Bibliothèque numérique Greenstone: du papier à la collection

Dr Michel Loots, Dan Camarzan et Ian H. Witten

ONG Human Info, Belgique
Simple Words, Roumanie
Université de Waikato, Nouvelle-Zélande

Greenstone est une suite logicielle destinée à la construction et à la distribution de collections de bibliothèques numériques. Cette suite fournit une nouvelle manière d'organiser l'information et de la publier sur l'Internet ou sur un cédérom. Greenstone est produit par le projet de bibliothèque numérique de Nouvelle-Zélande (dépendant de l'université de Waikato), et développé et distribué en coopération avec l'UNESCO et l'ONG Human Info.

C'est un logiciel Open Source, diffusé selon les termes de la licence publique générale de GNU, et qu'on peut obtenir à l'URL `http://greenstone.org`.

Nous souhaitons nous assurer que ce logiciel fonctionne bien pour vous.
Faites-nous part, en anglais, de tout problème à l'adresse électronique
`greenstone@cs.waikato.ac.nz`.

Greenstone gsdl-2.50 Mars 2004

À propos de ce manuel

Ce document explique comment créer des collections de cédéroms à partir de documents papier. Il décrit en détail les procédures et les coûts relatifs à la numérisation et aux processus de reconnaissance optique de caractères (ROC, ou *OCR*), de manière à obtenir des textes dans un format que Greenstone puisse accepter. Il décrit aussi comment créer et éditer le matériel associé à une collection.

Nous nous sommes efforcés d'être aussi clairs et complets que possible dans nos explication. Toute marque commerciale mentionnée ne l'est que dans un but d'illustration, et cela ne signifie en rien que nous conseillions, favorisions ou recommandions ce produit d'une quelconque manière.

Documents d'accompagnement

L'ensemble des documents de Greenstone comprend cinq volumes :

- *Guide d'installation de la bibliothèque numérique Greenstone*
- *Guide de l'utilisateur de la bibliothèque numérique Greenstone*
- *Guide du développeur de la bibliothèque numérique Greenstone*
- *Bibliothèque numérique Greenstone : du papier à la collection* (*ce document-ci*)
- *Bibliothèque numérique Greenstone : Utilisation de "L'organizer"*

Remerciements

L'opération de numérisation, l'Organizer et tout le savoir-faire relatif à la création de collections à but non lucratif par le fruit d'un travail de collaboration furent développés par Michel Loots, docteur en médecine, membre des projets Human Info NGO et HumanityCD, et par Dan Camarzan de Simple Words, ainsi que par leurs collaborateurs de Brasov, en Roumanie.

Le logiciel Greenstone a vu le jour grâce à un effort de collaboration entre de nombreuses personnes. Rodger McNab et Stefan Boddie en sont les principaux architectes et développeurs. Des contributions ont été faites par David Bainbridge, George Buchanan, Hong Chen, Michael Dewsnip, Katherine Don, Elke Duncker, Carl Gutwin, Geoff Holmes, Dana McKay, John McPherson, Craig Nevill-Manning, Dynal Patel, Gordon Paynter, Bernhard Pfahringer, Todd Reed, Bill Rogers, John Thompson, et Stuart Yeates. D'autres membres du Projet de bibliothèque numérique de Nouvelle-Zélande ont également donné des conseils

et inspiré les concepteurs du système : Mark Apperley, Sally Jo Cunningham, Steve Jones, Te Taka Keegan, Michel Loots, Malika Mahoui, et Lloyd Smith. Nous remercions aussi tous ceux qui ont contribué au développement des modules sous licence GNU GPL qui font partie de cette distribution : MG, GDBM, PDFTOHTML, PERL, WGET, WVWARE, XLHTML.

Table des matières

Liste des tableaux

1 Introduction

L'un des objectifs du logiciel de bibliothèque numérique Greenstone est de permettre à des entités telles que des universités, des agences des Nations Unies, des organisations non gouvernementales, des associations à but non lucratif, ou des gouvernements, de créer des collections variées d'informations qui pourront être diffusées en ligne ou sur cédérom.

