grande partie sur l'étape amont qui consiste à produire les documents et qui nécessite une parfaite maîtrise des outils et des concepts liés à la documentation structurée.

Il est prévisible que le document aura de plus en plus tendance à devenir un amalgame indissociable entre le corps du document et les métadonnées qui lui sont associées C'est l'une des conséquences majeures que provoquera à moyen terme 'émergence de XML. Ceci signifie en d'autres termes que les métiers de la documentation auront à se redéployer comme acteurs de la structuration du document au stade de sa pré-production ou de sa post-production.

Le document va se structurer et se baliser de façon de plus en plus riche. Cela en permettra la mise en relation structurée sur des réseaux beaucoup plus efficaces, mais cela demandera de prêter attention à la technique et à la réalité de sa structuration et à son balisage, en aval (et si possible en amont) de sa production; d'où la nécessité de lier les activités de la documentation à celle de la production et de l'édition du document.

Dès à présent, des projets d'édition électroniques de thèses (Cyberthèses, Silfide, ...) influent sur les pratiques documentaires des chercheurs.

16.3. revues universitaires électroniques

En abaissant les coûts de production par rapport à l'offset et en autorisant l'impression à la demande, le numérique permet à de nouveaux acteurs de se tourner vers l'édition scientifique.

Le monde universitaire saura sans doute tirer parti de cette opportunité, à l'image des américains qui réinventent les presses universitaires ; les bibliothèques ont une expertise à apporter dans cette démarche.

Citons déjà le SCD de Marne-la-Vallée qui se lance actuellement dans une première expérience d'édition électronique d'actes de colloque, assisté par la Documentation Française pour les aspects éditoriaux.

17. BIBLIOTHEQUES ET EDITION SCIENTIFIQUE

La situation qui prévaut aujourd'hui dans le monde de l'édition électronique est celle d'une grande effervescence, encore mal maîtrisée mais signe qu' une révolution a déjà commencé.

17.1. les nouveaux acteurs de l'édition scientifique

17.1.1. Etat des lieux

Que les universités manquent de moyens pour s'abonner aux revues qui publient leurs propres productions relève d'un véritable paradoxe :

- Hausse continuelle des abonnements / compressions budgétaires des bibliothèques et des centres de documentation.
- Augmentation et diversité croissante des revues scientifiques / acquisition limitée (dû aux coûts) des bibliothèques contradictoire avec leur mission de diffusion du savoir.
- Rapidité de production des résultats scientifiques / lenteur de la publication papier.
- Baisse des coûts de production pour une revue électronique / peu d'éditeurs proposant de vendre l'abonnement électronique disjoint de l'abonnement papier à des coûts nettement inférieurs.
- Discours politique sur la société de l'information / barrière économique de circuits éditoriaux commerciaux
- Facilité de transmission du savoir lié à Internet / durcissement juridique des éditeurs commerciaux
- Système lucratif de la fonction éditoriale / bénévolat des acteurs et système de subventions gouvernementales.

Des voix s'élèvent contre cet état de fait, tant chez les chercheurs que chez les bibliothécaires :

- dès 1997, B Lang de l'INRIA, remet en cause la pratique de cession des droits d'auteur pour la littérature scientifique :

  « A ce jour, l'essentiel du travail de publication est fait par les chercheurs : écriture des articles, organisation et gestion des comités éditoriaux, lecture et sélection des articles retenus. Le rôle des éditeurs est devenu dans la plupart des cas tout à fait minime, se bornant à l'impression et à la diffusion (et, parfois, un peu de présentation). Dans ce contexte, reste-t-il bien raisonnable de continuer à leur donner un contrôle exclusif de nos publications, nous interdisant par là-même une diffusion plus souple et moins chère par des moyens numériques, que nous sommes capables d'assurer nous-même, l'archivage pouvant être assuré par les divers centres de documentation et bibliothèques de la planète. En outre, cela réduirait le prix élevé de l'accès à ces informations, qui contribue à en tenir à l'écart les communautés et pays les plus pauvres, voire les PME » [1]

- J C Guédon, quant à lui, exhorte les universités à réagir face aux politiques expansionnistes d'éditeurs commerciaux comme Elsevier.

On ne peut certes pas ignorer le rôle social acquis au fil du temps par les revues commerciales : la carrière du chercheur en terme de promotion, d'honneur voire d'embauche dépend directement du prestige de la revue qui le publie.

Toutefois, le développement du numérique bouscule les habitudes prises par l'ensemble de la communauté scientifique : la propriété intellectuelle des résultats, les processus d'évaluation, les pratiques de communication des résultats de la recherche sont remis en question.

