La numérisation des documents au format pdf et la réalisation d'une bibliothèque électronique avec le logiciel Greenstone : la bibliothèque d'information scientifique et technique du Burkina Faso

Pier Luigi Rossi
IRD (Institut de recherche pour le développement), Secteur Documentation, Bondy, France
rossi@ird.fr

Résumé

Dans l'objectif de réaliser un transfert de compétences en matière de numérisation et de mise an accès des productions scientifiques vis à vis des institutions des pays en développement, l'IRD à mis en œuvre en 2004 un projet de création d'une bibliothèque électronique en partenariat au Burkina Faso. Cette bibliothèque électronique donne accès à plus de 120.000 pages de documents scientifiques produits par treize institutions présentes dans ce pays. Le site web a été conçu en utilisant le logiciel libre Greenstone.

Introduction

A partir du milieu des années '90 les évolutions des technologies de stockage sur disque, du matériel de numérisation et de l'Internet ont permis d'envisager la mise en accès de fonds de bibliothèques en texte intégral sur le web.
Dans ce contexte, en 1996, l'IRD a commencé la numérisation de son fonds documentaire patrimonial qui est constitué par les productions scientifiques de ses agents. Actuellement sont accessibles sur le site Internet de l'IRD environ 35.000 documents dans leur intégralité (environ 1.000.000 de pages), au format pdf, avec une recherche qui associé le texte des documents (l'ensemble de leur contenu) et les principaux champs documentaires (auteur, titre, descripteurs, date).
A partir de cette réalisation nous avons acquis, par une pratique sur le terrain, une maîtrise de la numérisation des documents scientifiques, de la transposition vers le format pdf de fichiers informatiques et de la création de bibliothèques électroniques.
Compte tenu des missions de notre Institut, notamment celles concernant le partenariat et la valorisation des activités et des productions scientifiques des Institutions des pays en développement, nous avons mis en œuvre un projet de réalisation d'une bibliothèque électronique de référence au Burkina Faso. Elle donne accès à un large ensemble de documents produit par des Institutions scientifiques présentes dans ce pays. Au cours de sa réalisation, sur une durée de 8 mois, nous nous sommes également employés à réaliser un transfert de compétences par la formation de nombreux collègues burkinabé.

La mise en place du projet et des ateliers de numérisation

Au cours de la mise en place du projet, nous avons proposé à un large éventail d'institutions de réaliser la numérisation d'environ 10.000 à 15.000 pages de documents relevant de leurs productions scientifiques. Nous avions informé les partenaires que le traitement des documents était pris en charge par nos personnels et par nos équipements, dans deux ateliers de numérisation localisés au CNRST (Centre national de la recherche scientifique et technique) et à l'EIER/ETSHER

(École d'ingénieurs de l'équipement rural / École des techniciens supérieurs de l'hydraulique et de l'équipement rural). Compte tenu des moyens et du temps dont nous disposions, nous avions prévu de traiter 100.000 pages.
Nous avions informé les partenaires que les documents électroniques résultant de la numérisation restaient la propriété de chaque Institution qui en disposerait de plein gré. Cependant, chaque institution devait donner son accord pour que les documents numérisés dans le cadre du projet soient également accessibles librement sur le site Internet constituant la bibliothèque électronique de la recherche scientifique du Burkina Faso.
Les institutions qui ont adhéré au projet sont : le Bureau national des sols, le Centre d'études et de recherches en lettres et sciences humaines et sociales, le Centre international de recherche-développement sur l'élevage en zone subhumide, le Centre national de la recherche scientifique et technique, le Centre régional pour l'eau potable et l'assainissement, le Département de géographie de l'Université de Ouagadougou, l'École d'ingénieurs de l'équipement rural / École des techniciens supérieurs de l'hydraulique et de l'équipement rural, Global water partnership / West Africa, le Ministère des infrastructures des transports et de l'habitat, les Presses universitaires de Ouagadougou, le Programme national de gestion des terroirs, l'Union économique et monétaire ouest africaine, l'Unité d'études et de recherches démographiques.