Les étapes à suivre seront typiquement :

i Sélection des documents à inclure
ii S'assurer des autorisations d'utilisation de ces documents (droits d'auteurs et copyrights) dans la bibliothèque numérique
iii Numérisation et ROC des documents papier non disponibles sous forme numérique de manière à disposer d'un format numérique parfait
iv Conversion de tous les documents dans un format (comprenant texte et images) qu'il est possible d'importer dans Greenstone (de préférence HTML ou Microsoft Word, mais d'autres formats sont également possibles, et traités plus ou moins précisément par un «greffon» (*plugin*) qui leur est propre (voir à ce sujet le *Guide de l'utilisateur de la bibliothèque numérique Greenstone*))
v Étiquetage des chapitres, paragraphes et images des documents numériques
vi Organisation de la collection sous forme d'une bibliothèque numérique structurée de manière optimale
vii Construction de la bibliothèque numérique à l'aide du logiciel Greenstone
viii Pressage et distribution de la collection sur cédérom et/ou distribution sur l'Internet

Pour créer une collection numérique, les publications doivent être disponibles dans un format numérique. Si des livres, revues ou d'autres documents ne sont disponibles que sous forme de papier, il faudra les numériser et les traiter sous un format lisible par ordinateur (étape iii). Généralement on procède par reconnaissance optique de caractères (ROC, ou *OCR*), mais parfois on saisit les documents à nouveau, au clavier. Ce processus est décrit dans les sections 2 à 4 du présent manuel.

L'étape v active les différentes portions d'un document qui pourront être sélectionnées indépendamment et affichées par les lecteurs dans la bibliothèque résultante, alors que l'étape vi correspond à l'affectation d'attributs à des documents tels que les catégories de thèmes, de mots-clefs, et les données bibliographiques, utilisées pour le classement et la recherche dans la bibliothèque. Ces étapes sont traitées au chapitre 5 de ce manuel.

Ce manuel présente les nombreux problèmes qui se posent lors du processus d'édi-

tion et de création d'une collection à partir de documents papier. Avant de poursuivre la lecture, posez-vous les questions suivantes :

- Quel but poursuit la collection ?
- Quel public est ciblé ?
- Quelle couverture géographique — locale, régionale, mondiale ?
- Combien de documents seront rendus disponibles ?
- Combien de pages ?
- Quelle quantité de contenus graphiques ?
- Le contenu se divise-t-il en portions que seul un public limité consultera et en portions à disséminer plus largement ?
- Les documents sont-ils déjà disponibles sous forme électronique ?
- Si tel est le cas, sous quels formats ? (remarquez au passage que des fichiers au format PDF ne sont pas toujours équivalents à leur contenu textuel sous forme numérique, puisqu'ils ne contiennent souvent que des images, une par page)
- Quels sont les droits qui s'appliquent aux documents ?
- Qui possède des droits ?
- Existe-t-il d'autres organisations ciblant le même public ?
- Souhaitez-vous collaborer avec d'autres groupes ?
- Quel est le budget global du projet ?
- De quelles ressources humaines (en hommes-mois) disposez-vous pour la coordination, l'édition, la numérisation et la programmation ?
- De combien d'ordinateurs disposez-vous pour ce projet ?
- Combien de cédéroms souhaitez-vous distribuer ?
- Seront-ils gratuits ou payants ?

2 Les scanners et la numérisation

La première étape dans la conversion de documents papier en une collection pour bibliothèque numérique est d'obtenir des images de toutes les pages de toutes les publications sous forme numérique. L'étape suivante consiste à appliquer une reconnaissance optique de caractères (ROC, ou *OCR*), pour la réussite de laquelle il est essentiel de disposer d'images propres et de bonne qualité. Le processus de numérisation utilise un scanner qui travaille à une résolution minimale de 300 dpi (points par pouce, ou *dots per inch*). La plupart des numérisations peuvent se faire en noir et blanc, mais en présence d'illustrations en couleurs il faut utiliser un scanner couleur. Dans la plupart des cas les couvertures des livres sont en couleurs et il faudra les numériser en tant que photographie couleur.

2.1 Les scanners

On trouve des scanners à tous les prix, de toutes tailles et de toutes formes. Ils coûtent de 100$ pour les scanners à plat à 50 000$ pour de gros scanners industriels de fabricants tels que Bell & Howell[1]. On trouve de nombreux sites web qui proposent de nombreux scanners à vendre. Vous les trouverez facilement en tapant le mot-clef «scanner» dans des moteurs de recherche tels que Google, Altavista, ou Yahoo.

Le format de sortie des scanners est un fichier informatique dont le format est souvent TIFF ou bitmap. Le meilleur format est le TIFF IV compressé : une page numérisée et convertie en ce format n'occupe en moyenne que 50 kilo-octets, à comparer aux 2 méga-octets qu'elle occuperait sous forme bitmap non compressée.

2.1.1 Scanners à plat de bas de gamme

Les scanners à plat sont les moins onéreux et les plus communs. De nombreuses marques en proposent : HP, Agfa, Acer, etc. Les prix s'étalent de 100 à 300$. On peut numériser à la fois des images en noir et blanc ou en couleurs.

Le faible prix permet d'équiper chaque ordinateur de son propre scanner.