17.1.2. Publications scientifiques sur Internet

Depuis quelques années, le développement d'Internet a entraîné de grands bouleversements dans la communication scientifique. Des projets initiés dans le monde entier, dans différents secteurs de la science, ébranlent l'ordre établi dans la diffusion des résultats de la recherche et dans le processus du partage des connaissances.

---

[1] Dans « pour une politique de contrôle des droits d'accès », http://pauillac.inria.fr/~lang/ecrits/copyright/

On trouve différentes formes de presse scientifique sur Internet :

- les revues
- les pré-publications
- les rapports de laboratoires
- les lettres d'informations, les bulletins d'informations

Mais la littérature scientifique officielle, constituée des ouvrages publiés par les éditeurs et des articles de revues, n'est en général pas directement disponible sur la Toile. Les deux raisons principales en sont : le coût de la numérisation de documents imprimés déjà édités et la difficulté à faire respecter les droits de l'auteur et de l'éditeur.

De fait, il existe très peu d'ouvrages imprimés disponibles en ligne, sauf quelques exceptions comme les ouvrages anciens à caractère historique. A titre d'exemple, la Bibliothèque Nationale de France a numérisé de nombreux ouvrages du 19e siècle et en met une sélection à disposition sur son serveur Gallica[1]. On peut ainsi y consulter 33 ouvrages de physique dont : "Sur [la] diffraction de la lumière" par A. Fresnel et aussi 41 ouvrages en sciences appliquées et technologie.

En ce qui concerne les articles de revue, une évolution récente est perceptible. De nombreux éditeurs proposent à la fois une version électronique et une version papier avec des modes d'abonnement différents. De nouveaux titres sont créés uniquement sous forme électronique, disponible en ligne[2]. On comptait déjà environ 3000 revues et lettres d'information sous forme électronique en janvier 1997.

Une tradition de pré-publications s'est installée dans certains secteurs de la recherche comme en physique des hautes énergies ou en astronomie. Elle consiste à mettre à disposition de la communauté, en pratique de tout le monde, les versions des articles avant leur publication dans une revue reconnue.

A titre d'exemple, les archives de publications et pré-publications électroniques en physique sont conservées par le Los Alamos National Laboratory ; leur diffusion est relayée en France à Jussieu[3]. De même que l'on trouve un grand nombre d'informations en astrophysique par le service ADS financé par la NASA[4].

---

[1] http://gallica.bnf.fr

[2] voir à ce sujet l'analyse de Ghislaine CHARTRON, http://www.ccr.jussieu.fr/urfist/revues.htm

[3] http://xxx.lpthe.jussieu.fr

[4] http://adsabs.harvard.edu

Il s'agit là d'un mode de diffusion moins formel que la publication classique mais beaucoup plus efficace du point de vue de la dissémination.

Cette pratique de la pré-publication commence à se répandre dans d'autres communautés de chercheurs ; certains articles deviennent désormais disponibles sur les serveurs des équipes de recherche ou même dans les pages personnelles des chercheurs (on en trouve de nombreux exemples en informatique).

Plus généralement , la banalisation d'Internet conduit à une évolution de la publication et de la diffusion dans le domaine scientifique et technique. Les principales modifications observables peuvent se résumer en trois points :

- une plus grande disponibilité des documents, non seulement pour les chercheurs, mais aussi pour le grand public,
- une plus grande rapidité de diffusion,
- la mise en valeur de la littérature grise

Verra-t-on maintenant se dynamiser des structures éditoriales électroniques " académiques ", presses universitaires ou sociétés savantes ?

Trois conditions sont nécessaires au développement de ces nouveaux vecteurs de communication et d'information scientifique :

- l'adhésion des chercheurs aux pratiques de communication et d'information électroniques.
- la qualité des écrits qui y circulent.
- des mesures politiques de reconnaissance de ces nouveaux dispositifs.

17.2. et les bibliothèques dans tout cela ?

Les BU doivent se repositionner face au monde de la recherche ; l'Association des directeurs de la documentation et des bibliothèques universitaires (ADBU) a abordé cet aspect lors de sa journée d'étude de septembre 1999 consacrée au

thème « le rôle des bibliothèques par rapport aux besoins en information scientifique et technique des chercheurs »[1].

Par ailleurs, la technologie numérique redistribue les points de repère du monde documentaire, basé depuis cinq siècles sur les valeurs essentielles de l'imprimerie : matérialité du support, opposition manuscrit unique/imprimé multiple, séparation texte/image.