Nous avons consacré le démarrage du projet à la mise en place d'un atelier de numérisation à de la Direction de l'information scientifique et technique du CNRST. Au cours d'une période de trois semaines nous avons formé une équipe de cinq personne à la numérisation et à la production de documents électroniques et nous avons numérisé environ 10.000 pages : notamment la revue Eureka et la revue Sciences et Techniques.
En parallèle à cette activité, nous avons mis en place un atelier de numérisation au service des institutions ayant décidé de fournir leurs documents pour alimenter la bibliothèque électronique. Cet atelier de numérisation, qui a fonctionné environ 6 mois au cours du projet, était localisé dans un bureau de l'EIER.
Son fonctionnement était assuré par deux techniciens que nous avons formé et encadré au cours du déroulement du projet[1].
Pour assurer les traitements de numérisation nous disposions d'un scanner de production (Fujitsu FI 4120C), d'un scanner à plat A3 (Mustek Scanexpress A3 SP), de trois micro ordinateurs (1 sous XP et deux sous Windows 98), des logiciels Acrobat et Photoshop.

L'organisation de l'atelier et les procédures de numérisation

La configuration de l'atelier qui était au service de la numérisation des documents des institutions ayant adhéré au projet était basée sur la présence de deux personnes, de deux scanners et de trois postes de travail. Cette configuration a permis de structurer les différentes taches et d'optimiser les temps de traitement. Par exemple, la reconnaissance optique des caractères, qui implique des longs temps de traitement, a été réalisée sur un poste de travail surtout en dehors des plages de travail.

[1] Trois mois après le démarrage du projet, l'encadrement a été assuré à distance et ponctuellement sur place lors de deux missions d'une semaine chacune.

L'organisation du travail sur plusieurs postes avec une structuration des tâches a permis aux deux techniciens de varier leurs fonctions et de prendre en charge l'ensemble des opérations de traitement. Sur une période effective de cinq mois de travail nous avons pu traiter 120.000 pages, soit 20 % de plus de ce que nous avions prévu pour ce projet.
Le fonctionnement et l'organisation du travail d'un atelier de numérisation s'articulent donc en plusieurs phases. Il est très utile que les personnels chargés de l'atelier puissent les assurer à tour de rôle et que une bonne complémentarité soit acquise. Certaines tâches sont parfois répétitives, d'autres demandent une grande attention ou une bonne technicité.
Les principales phase concernant la numérisation des documents sont, schématiquement, les suivantes :

La préparation des documents

La première démarche pour le fonctionnement de l'atelier, consiste dans la récupération des documents à numériser. Pratiquement, nous nous rendions auprès des institutions ou nous réceptionnions les documents à l'atelier.
La préparation des documents consiste principalement à constituer des ensembles de feuilles séparées, puisqu'un scanner de production numérise les pages en les faisant glisser devant ses capteurs.
Nous avons assuré le massicotage des documents à reliure rigide ou le désassemblage des reliures légères. Une vérification manuelle de l'efficacité des ces opérations est nécessaire pour obtenir des dossiers avec des feuilles entièrement déliassées.

La numérisation

Le scanner Fujitsu FI 4120C

Les traitements de numérisation ont été réalisés avec un scanner de production Fujitsu FI 4120C. Il traite environ 20 pages à la minute, le format maximum possible est le A4. Le scanner fonctionne avec son logiciel spécifique qui est lui-même piloté par le logiciel Acrobat.
La numérisation est réalisée avec une définition de 300 dpi[2], en noir et blanc, avec calibrage de la luminosité et du contraste. En fonction de la nature des documents, ils sont traités en recto seul ou en en mode recto verso. Le scanner peut traiter des lots de 50 feuilles et le technicien doit donc l'alimenter au fur et à mesure.