Ces scanners ont pour inconvénients une qualité moyenne du résultat, une faible vitesse de numérisation, une fiabilité réduite quand la température s'élève, et des

[1] Tous les prix mentionnés représentent des dollars américains (USD) et sont donnés sur la base du marché en 2001.

pannes fréquentes. Il faut numériser les pages manuellement, une à une. Chaque page doit être placée avec attention sur la vitre de numérisation de manière à obtenir un alignement correct. Ces scanners ont une faible productivité. Bien que les fabricants prétendent qu'on peut scanner une page en moins d'une minute, il est rare dans la pratique de pouvoir scanner plus de douze pages en une heure. Le processus de numérisation monopolise l'ordinateur utilisé.

C'est pourquoi de tels scanners ne sont utiles que pour des petits travaux de numérisation, avec de 200 à 400 pages par mois de manière régulière, ou de 1000 à 2000 pages de manière exceptionnelle.

2.1.2 Scanners de bas de gamme avec chargeur

Ces scanners coûtent de 500 à 1200$. On peut insérer, scanner et traiter de 10 à 50 pages d'un coup : l'opérateur ne doit pas rester constamment aux côtés de la machine. Cela permet d'atteindre une productivité de 150 à 200 pages par jour. Ces scanners sont plus robustes et tombent en panne moins souvent — ils permettent généralement de traiter de 30 à 50 000 pages avant de nécessiter une réparation.

Ils ont pour inconvénient de ne numériser qu'une face de chaque feuille à la fois — il faut retourner la pile de feuilles et la numériser de nouveau pour obtenir les versos. Ceci crée des problèmes car les chargeurs causent souvent des ennuis des bourrages papier.

Ces scanners sont utiles pour des volumes de 1500 à 3000 pages par mois.

2.1.3 Scanners couleur

Toute opération de numérisation se heurte tôt ou tard au problème des images couleur, c'est pourquoi il faudra toujours disposer d'un scanner couleur. De manière générale, moins de 5% de toute publication contient des images couleur, sans compter la couverture. C'est pourquoi un scanner à plat tel que décrit ci-dessus suffira. Nous conseillons de choisir un scanner capable d'une résolution de 600 dpi.

2.1.4 Scanners professionnels bi-faces

Les scanners professionnels sont des machines fiables et industrielles, capables de traiter un grand volume — de 2 à 10 000 pages par jour. Ils disposent d'un système de chargeur automatique à plateau qui peut recevoir des paquets de 50 à

200 pages. Les meilleurs et les plus rapides sont des machines duplex qui peuvent numériser les deux faces d'une feuille d'un coup.

Les scanners professionnels duplex ont besoin d'un ordinateur puissant équipé d'un disque dur de 10 à 20 giga-octets. Les prix s'étalent de 5 à 50 000$. Par exemple, le scanner duplex Canon DR-6020 coûte 5000$ et peut traiter des documents recto-verso. Il a une capacité d'environ 2000 pages par jour et une durée de vie de 600 à 800 000 pages. Les scanners Bell & Howell et Fujitsu coûtent de 10 à 50 000$ et ont une durée de vie de plusieurs millions de pages.

Les scanners à micro-fiches coûtent de 15 000$ pour une unité semi-automatique à 80 000$ pour une unité entièrement automatisée.

2.1.5 Logiciels de numérisation

Tous les scanners sont fournis avec leur propre logiciel, qui doit donc être installé sur l'ordinateur qui pilote le scanner. Certains scanners disposent d'une carte à enficher sur l'ordinateur pour accélérer l'opération de numérisation.

2.2 La préparation des documents

Avant d'être numérisés, les documents doivent être correctement préparés. Il faut dépoussiérer, sécher, enlever les agrafes, et aplanir les pages pliées.

Il faut ôter la reliure des livres par une coupure nette, droite et précise. Les livres des bibliothèques devront souvent être reliés à nouveau, auquel cas la plus grande attention est nécessaire lors de cette opération, ce qui facilitera la mise en place de la nouvelle reliure.

Pour un nombre limité de documents, la coupe peut se faire à la règle et au cutter. Faites attention à vos mains ! Pour des plus gros volumes, il existe des machines à couper faites pour.

Pour des gros volumes (à partir de 20 documents) nous recommandons de demander à un imprimeur ou une boîte à copies l'autorisation d'utiliser leur massicot professionnel. N'oubliez pas d'ôter toute agrafe ou trombone ; ils pourraient endommager les lames.

2.3 L'opération de numérisation

Quand on utilise le logiciel fourni avec le scanner, chaque page est numérisée et transformée en image numérique au format TIFF ou bitmap. Ces images doivent alors être stockées sur le disque dur sous des noms de fichiers habituels. La reconnaissance optique de caractères débute à la fin de la numérisation de tout ou partie d'un lot de documents.