Dans le monde de l'édition classique, les rôles étaient bien définis :

- l'imprimeur : création du support imprimé
- l'éditeur : relecture, validation du fond et de la forme
- le libraire : valorisation, mise à disposition des ouvrages
- le bibliothécaire : référencement, description des ouvrages

Mais à une époque où les producteurs se mettent à documenter leur écrits et où les utilisateurs s'efforcent de chercher tout seuls, le bibliothécaire voit une partie de son rôle traditionnel lui échapper. Il est secoué dans ses habitudes et poussé à réussir la mutation de son savoir-faire.

Henri Hudrisier (http://www.aupelf-uref.org/initiatives/colloque/COM/COM-Hudrisier.rtf) constate un double mouvement de convergence en ce qui concerne les métiers de la création multimédia et ceux de la documentation :

- Les éditeurs de documents ont de plus en plus besoin d'avoir des compétences qui appartenaient auparavant au savoir-faire des documentalistes et des bibliothécaires. Ceci tient à ce que les contraintes de la structuration, de la normalisation tant sémantique que formelle deviennent tellement complexes qu'elles exigent l'intervention d'un professionnel au savoir documentaire relativement spécialisé.
- D'autre part, les spécialistes de l'information sont de plus en plus invités à mettre leurs collections sur réseaux ou à les éditer sous une forme directement accessible donc à les transformer en corpus de documents (voire en document encyclopédique) banalisés. Pour ce faire, ils ont besoin d'un savoir éditorial

[1] Les actes de cette journée sont consultables depuis le site http://www-sv.cict.fr/adbu/actes_et_je/je99/Polity.html

Pour la gestion documentaire, le numérique ne se traduit pas seulement par une modification du support et de la diffusion ; il déstabilise toute la fonction de médiation du spécialiste de l'information entre un producteur et un utilisateur car les besoins documentaires sont relatifs aux nouvelles possibilités que leur offre directement la technologie.

Le spécialiste de l'information a toujours un rôle à jouer, tant par ses connaissances des ressources documentaires, pour lesquelles internet n'est qu'un nouveau support de diffusion, que par son expertise dans le traitement intellectuel des documents.

Partie VI - conclusion

On peut considérer que la littérature grise, c'est-à-dire l'ensemble des documents imprimés sans référence et tirés jusqu'à présent à un petit nombre d'exemplaires comme les thèses, les rapports de recherche, les études, les notes, bénéficiera avec Internet de la possibilité d'une diffusion accrue au moindre coût.

Il est opportun pour les bibliothèques universitaires de chercher à acquérir un rôle majeur en ce domaine, l'un des rares secteurs de production intellectuelle non encore totalement soumis à une logique commerciale.

Elles n'y parviendront qu'au sein de structures coopératives. C'est pourquoi la réussite du projet de Lille 1 sera conditionnée par l'adoption de normes internationales communes, tant pour pour la représentation, l'indexation et la communication de l'information qu'en ce qui concerne les protocoles d'accès et de communication :

Le SCD-BU prévoit de réaliser un prototype dans le courant de l'année 2000. L'étude poursuivie durant le stage a abouti aux choix suivants :

- Pour le stockage des documents primaires, le format PDF d'Adobe (standard de fait) sera retenu dans un premier temps.
- Pour l'information secondaire, traduction en XML de notices US-MARC enrichies.
- La communication avec le web se fera a priori selon le protocole Z39.50

Outre ces aspects techniques, l'étude a mis en évidence les difficultés inhérentes au projet :

D'une part, les questions juridiques autour des droits d'auteur sont loin d'être résolues pour les documents numériques. Il conviendra donc d'être extrêmement vigilants lors de la collecte.

D'autre part, il ne sera pas simple pour le SCD-BU de se positionner en médiateur reconnu par les chercheurs. La réalisation du prototype devra être accompagnée d'une réelle démarche auprès de la communauté scientifique.

Partie VII - bibliographie

CELOG. *Code commenté de la propriété intellectuelle*. [en ligne] [consulté le 19/07/2000]. Adresse URL : http://www.celog.fr/cpi/sommaires/livre_1.htm

TREAN, Claire. Le français, « langue de contre-pouvoir » pour M Jospin, *Le Monde*, 2000, dimanche 23 juillet, p 3

Documents numériques

CHABIN, Marie-Anne. Exigences numériques et besoins documentaires. *Revue Solaris* [en ligne]. 2000, n° 6 [consulté le 06/04/2000]. Adresse URL : http://www.info.unicaen.fr/bnum/jelec/Solaris/d06/6chabin.html

DESS SID de l'université de Lille 3. *Les enjeux du management de l'information dans les organisations : usages, outils, techniques*. Paris : ADBS, 1999. 161 p.