La reconnaissance optique de caractères

La reconnaissance optique des caractères a été assurée par les modulés spécifiques du logiciel Acrobat. Puisque les temps de traitement sont importants, cette opération est souvent effectuée la nuit ou en fin de semaine sur des lots de documents.
La finalité des opérations de numérisation est d'obtenir à l'écran une représentation des documents conforme à l'original. Pour cette raison, les traitements de reconnaissance optique des caractères sont paramétrés pour produire une image qui

[2] Dpi = points par pouce. 1 pouce = 2,54 cm. 300 dpi est la définition qui donne une bonne qualité de l'image notamment dans la perspective de réaliser des traitements de reconnaissance optique de caractères.

est associée au texte qui a été reconnu. L'image générée par la numérisation est donc conservée et s'affiche systématiquement à l'écran. Le texte reconnu, sur lequel aucun correction n'est effectuée, n'apparaît pas à l'écran : il sert à effectuer la recherche de texte dans le document et peut être exporté par l'utilisateur final pour qu'il puisse effectuer des traitements spécifiques.

La vérification et finalisation des documents électroniques

La vérification des documents après les traitements que nous venons de décrire sert: :

- à constater la présence de toutes les pages (certaines pourraient passer en double dans le scanner ou manquer dans le cas d'un document défectueux),
- à constater l'orientation des pages notamment après le traitement de reconnaissance optique des caractères (les pages sont orientées par le logiciel selon le sens de la lecture, mais cette opération peut échouer pour certaines pages),
- à réaliser la numérisation des pages en couleur et en nuances de gris : pour garder l'intégrité et la représentation du document conforme à l'original, ces pages sont numérisées en fonction de leur contenu en couler (graphiques, illustrations, photos) ou en nuances de gris (photos) avec des procédures spécifiques de compression des images,
- à réaliser l'assemblage et le désassemblage des documents : pour la reconnaissance optique des caractères on traite des lots de documents et, après traitement, il faut désassembler les lots pour constituer les documents définitifs. Il en est de même pour les pages en couleur, en nuances de gris, les couvertures, les pages hors format[3] qui doivent être réinsérées dans le document final,
- à assurer la saisie des méta données (auteur, titre, date de publication) pour chaque fichier électronique.

Le reconditionnement des documents

Les documents qui ont servi à la numérisation qui ne constituent pas des exemplaires en surnombre (éliminés après leur massicotage) doivent être remis en état avec une reliure légère ou un encollage au dos. Lors du déroulement du projet, nous avons assuré ces opérations avec l'aide des personnels de l'atelier de reprographie de l'EIER.

[3] Les carte géographiques et thématiques sont souvent des documents qui dépassent le format A3. Nous avons assuré la numérisation de ces documents sur le site IRD de Bondy en France ou nous disposons d'un scanner A0.

La bibliothèque électronique conçue avec le logiciel Greenstone

Pour la création du site de la bibliothèque d'information scientifique et technique du Burkina Faso nous avons utilisé le logiciel Greenstone.

« Greenstone est un ensemble de logiciels dont le but est de donner accès à des collections d'informations constituant une bibliothèque numérique. Greenstone fait partie du Projet de Bibliothèque Numérique de Nouvelle-Zélande à l'Université de Waikato, il est développé et distribué en coopération avec l'UNESCO et le Projet de Bibliothèques Humanitaires et de développement. C'est un logiciel libre (Open Source) et qui est disponible pour téléchargement sur le site web http://greenstone.org, la licence d'utilisation est la licence publique générale de GNU (GPL). »