Il faudra une résolution de 300 dpi pour la numérisation, même si parfois 200 dpi peuvent suffire.

2.3.1 Contrôle qualité

La numérisation a pour but de passer les pages à la ROC pour produire des versions traitement de texte ou HTML des publications, ou de produire des fichiers image de bonne qualité tels que des fichiers image PDF. Dans tous les cas, la qualité de la numérisation est cruciale : une mauvaise qualité produira des images peu jolies, qui occuperont plus de mémoire. La netteté des images affecte énormément le processus de ROC : la productivité peut chuter jusqu'à 40% si la qualité n'est pas au rendez-vous. Sachant que l'opération de ROC représente plus de 90% du coût total de la transformation du papier en collection, on constate que la qualité de la numérisation peut avoir des effets très marqués sur le coût global.

On peut améliorer la qualité du fichier TIFF en ajustant le processus de numérisation à chaque type de papier, en utilisant les réglages fournis par le logiciel de numérisation. Les papiers relativement transparents requerront des réglages plus clairs ; il faut ajuster le contraste en fonction de la qualité d'impression, etc.

Commencez par diviser les feuilles en lots de texture de papier et de qualité d'impression comparables. Faites des tests de ROC sur un échantillon du premier lot pour déterminer les réglages optimaux. Numérisez ensuite tout ce lot avant de passer au suivant.

2.3.2 Conventions sur les noms de fichier

Attribuez à chaque livre ou document un numéro de traitement ou un code unique, qui sera utilisé en tant que nom du répertoire contenant tous les fichiers TIFF relatifs au document. Selon votre système d'exploitation (DOS, Windows, Unix, Linux, etc.), vous pourrez utiliser de 8 à 128 caractères pour les noms de fichiers. Nous vous recommandons de vous limiter à 8 ou 16 caractères pour cet identifiant de document. Les 5 premiers caractères pourront par exemple identifier le

document, la lettre suivante pourra contenir un code de langue de rédaction du document, et les caractères restants pourront identifier la page. Par exemple, l'identifiant `u7548f12.tif` pourra représenter l'image TIFF de la page 12 d'un livre écrit en français, de code u7548.

Réservez un répertoire du disque dur aux opérations de numérisation (appelons-le par exemple `scanjobs`). Puis créez un sous-répertoire pour chaque lot. Créez à nouveau un sous-répertoire pour chaque publication, comme par exemple `u7548f` pour le document ci-dessus évoqué. Stockez toutes les images TIFF de la publication, y compris les images couleur, dans ce répertoire.

2.4 De la productivité et des ressources nécessaires

Ne sous-estimez pas l'ampleur de l'opération de numérisation — et en particulier celle de la ROC qui la suivra. Il vaut mieux considérer la numérisation et la ROC comme deux activités complètement séparées. Les choix optimaux, tant économiques que pratiques, seront faits indépendamment.

Les aspects auxquels réfléchir incluent les investissements nécessaires en matériel (scanners et ordinateurs) ; la disponibilité d'un espace disque suffisant ainsi que d'assez de ressources humaines ; la formation des agents ; les coûts salariaux ; les nombres de pages initial et final à numériser, les dates maximales de fin de travaux ; et les possibilités de sous-traitance.

Coûts de numérisation

Une décision importante consiste à trancher entre investir en achat d'équipement de numérisation et effectuer toute la numérisation soi-même, ou sous-traiter cette tâche à une société spécialisée. Les critères principaux sont :

– urgence du travail de numérisation ;
– nombre total de pages ;
– coûts salariaux des opérateurs de numérisation.

Les opérateurs doivent être extrêmement motivés, doués techniquement, et soucieux de qualité.

Une société spécialisée réclamera en moyenne 0.06$ par page. Il faut ajouter à cela le coût du transfert des documents, qui peut s'élever à 0.03$ pour un envoi d'un pays en voie de développement vers un pays développé, et 0.015$ pour un envoi entre pays développés.

Le tableau 1 donne une estimation du coût du processus de numérisation s'il est effectué par vos soins, en utilisant différents types de scanners. Vous remarquerez que ces chiffres sont des estimations, et ne sont fournis qu'en tant que référence approximative en se fondant sur l'expérience des auteurs. Les trois premières colonnes traitent des coûts salariaux. La première présente la capacité en pages par mois, sur la base d'un temps plein. On trouve les ressources nécessaires en homme-mois par page en divisant le nombre mensuel d'heures travaillées par la capacité en pages par mois de la deuxième colonne. Il est présenté dans la troisième colonne, sur la base de 180 heures travaillées par mois.