HAIGH, Susan. Glossaire des normes, des protocoles et des formats liés à la bibliothèque numérique. *Flash réseau, n° 54* [en ligne]. Bibliothèque Nationale du Canada. [modifié le 6 mai 1998]. Adresse URL : http://www.nlc-bnc.ca/pubs/netnotes/fnotes54.htm

JACQUESSON, Alain et RIVIER, Alexis. *Bibliothèques et documents numériques*. Paris : Editions du cercle de la librairie, 1999. 377 p.

SALSA, Patrick. *Réaliser une bibliothèque numérique* [en ligne]. [consulté le 27/06/2000]. Adresse URL : http://membres.tripod.fr/salsa/index.htm

TEASDALE, Guy. Incursion dans le format image . *Lettre du bibliothécaire québécois* [en ligne]. 1999, n° 16 [consulté le 04/04/2000]. Adresse URL : http://www.sciencepresse.qc.ca/lbq/lbq16.5.html

XML

ALIS. *Langage de balisage extensible (XML) 1.0 : traduction française de la recommandation du W3C* [en ligne]. [modifié le 10 février 1998]. Adresse URL : http://babel.alis.com/web_ml/xml/REC-xml.fr.html

CLARK, James. *World Wide Web Consortium : Comparison of SGML and XML* [en ligne]. [modifié le 15 décembre 1997]. Adresse URL : http://www.w3.org/TR/NOTE-sgml-xml-971215

DUCLOY, Jacques. *Supports pédagogiques* [en ligne]. Paris : URFIST, 2000 [consulté le 07 avril 2000]. Adresse URL : http://loria.fr/~ducloy/COURS/URFIST.html

ELLIOTTE, Rusty Harold. *XML le guide de l'utilisateur*. Paris : OsmanEyrolles Multimedia, 2000. 889 p.

HUDRISIER Henri, ROMARY Laurent. L'évolution des métiers et des formations dans les nouvelles méthodes de production des connaissances : le document numérique normalisé XML, TEI, Unicode. *Colloque INITIATIVES'99*, Edmunston, 1999 [en ligne]. [consulté le 26/03/2000]. Adresse URL : http://www.aupelf-uref.org/initiatives/colloque/COM/COM-Hudrisier.rtf

Métadonnées

GRADMANN, Stefan. Catalogage et métadonnées : du vin vieux dans des bouteilles neuves ?. *64ème conférence IFLA, août 1998* [en ligne]. [consulté le 06/04/2000]. Adresse URL : http://ifla.inist.fr/IV/ifla64/007-126f.htm

VERCOUSTRE, Anne-Marie. *Eléments de métadonnées du Dublin Core, version 1.1: description de Référence*. Versailles : INRIA, 2000. [modifié le 20 avril 2000]. Adresse URL : http://www-rocq.inria.fr/~vercoust/METADATA/DC-fr.1.1.html

W3C. *Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendation* [en ligne]. [modifié le 22 février 1999]. Adresse URL : http://www.w3.org/TR/REC-rdf-syntax/

Publications scientifiques

BLONDEL, François-Marie. *Internet, science, technologie et éducation.* [en ligne] Versailles : CRDP, 1998. [consulté le 13/04/2000]. Adresse URL : http://www.inrp.fr/Acces/Biogeo/univete/fmb.htm

CHARTRON, Ghislaine. *Revues scientifiques et Internet* [en ligne]. Paris : URFIST, 1998. [visité le 15/02/2000]. Adresse URL : http://www-scd-ulp.u-strasbg.fr/urfist/revues_sur_internet/revues-revues.htm

GRADMANN, Stefan. La gestion des thèses électroniques [en ligne]. *revue Arabesques*, septembre 1998. [visité le 15/02/2000]. Adresse URL :]http://www.abes.fr/ara11.htm

DIDIER, Bruno. Services documentaires et recherche scientifique : métiers et relations en mutation. *Micro Bulletin Thématique*, avril 1999.