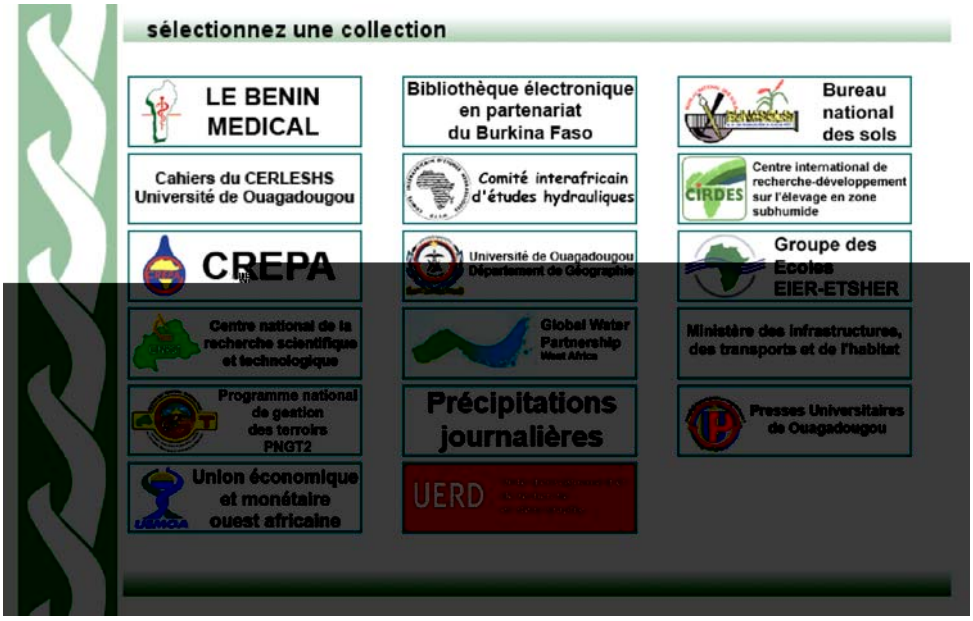
Il permet de rendre accessibles sur un site web des ensembles de documents électroniques ou des bases de données bibliographiques.

Les collections de documents dans le logiciel Greenstone

Pour la création de la bibliothèque d'information scientifique et technique du Burkina Faso nous avions prévu de réaliser un point d'entrée pour chaque Institution et un point d'entrée commun pour l'ensemble des Institutions.

Ce principe correspond bien aux solutions mises en œuvre par Greenstone puisque ce logiciel se base sur le principe de constructions de collections et d'ensembles de collections.

Extrait de la page d'accueil de la bibliothèque électronique en partenariat du Burkina Faso



Une collection se compose, dans notre cas, d'un ensemble de fichiers pdf qui sont les fichiers correspondant aux documents de chaque institution que nous avons numérisé.

Chaque collection est présentée par un logo et par un texte qui fourni des informations concernant l'Institution qui a produit ces documents, le contenu de la collection et les différentes modes de navigation et d'interrogation.

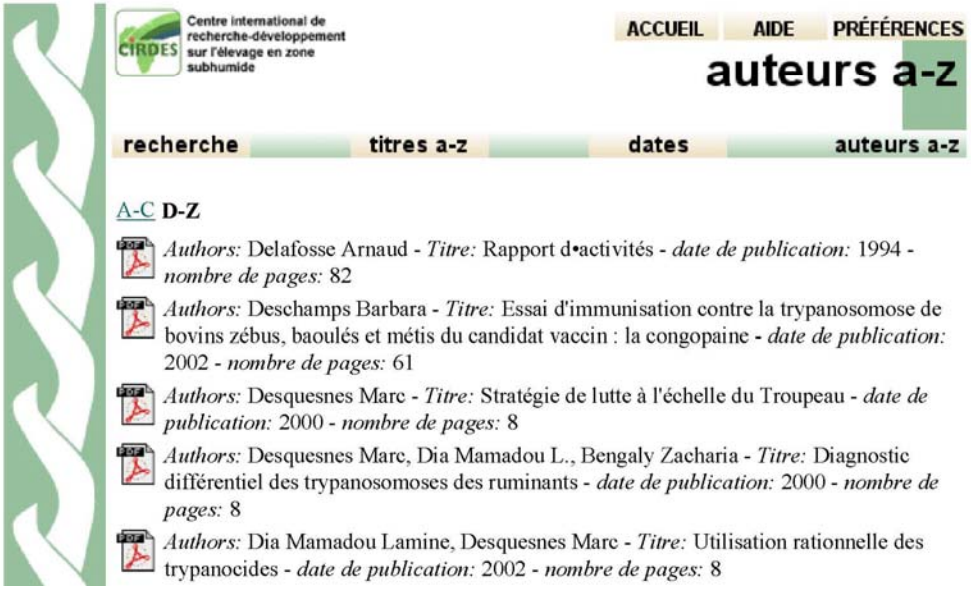
Ces textes de présentation et d'explication peuvent être affichés dans la langue choisie selon les préférences de l'utilisateur puisque Greenstone gère le multilinguisme.