TAB. 1 – Coût de la numérisation

	Capacité (pages/mois)	Heures/page (180 h/mois)	Coût/page (4$/h)	Achat du scanner	Durée de vie du scanner (en pages)	Coût du scanner (0.06$/p.)
Scanner à plat	2 500	0.072	0.288$	300$	7 000$	5 000
Scanner à chargeur	8 000	0.0225	0.09$	800$	30 000	13 000
Scanner pro. bas de gamme	40 000	0.0045	0.018$	6 000$	600 000	100 000
Scanner pro. haut de gamme	150 000	0.0012	0.0048$	50 000$	8 000 000	833 000

Le prix par page s'obtient en multipliant le coût salarial horaire global dans votre cas par la deuxième colonne du tableau 1. Nous avons donné en exemple, dans la troisième colonne, le prix d'une numérisation faite sur place avec un coût salarial de 4$ par heure — ce qui ne comprend pas le coût des investissements.

Ces calculs supposent que le scanner est utilisé pour un volume suffisant, justifiant l'investissement. Les trois dernières colonnes du tableau 1 fournissent plus d'informations sur le coût du scanner à proprement parler. La première montre le coût d'achat du scanner, et la deuxième fournit sa durée de vie moyenne. La dernière colonne montre le nombre de pages qu'on pourrait faire numériser par un sous-traitant, au prix de 0.06$ par page, pour le prix du scanner seul.

Bien sûr, de nombreux autres facteurs guident le choix du scanner : disponibilité des fonds, souhait d'indépendance, volonté de développer un pôle de compétences local, règles des bibliothèques imposant une numérisation locale sans possibilité de transport des livres, etc.

Ces chiffres donnent une idée du volume de pages à traiter pour justifier différents niveaux d'investissement. Un institut ou une organisation aura rarement besoin de numériser plus de 800 000 pages. À de tels niveaux, des paramètres plus complexes entrent en ligne de compte — tels que la maintenance et la possibilité de rentabiliser l'investissement en revendant des services de numérisation — para-

mètres dont nous ne traiterons pas ici.

On peut être séduit par l'idée de développer une activité commerciale de numérisation, surtout dans les pays en voie de développement. Mais gardez à l'esprit que la numérisation est une activité non récurrente : une fois les documents numérisés, les clients ne passeront jamais une autre commande pour la numérisation des mêmes documents, même si les relations sont excellentes. D'un point de vue commercial, il faut prévoir d'intenses efforts de mercatique. Nous déconseillons aux ONG et aux autres organisations à but non lucratif de s'aventurer dans un tel projet sans essais liminaires très complets et un projet financier extrêmement réfléchi.

En conclusion, on peut dire qu'il vaut mieux sous-traiter si la quantité de pages à numériser s'étale de 10 à 50 000 pages. Un scanner professionnel de bas de gamme, coûtant environ 6000$, ne se justifie que s'il faut numériser plus de 100 000 pages. On peut aussi envisager une association de plusieurs institutions (telles que des ONG ou des bibliothèques) pour l'achat groupé d'une telle machine.

3 ROC : reconnaissance optique de caractères

Un système de reconnaissance optique de caractères, ou ROC, transforme une image numérisée en texte. Il accepte en entrée une image numérique au format TIFF ou bitmap, de préférence propre et de bonne qualité. Il fournit en sortie un fichier de traitement de texte ou pour le web, aux formats RTF, Word, ou HTML.

La conversion de documents papier sous forme électronique est un processus en quatre étapes :

1. numérisation ;
2. analyse de la mise en page ;
3. reconnaissance ;
4. numérisation des images et des tableaux.

En suivant ces étapes, il faut effectuer des contrôles qualité sur les fichiers produits, et les sauvegarder dans le format approprié.

On trouve beaucoup de bons programmes de ROC sur le marché, et leurs prix varient de 100 à 400$[2]. On trouve par exemple, entre (nombreux) autres :

– Read-Iris (`http://www.readiris.com/`)
– Omnipage (`http://www.omnipage.com/`)
– Fine-Reader (`http://www.finereader.com/`)

Les sites web des éditeurs vous fourniront toutes les informations nécessaires, y compris la liste des revendeurs dans votre région. L'expérience des auteurs les amène à recommander pour leur bonne ergonomie les programmes Fine-Reader et Omnipage. Fine-Reader est le moins cher, à environ 100$. Il est très souple, et a les options de langue les plus variées.

Il faut choisir entre effectuer la numérisation et la ROC sur place, ou sous-traiter ces opérations à une organisation spécialisée. Un travail sur place nécessite un scanner, un logiciel de ROC, des compétences en ROC (donc des formations), et des opérateurs soucieux de qualité et extrêmement motivés.