Partie VIII - annexes

Annexe 1

exemple d'en-tête TEI

```xml
<?xml version="1.0" encoding="UTF-8"?>
<!-- edited with XML Spy v3.0 (http://www.xmlspy.com) by Marie-France
Claerebout (private) -->
<?xml-stylesheet type="text/css" href="chamm1.css"?>
<tei.2>
   <teiHeader>
      <fileDesc>
         <titleStmt>
            <title type="main">Identification de contraintes... </title>
            <author>
               <name type="nom">Nom1</name>
               <name type="prenom">Prenom1</name>
            </author>
            <author>
               <name type="nom">Nom2</name>
               <name type="prenom">Prenom2</name>
            </author>
            <respStmt>
               <resp>numerisation par</resp>
               <name>MIKROS</name>
            </respStmt>
            <respStmt>
               <resp>codage TEI/XML par</resp>
               <name>Claerebout, MF</name>
            </respStmt>
         </titleStmt>
         <extent> 25 pages PDF, 1183 Kilo-octets </extent>
         <publicationStmt>
            <publisher>BUSTL</publisher>
            <distributor>CRI</distributor>
            <idno>F1999-55</idno>
            <availability status="FREE">
               <p>ce document est libre de droits</p>
            </availability>
            <date>mai 2000 (date de mise a disposition du fichier)</date>
         </publicationStmt>
         <notesStmt>
            <note type="uniform_title"/>
            <note type="typedoc">Rapport de recherche </note>
            <note type="url">http://www...  /F1999-55.pdf</note>
            <note type="fac"> nom de l'ecole ou de la fac </note>
            <note type="dept">nom du labo</note>
         </notesStmt>
         <sourceDesc>
            <p>BUSTL: M/F1999-55</p>
         </sourceDesc>
      </fileDesc>
      <profileDesc>
         <creation>
            <date value="date0086s">date au format jj/mm/ssaa</date>
            <name type="pays00815-17">ci le code pays sur 3 caract</name>
         </creation>
         <langusage>
            <language id="FR">francais</language>
         </langusage>
      </profileDesc>
   </teiHeader>
   <text>
      <front/>
      <!--la sous-section BODY contient le corps du document. -->
      <!--Dans un premier temps, il s'agit du lien vers l'image PDF.-->
      <body>
         <div1 id="TH.0.1" type="chapter" n="1">
            <head>lien vers le document numerise</head>
            <p>
               document PDF (890K) telechargeable page par page :
               <xref>F1999-55_01.pdf</xref>
            </p>
         </div1>
      </body>
   </text>
</tei.2>
```