Extrait de la page de présentation d'une collection de documents d'une Institution

La navigation et la recherche avec le logiciel Greenstone

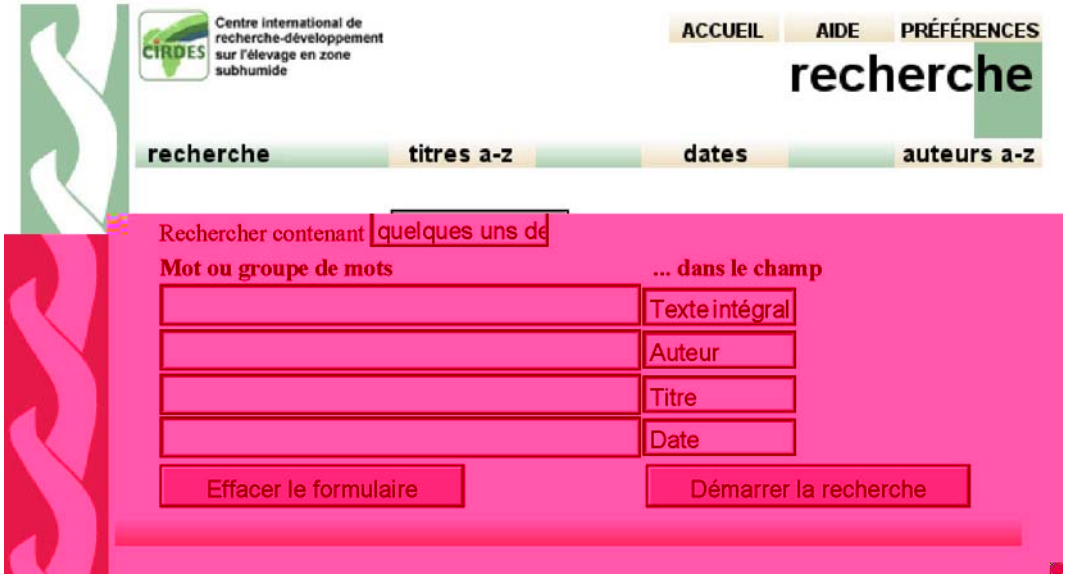
Pour chaque document numérisé nous avons renseigné les champs titre, auteurs et date de publication. Greenstone extrait et gère ces métadonnées qui permettent une interrogation par champs et permettent également la génération de listes de présentation et de navigation selon ces différents critères.

Extrait de la page de présentation d'une liste de références par le critère auteurs



La recherche s'effectue grâce à un formulaire spécifique et paramétrable par l'utilisateur. Il est possible d'effectuer une recherche dans le texte intégral des documents et dans les champs auteurs, titre, date de publication.

Extrait de la page de recherche en texte intégral et par champs



En ce qui concerne la collection constituée par l'ensemble des documents disponibles, point d'entrée commun de la bibliothèque électronique, il est possible d'effectuer une recherche dans le texte intégral des documents et dans les champs auteurs, titre, date de publication. L'utilisateur peut aussi paramétrer cette recherche en la limitant aux collections qu'il juge pertinentes pour ses recherches. Pour cette collection commune la navigation par listes n'est pas possible.
Le résultat de l'interrogation est triée selon les critères de fréquence des mots recherchés et par la proximité, lorsqu'on effectue une recherche en utilisant les opérateurs prévus à cet effet.