3.1 Le processus de ROC

Le processus de ROC change d'un programme à l'autre, et chacun est très long à apprendre et à maîtriser. Le manuel du logiciel expliquera ce processus en détail.

[2]Rappel : tous les prix mentionnés représentent des dollars américains (USD) et sont donnés sur la base du marché en 2001.

Quatre aspects méritent qu'on leur accorde une attention particulière : le contrôle qualité, les tableaux, les images, et les contenus spécialisés tels que formules, caractères d'autres alphabets ou langues, etc.

3.1.1 Contrôle qualité

On n'insistera jamais assez sur ce point. Il vaut mieux faire mener ces contrôles qualité par des locuteurs dont la langue traitée est la langue maternelle, ou des gens maîtrisant parfaitement cette langue. Les meilleurs candidats se recruteront à l'université ou au lycée. Remarquons que des relecteurs plus jeunes seront capables d'une concentration plus soutenue pour ce type de travail.

On compte normalement quatre contrôles qualité.

Le premier est effectué au moment de la ROC. Tout programme de ROC dispose d'un vérificateur orthographique intégré qui met en valeur toutes les lettres suspectes. Il affiche en même temps l'image du mot concerné, ce qui facilite le travail de vérification et de correction de l'erreur.

Le deuxième est une vérification globale du texte à la fin du processus de ROC. Il arrive souvent d'oublier une page, un paragraphe, un titre de chapitre, etc. Un examen global est nécessaire pour détecter d'éventuelles pages manquantes. Il est essentiel de vérifier les titres, les en-têtes de chapitres, les paragraphes, et les tableaux.

Le troisième est une vérification orthographique utilisant Microsoft Word : ce programme dispose en effet d'un dictionnaire souvent plus sophistiqué que ceux qui sont embarqués dans les programmes de ROC. En important le livre dans Word et en y effectuant une vérification orthographique, on peut trouver et corriger des erreurs supplémentaires. Veillez à enseigner au vérificateur orthographique tous les mots particulièrement difficiles ou sujets à erreurs, ou encore les termes scientifiques et techniques communs dans le type de publication considéré.

Enfin, le document complet devrait être vérifié par un relecteur indépendant, qui en extrait des échantillons et contrôle l'absence d'erreurs, de problèmes avec les tableaux et les images, la typographie, et l'aspect général du texte résultant. Ce n'est qu'après cette étape qu'un livre peut être déclaré apte à la dissémination numérique.

3.1.2 Tableaux

Les programmes de ROC peinent à traiter les tableaux. De plus, il est difficile de les vérifier : ils contiennent de nombreux chiffres, souvent avec des espaces, points ou virgules, et il est aisé de décaler des cellules d'une ligne ou d'une colonne. Il leur faut un effort de concentration particulier et une relecture obstinée et intense, une vérification soigneuse, et un bon contrôle qualité. On peut les traiter de trois manières extrêmement différentes.

D'abord, les tableaux peuvent être traités en tant qu'images. Cela implique de les numériser sous forme d'image noir et blanc et de les placer au bon endroit dans le document résultant. C'est la solution la plus facile : aucun risque d'erreur, et le temps nécessaire est uniquement celui de la création de l'image. Cependant, cette solution est plus gourmande en mémoire. Un autre problème concerne la résolution, pas toujours suffisante lorsqu'il faut afficher de grands tableaux sur un écran d'ordinateur : si on affiche tout le tableau, il est illisible, et si on agrandit l'image (en la faisant sortir de l'écran) pour la rendre lisible, l'utilisateur doit s'y déplacer pour pouvoir en lire toutes les lignes et colonnes, et manque d'une vue d'ensemble.

Ensuite, les tableaux peuvent être recréés manuellement en créant une table comptant le même nombre de lignes et de colonnes et en tapant les cellules une à une, caractère par caractère.

Enfin, le tableau peut passer à la ROC. C'est plus rapide que la saisie manuelle, mais présente un plus grand risque d'erreurs. Les colonnes sont parfois fusionnées, et les points et les virgules ne sont pas bien reconnus.

3.1.3 Images

Les publications contiennent trois grands types d'images différents :

– croquis en noir et blanc ;
– photographies en noir et blanc ;
– photographies en couleurs.

Les croquis noir et blanc se numérisent en mode «croquis» et seront sauvegardés au format GIF ou PNG. Les photographies noir et blanc se numérisent en mode «niveaux de gris» et seront sauvegardées au format GIF ou JPEG. Les photographies couleur se numérisent en mode «couleur» et seront sauvegardées au format JPEG. Dans la plupart des cas, le JPEG de qualité moyenne fournit une résolution suffisante.