Annexe 2

modèle de notice MARC XML

```
<mrcbfile>
  <mrcb>
    <mrcbldr-bd>
      <mrcb008>champ fixe, a detailler</mrcb008>
    </mrcbldr-bd>
    <mrcb-control-fields>champs 001 a 009</mrcb-control-fields>
    <mrcb-numbers-and-codes>
      <!--champs  010 a 091, hors champs locaux-->
      <mrcb020>
        <mrcb020-a>ISBN</mrcb020-a>
      </mrcb020>
      <mrcb022>
        <mrcb022-a>ISSN</mrcb022-a>
      </mrcb022>
      <mrcb041>
        <mrcb041-a>langue du document</mrcb041-a>
      </mrcb041>
      <mrcb049-bustl>
        <mrcb049-bustl-a>numero d'inventaire (utile, avec 949 ?) </mrcb049-bustl-a>
      </mrcb049-bustl>
      <mrcb072>
        <!--voir si on peut y stocker le type de document-->
        <mrcb072-a> "code categorie sujet" </mrcb072-a>
      </mrcb072>
      <mrcb092-bustl>
        <!--repetable - pas de 092 dans la DTD - -->
        <mrcb092-bustl-a>indice Dewey local</mrcb092-bustl-a>
      </mrcb092-bustl>
      <mrcb099-bustl>
        <mrcb099-bustl-a>categorie Astropar</mrcb099-bustl-a>
      </mrcb099-bustl>
    </mrcb-numbers-and-codes>
    <mrcb-main-entry>
      <!--champs 100 a 130-->
      <mrcb100>
        <mrcb100-a>auteur pers physique</mrcb100-a>
      </mrcb100>
      <mrcb110>
        <mrcb110-a>auteur collectivite</mrcb110-a>
      </mrcb110>
      <mrcb111>
        <mrcb111-a>congres</mrcb111-a>
      </mrcb111>
    </mrcb-main-entry>
```

```
<mrcb-title-and-title-related>
  <!--champs 210 a 247-->
  <mrcb245 i1="i1-0" i2="nb de caract. non significatifs (i2-0 a i2-9)">
    <mrcb245-a>titre</mrcb245-a>
    <mrcb245-b>sous-titre</mrcb245-b>
    <mrcb245-c>auteurs</mrcb245-c>
  </mrcb245>
</mrcb-title-and-title-related>
<mrcb-edition-imprint-etc>
  <!--champs 250 a 270-->
  <mrcb250>
    <mrcb250-a>numero d'edition ou de revision</mrcb250-a>
  </mrcb250>
  <mrcb260>
    <mrcb260-a>lieu de production (ville)</mrcb260-a>
    <mrcb260-b>nom du producteur (etablissement)</mrcb260-b>
    <mrcb260-c>date de production (annee)</mrcb260-c>
  </mrcb260>
</mrcb-edition-imprint-etc>
<mrcb-physical-description>
  <!--champs 300 a 362-->
  <mrcb300>
    <mrcb300-a>nombre de pages</mrcb300-a>
    <mrcb300-b>mention d'illustration</mrcb300-b>
    <mrcb300-c>format</mrcb300-c>
  </mrcb300>
</mrcb-physical-description>
<mrcb-series-statement>
  <!--champs 400 a 490-->
  <mrcb490>
    <mrcb490-a>titre de la collection</mrcb490-a>
    <mrcb490-v>numero dans la collection</mrcb490-v>
    <mrcb490-x>ISSN eventuel</mrcb490-x>
  </mrcb490>
</mrcb-series-statement>
<mrcb-notes>
  <!--champs 500 a 590-->
  <mrcb500>
    <mrcb500-a>texte de note locale</mrcb500-a>
  </mrcb500>
  <mrcb502>
    <mrcb502-a>note de these (nature,discipline,univ,date)</mrcb502-a>
  </mrcb502>
  <mrcb504>
```

```
    <mrcb504-a>bibliographie</mrcb504-a>
  </mrcb504>
  <mrcb505>
    <mrcb505-a>note de depouillement</mrcb505-a>
  </mrcb505>
  <mrcb520>
    <mrcb520-a>nature de la note mrcb-ci-520 (resume, intro,...)</mrcb520-a>
  </mrcb520>
  <mrcb-ci-520>
    <mrcb520-a>note descriptive</mrcb520-a>
  </mrcb-ci-520>
</mrcb-notes>
<mrcb-subject-access>
  <!--champs 600 a 658-->
  <mrcb690-bustl>
    <!--mot-cle (element repetable)-->
    <mrcb690-bustl-a>sujet local</mrcb690-bustl-a>
  </mrcb690-bustl>
</mrcb-subject-access>
<mrcb-added-entry>
  <!--champs 700 a 755-->
  <mrcb700>
    <mrcb700-a>nom, prenom d'auteur secondaire</mrcb700-a>
    <mrcb700-e>fonction abregee (dir.,ed.,etc)</mrcb700-e>
  </mrcb700>
</mrcb-added-entry>
<mrcb-linking-entry>
  <!--champs 760 a 787-->
  <mrcb760>
    <mrcb760-a>titre serie (si catalogage d'1 sous-serie)</mrcb760-a>
  </mrcb760>
  <mrcb773>
    <mrcb773-a>publication hote</mrcb773-a>
  </mrcb773>
</mrcb-linking-entry>
<mrcb-series-added-entry>champs 800 a 840</mrcb-series-added-entry>
<mrcb-holdings-notes>champs 841 a 845</mrcb-holdings-notes>
<mrcb-location>
  <!--champs 850 a 852-->
  <mrcb850>
    <mrcb850-a>code de la BU (si CCO)</mrcb850-a>
  </mrcb850>
</mrcb-location>
<mrcb-captions-and-patterns>champs 853 a 855</mrcb-captions-and-patterns>
```

```
    <mrcb-access>
      <!--champ 856 : adresse electronique - acces-->
      <mrcb856 i1="mode d'acces" i2="i2-0">
        <mrcb856-a>nom du serveur</mrcb856-a>
        <mrcb856-b>numero d'acces (adresse IP)</mrcb856-b>
        <mrcb856-c>eventuel pg de decompression</mrcb856-c>
        <mrcb856-d>chemin d'acces</mrcb856-d>
        <mrcb856-f>nom du fichier</mrcb856-f>
        <mrcb856-h>processeur de demande</mrcb856-h>
        <mrcb856-i>Instructions</mrcb856-i>
        <mrcb856-j>vitesse de transmission (en bps)</mrcb856-j>
        <mrcb856-k>mot de passe</mrcb856-k>
        <mrcb856-l>logon</mrcb856-l>
        <mrcb856-m>personne contact pour assistance</mrcb856-m>
        <mrcb856-n>nom en clair du serveur cite en -a</mrcb856-n>
        <mrcb856-o>systeme d'exploitation</mrcb856-o>
        <mrcb856-p>port</mrcb856-p>
        <mrcb856-q>format de fichier (d'ou mode de transfert)</mrcb856-q>
        <mrcb856-r>parametres</mrcb856-r>
        <mrcb856-s>taille du fichier</mrcb856-s>
        <mrcb856-t>emulation de terminal</mrcb856-t>
        <mrcb856-u>URI</mrcb856-u>
        <mrcb856-v>horaires d'acces</mrcb856-v>
        <mrcb856-x>note cachee</mrcb856-x>
        <mrcb856-z>note affichable</mrcb856-z>
        <mrcb856-2>mode d'acces (WAIS,...)</mrcb856-2>
      </mrcb856>
    </mrcb-access>
    <mrcb-enumeration-and-chron>champs 863 a 865</mrcb-enumeration-and-chron>
    <mrcb-textual-holdings>champs 866 a 868</mrcb-textual-holdings>
    <mrcb-variant-names>champs 870 a 873</mrcb-variant-names>
    <mrcb-item-information>champs 876 a 878</mrcb-item-information>
    <mrcb-linkages>champs 880 a 886</mrcb-linkages>
    <mrcb-bustl>
      <!--champs 9xx   (locaux et donc absents de la DTD de base)-->
      <mrcb920-bustl>date de creation notice</mrcb920-bustl>
      <mrcb930-bustl>date corr Bib (?)</mrcb930-bustl>
      <mrcb949-bustl>cote locale (si CCO)</mrcb949-bustl>
    </mrcb-bustl>
  </mrcb>
</mrcbfile>
```