A partir des listes des réponses à une interrogation ou des listes d'affichage par le titre, les auteurs ou la date de publication, l'utilisateur peut choisir l'accès aux documents en texte intégral avec un clic sur l'icône du document.

Avantages et inconvénients du logiciel Greenstone

Il nous semble utile de rappeler que, comme tous les produits, le logiciel Greenstone possède des avantages et des inconvénients. Nous essayons de citer ceux qui nous sont apparus lors de son utilisation pour la réalisation du projet.

Avantages du logiciel Greenstone :

- Le logiciel est basé sur des standard ouverts (langage XML), est un logiciel libre, est multiplateformes et il intègre le multilinguisme.
- Son installation, son paramétrage, son utilisation sont d'un niveau accessible à un très grand nombre d'utilisateurs et il est bien documenté.
- Le logiciel intègre le concept de collection de documents avec une bonne modularité.
- Il permet d'effectuer des interrogations par champs et en texte intégral et la création de listes de navigation.

- Il permet d'intégrer plusieurs type de fichiers venant d'applications très variées et d'intégrer également des bases de données bibliographiques.
- Son usage est très répandu, notamment dans le monde anglophone et hispanophone, et il existe un forum des utilisateurs.

Inconvénients du logiciel Greenstone :

- La mise à jour d'une collection nécessite une nouvelle création.
- Pour les fichiers pdf, Greenstone n'intègre pas la mise en surbrillance du texte recherché.
- Le forum de discussion est en anglais.
- Le paramétrage du logiciel est étendu et il est donc difficile d'avoir une appropriation complète des outils.
- La gestion et la modification des scripts de paramétrage nécessitent un bon niveau informatique.
- La modification et l'adaptation des pages de navigation nécessitent un investissement conséquent de la part de l'administrateur du site.

Conclusions

Cette expérience de réalisation d'une bibliothèque électronique au Burkina Faso, basée sur les savoir-faire mis à disposition par l'IRD, montre que le modèle qui a été utilisé est transposable dans le contexte des pays en développement.
La création d'un atelier commun de numérisation, assurant l'ensemble des étapes de la chaîne de traitement, a recueilli l'adhésion des différents partenaires et a permis de traiter un nombre conséquent de documents sur un temps très court (cinq mois).
Cette solution nous paraît adaptée pour les Institutions qui disposent de fonds patrimoniaux de petite taille, pour lesquels l'atelier a pu numériser l'intégralité de leur production, et pourrait également être généralisée pour conduire à des économies d'échelle, à des capacités importantes de traitement et à une excellente qualité des résultats.
Pour des projets de numérisation de fonds scientifiques, les savoir-faire à mettre en œuvre et à acquérir sont très spécifiques. Il en est de même pour les équipements (scanners de productions, scanner à plat A0, scanner de livres), si l'on souhaite répondre à toutes les problématiques de la numérisation (documents hors format, documents à reliure rigide, gros volumes de documents, …). Pour ces différents aspects, un modèle coopératif, fondé sur un atelier commun de numérisation, nous paraît parmi les meilleures solutions à rechercher.

L'utilisation de logiciel Greenstone pour la réalisation du site web de la bibliothèque électronique, nous paraît répondre aux problématiques techniques du projet et aux problématiques de communication de chaque partenaire (identification et présentation spécifique pour chaque institution).
Le logiciel permet de mettre en œuvre une navigation aisée dans les collections des documents, d'effectuer des recherches dans le texte intégral, dans les champs et de réaliser une interface commune d'interrogation pour l'ensemble des collections disponibles.
Il est très modulaire, dispose d'une architecture sophistiquée et gère le multilinguisme. Comme d'autres moteurs de recherche, il n'intègre pas la mis en surbrillance du texte recherché dans les fichiers pdf, mais cette fonction peut être lancée par l'utilisateur lors de l'affichage du document sur son poste de travail.