Pour la plupart des collections, ce sont les images qui consomment le plus de place sur le disque dur ou sur le cédérom. C'est pourquoi il est important d'optimiser chaque image du point de vue de la clarté et de la lisibilité, tout en minimisant sa taille. Vous économiserez de l'espace disque en négligeant de reprendre tout ou partie des images, de préférence celles qui ne sont pas pertinentes par rapport au texte.

Il faut numériser les images séparément, une à une. Nous vous recommandons de donner aux fichiers image un nom consistant en les 5 ou 6 premiers caractères utilisés pour identifier le document, suivis du numéro de page où apparaît l'image. Une autre solution, dans l'hypothèse où chaque document dispose de son propre répertoire, est de se contenter d'utiliser la lettre `p`, suivie du numéro de page de l'image. Si plusieurs images apparaissent sur la même page, on ajoutera au nom de fichier une lettre supplémentaire : `a`, `b`, `c`... Si par exemple une image JPEG apparaît page 36 de la publication `u7548f` ci-dessus évoquée, elle sera placée dans un fichier appelé `u7548e36.jpg` ou `p36.jpg`.

Après la numérisation des images, on peut mettre au travail les programmes de traitement par lots (*batch*), afin de changer les tailles ou d'améliorer la qualité de toutes les images en une seule passe.

3.1.4 Contenus spécialisés

De nombreux documents renferment des contenus spécialisés tels que des caractères spéciaux, des formules, ou des pages difficiles. Les caractères spéciaux sont souvent issus de langues étrangères ou pourvus de signes diacritiques. Il faut alors utiliser l'option de langue du programme de ROC utilisé et lui indiquer la langue à reconnaître. Les formules devront être recréées manuellement. Parfois cette opération n'est pas possible dans le programme de ROC, et uniquement faisable dans un traitement de texte tel que Microsoft Word. Les pages difficiles, au contenu complexe ou si endommagées qu'on n'a pas pu en obtenir une image nette lors du processus de numérisation, devront parfois être retapées.

3.2 De la productivité et des ressources nécessaires

Comme on l'a déjà signalé, il ne faut pas sous-estimer la difficulté du processus de ROC. Même si ses aspects économiques et pratiques doivent être traités indépendamment de ceux relevant de la numérisation, on trouve des points communs : le nécessaire investissement en ordinateurs ; la disponibilité en ressources humaines et leur encadrement ; la formation des opérateurs ; les coûts salariaux ; le nombre

total de pages à traiter ; et la possibilité de sous-traiter des documents.

Dans cette section, nous faisons partager notre expérience d'opérations de ROC en Belgique, en Roumanie et en Inde. Toutes les études de cas, les calculs et les chiffres présentés font les hypothèses implicites de situations moyennes et de documents de difficulté standard (incluant images et tableaux) tels qu'on en trouve dans la plupart des archives et des bibliothèques, des résultats de très bonne qualité, et une opération à moyen ou long terme.

3.2.1 La ROC intensive

La ROC est une activité difficile, qui requiert une grande concentration et beaucoup de compétences. Avant d'atteindre une productivité et une qualité de croisière, il faut prévoir une période d'apprentissage d'environ six semaines.

Les premières heures de chaque jour sont souvent les plus fructueuses en matière de résultats et de productivité. Après trois heures de travail de ROC, la productivité décroît très rapidement, jusqu'à 50% du niveau initial. Après six heures de travail, la plupart des gens sont très fatigués.

La même courbe se dessine au niveau supérieur, celui des semaines. Les premières semaines, tout le monde travaille vite et bien, mais vient un moment où les deux tiers des agents s'ennuient et deviennent frustrés. Ces personnes abandonnent le projet ou se mettent à travailler de façon médiocre, en quantité comme en qualité. Même ceux qui passent le cap critique des 3 ou 5 semaines de travail et intègrent l'équipe, partent souvent après 6 à 12 mois, à la recherche d'un meilleur poste.

Les remarques de la section 3.1 concernant le personnel sont particulièrement avérées dans le cadre d'un travail de ROC intensif.

Il vaut mieux faire mener les contrôles qualité par des locuteurs dont la langue traitée est la langue maternelle, ou des gens maîtrisant parfaitement cette langue. Des relecteurs plus jeunes seront capables d'une concentration plus soutenue pour des tâches de ROC. Empiriquement, on a constaté que des personnes âgées de 18 à 23 ans convenaient mieux que des personnes de plus de 25 ans.

Enfin, la ROC peut être un travail fastidieux, ce qui donne une importance exceptionnelle aux questions de motivation et de goût du travail bien fait.