Annexe 3

quelques sites de littérature grise

IRCAM

LE PROJET	
organisme	médiathèque de l'IRCAM, centre G Pompidou
public ciblé	
soutien, financement	
coopération	
DOCUMENTS PRIMAIRES	
typologie	Livres, Articles, Cédérom, Revues, Archives sonores Ircam, Disques, Mémoires thèses, Partitions, Vidéo
disciplines	musique et disciplines s'y rattachant
volume	3 DEA, 2 thèses, 1 maîtrise
langue	
DOCUMENTS NUMERIQUES	
format de stockage	stockage sous DORIS, interface DORIS-WEB
langage de balisage	
métadonnées	
plan de classement	visible sous http://varese.ircam.fr/catalogue/plan-de-classement : eg 18 = littérature grise (180.1=stage DEA, 181=thèse, 182=mémoire maîtrise)
INTERFACE WEB	
adresse URL	http://varese.ircam.fr/catalogue/index.html
consultation et recherche	recherche détaillée sur notice (apparemmment générée par DORIS-WEB)
critères de recherche	type de document
navigation proposée	accès à liste puis à notice détaillée
notice à l'écran	
formats de restitution	Aucun des documents n'est en ligne (contiennent tous du son)
autres services proposés	
info sur les droits ?	
REMARQUES	

Fourier Prépublications

LE PROJET	
organisme	Institut Fourier de Grenoble
public ciblé	accessible à tous publics
soutien, financement coopération	
DOCUMENTS PRIMAIRES	
typologie	prépublications de l'Institut
disciplines	Mathématiques
volume	500 titres (de 1984 à 2000)
langue	Français et Anglais
DOCUMENTS NUMERIQUES	
format de stockage	
langage de balisage métadonnées	
plan de classement	le code interne est un numéro d'ordre chronologique (d'enreg ?, de publication ?)
INTERFACE WEB	
adresse URL	http://www-fourier.ujf-grenoble.fr/PREP/prep_if.html
consultation et recherche	accès par année (ordre chrono inverse) ou par auteur
critères de recherche	par année : affiche tous les titres depuis la date sélectionnée (année en cours par défaut)
navigation proposée	pour l'année : affiche la liste des prépublications (auteurs, titre, code interne,lien sur notice, lien sur texte PS) clic sur l'auteur -> liste de ses publications (auteurs,titre, code interne, année, lien sur notice, lien sur texte PS) clic sur lien 'résumé' -> accès à la notice (titre,auteurs,date, résumé, classification,mots-clés, lien sur texte PS, retour liste) clic sur lien "texte' -> téléchargement au format PS
formats de restitution	
autres services proposés	propose un contact avec l'auteur (par messagerie ?)
info sur les droits ?	
REMARQUES	
	plan de classement à étudier

Michigan

LE PROJET	The University of Michigan Dissertation and Thesis Library
organisme	Université du Michigan (USA)
public ciblé	
soutien, financement	
coopération	appartient à la bibliothèque numérique nationale de T&D, initiée par le département de l'éducation des USA
DOCUMENTS PRIMAIRES	
typologie	mémoires de l'université (était prévu l'ajout de thèses en 1998)
disciplines	économie, histoire de l'art, anglais
volume	5 documents de 1996-1997
langue	anglais
DOCUMENTS NUMERIQUES	
format de stockage	texte intégral
langage de balisage	SGML/TEI
métadonnées	à voir ds le source
plan de classement	
INTERFACE WEB	
adresse URL	http://dns.hti.umich.edu/misc/diss.example/
consultation et recherche	
critères de recherche	mot-clé, auteur, année
navigation proposée	sur mot-clé : recherche en texte intégral, restitution des documents/chapitres/extraits, auteur, année, liens vers doc SGML voire HTML (voir http://dns.hti.umich.edu/misc/diss.example/kwic.html) sur auteur : renvoie titre, auteur, année, résumé, liens vers doc SGML voire HTML pour tous documents pertinents sur année : renvoie titre, auteur, année, résumé, liens vers doc SGML voire HTML pour tous documents pertinents
autres services proposés	
info sur les droits ?	
REMARQUES	
	Aucune notion du nb de pages. Le téléchargement d'un document (sur St-Riquier) au format SGML/TEI prend 1/2 heure environ ... !! Voir passage au XML sous http://www.umdl.umich.edu/um_diss_study.html

Montréal

LE PROJET	**Prototype pour la diffusion électronique des thèses, lancé en février 1999**
organisme	Presses de l'Université de Montréal
public ciblé	
soutien, financement	
coopération	universités Condordia (USA), Laval (Québec), Lyon2(France), Senghor(Egypte)
DOCUMENTS PRIMAIRES	
typologie	thèses de doctorat de l'Université de Montréal
disciplines	
volume	prévu pour l'an 2000 : un système de publication en série qui permettra le traitement d'une année de production, soit environ 350 thèses.
langue	
DOCUMENTS NUMERIQUES	
format de stockage	
langage de balisage	Le système d'information développé pour ce projet est basé sur le format normalisé SGML (ISO 8879 : 1986). Les caractéristiques de ce format permettent de répondre aux besoins d'archivage, de diffusion, de recherche et de réutilisation des données. (INTIF)
métadonnées	
plan de classement	
INTERFACE WEB	
adresse URL	http://www.pum.umontreal.ca/theses/
consultation et recherche	
critères de recherche	
navigation proposée	
notice à l'écran	
formats de restitution	
autres services proposés	
info sur les droits ?	
REMARQUES	
	L'objectif à moyen terme de ce projet est la mise en réseau d'universités ayant adoptées un système compatible de traitement et de diffusion électroniques des thèses.

PELLEAS

LE PROJET	
organisme	SCD de Marne-la-Vallée
public ciblé	tous les postes du Campus
soutien, financement	
coopération	
DOCUMENTS PRIMAIRES	
typologie	documents produits par l'Université, de la revue scientifique au support de cours, de la thèse au didacticiel pour l'autoformation
disciplines	
volume	
langue	
DOCUMENTS NUMERIQUES	
format de stockage	le plus riche possible en prévision d'une réexploitation future
langage de balisage	normalisé (références à la DTD ISO 12083/84 et à la TEI)
métadonnées	
plan de classement	
INTERFACE WEB	
adresse URL	
consultation et recherche	
critères de recherche	
navigation proposée	
notice à l'écran	
formats de restitution	
autres services proposés	
info sur les droits ?	
REMARQUES	

THESAURIA

LE PROJET	THESAURIA
organisme	INRIA (recherche en Informatique et Automatique)
public ciblé	
soutien, financement	"Nous participons à l'action GRISELI (constitution d'une base de données de la littérature grise scientifique française)" [http://www.inria.fr/Ra-Tech96/moyens-info/node49.html]
coopération	
DOCUMENTS PRIMAIRES	
typologie	essentiellement des thèses et HDR, quelques rapports de l'INRIA
disciplines	informatique
volume	
langue	anglais et français
DOCUMENTS NUMERIQUES	
format de stockage	il est possible de consulter les signalements des documents parus au cours d'un mois, de faire une recherche par mots du titre ou nom d'auteur sur l'ensemble de la base. A partir de la notice complète avec résumé, on accède au texte intégral, stocké sur les ordinateurs des différents laboratoires.
langage de balisage	
métadonnées	
plan de classement	
INTERFACE WEB	
adresse URL	http://www.inria.fr/Griseli/griseli.html
consultation et recherche	recherche libre ou par mois
critères de recherche	donne accès à une liste minimale : réf interne, auteurs, et titre, lui-même lien vers la notice détaillée
notice à l'écran	titre en 2 langues nature (thèse, HDR, rapport, ??) date auteur, avec lien sur liste de l'ens de ses travaux organisme, avec lien vers le site de l'organisme résumé en 2 langues mots-clés en 2 langues lien vers document primaire : format ps ou format papier à commander par mail (gratuit)
formats de restitution	Chaque trimestre, une version papier de Thesauria à destination des PED est réalisée à partir de la base Texto. Les fichiers postscript correspondants sont également disponibles sur le service web.
autres services proposés	
info sur les droits ?	
REMARQUES	visité le 27/04/2000