Ces remarques sur la ROC mènent aux préceptes suivants :

– Les jeunes gens âgés de 18 à 25 ans sont les plus indiqués pour ce travail.
– Les premières heures étant toujours les plus productives, il faut organiser le travail à temps partiel ou ne retenir que les gens les plus motivés et concentrés pour un travail à temps plein.

– Les deux tiers des gens abandonnent ou s'ennuient après trois à cinq semaines, ce qui se traduit par une qualité et une productivité en baisse les dernières semaines.
– Il faut veiller à fournir un travail régulier pour justifier la formation nécessaire, pour maintenir la concentration, et pour que restent hauts les cœurs.

3.2.2 Productivité possible

Le tableau 2 donne les statistiques moyennes de productivité pour la ROC. Les documents sont de toutes tailles et de toutes qualités, et ces chiffres supposent que le lot de documents contient un nombre moyen d'images et de tableaux — disons une image et une tableau de 5 lignes par 5 colonnes toutes les 8 pages. Ils supposent aussi que les images des pages sont de qualité moyenne à bonne (comme on l'a déjà signalé, ceci dépend de la numérisation) et que les opérateurs maîtrisent bien la langue.

TAB. 2 – Productivité de l'OCR

	Heures travaillées par jour	Pages par jour	Pages par mois
Formation initiale (6 semaines)	3	6	120
Niveau de productivité optimale	3	9	150 à 200
	7	28	500 à 600

Le tableau 2 distingue les cas des opérateurs en formation et celui des opérateurs ayant atteint leur niveau de productivité optimal. Si un agent administratif devait passer 3 heures par jour à des activités de ROC, il pourrait produire 180 à 200 pages par mois. Dans le cas de personnel employé à temps plein, ayant reçu une formation adéquate, avec une concentration élevée et un goût du travail bien fait, on peut obtenir de 500 à 600 pages par mois.

Cependant, les taux obtenus sur des pages difficiles, de qualité médiocre, contenant beaucoup d'images ou de tableaux, sont bien plus faibles — peut-être de 300 à 400 pages par mois pour un travail à temps plein.

Supposons que les coûts salariaux d'opérateurs de ROC motivés et soucieux de qualité travaillant à temps plein s'élèvent à 400$ par mois, et que les frais d'infra-

structure (comprenant les coûts d'encadrement, les ordinateurs, les bureaux, les fournitures, etc.) s'élèvent à 300 à 400$ par personne et par mois. Alors le coût de la ROC est de 1.2$ à 1.6$ par page. Si on prend en compte le temps de formation, le volume total, la durée de l'opération, et les coûts de licenciement si l'opération devait prendre fin par manque de travail, ces coûts atteignent 1.5$ à 2.5$ par page.

Il faut comparer le coût d'une ROC menée sur place à celui d'une ROC sous-traitée à un professionnel. De telles sociétés demandent en général de 1.5$ à 4$ par page, en comptant les images et les tableaux. L'ONG Human Info/Simple Words dispose d'une telle unité en Roumanie, et pratique un tarif spécial pour les organisations humanitaires ou à but non lucratifs : de 1.2$ à 2$ par page. Contactez-nous à l'adresse électronique `scanning@humaninfo.org` si vous souhaitez obtenir des informations ou des conseils complémentaires.

3.3 Pour éviter la ROC

Il existe deux solutions qui permettent d'éviter la ROC, et nous les présentons toutes deux ici.

3.3.1 Saisie manuelle

La première, qui élimine également la plupart des opérations de numérisation, consiste à retaper les documents à la main, en utilisant un traitement de texte. Il faut quand même numériser la couverture et les images, mais les autres pages n'ont pas besoin d'être numérisées, ce qui évite l'achat d'un scanner puissant et de logiciels de ROC.

Les opérateurs n'ont pas besoin de comprendre le texte : il leur suffit d'être des dactylographes précis, qui tapent exactement ce qu'ils voient. La saisie génère des erreurs, qu'on trouve et détecte par la méthode dite de double saisie. Elle consiste à demander à deux personnes de saisir le même document indépendamment, suite à quoi on comparer les deux versions numériques mot à mot à l'aide d'un logiciel spécial manipulé par un opérateur disposant du document original. On suppose implicitement qu'un mot tapé indépendamment deux fois de la même manière est nécessairement correct. Mais cela ne suffit pas toujours, et on peut avoir recours à de la triple saisie pour obtenir une précision extrêmement élevée.

L'avantage de la saisie est l'économie effectuée : nul besoin d'un programme de ROC (qui nécessite des ordinateurs puissants), aussi des ordinateurs plus anciens, ou d'occasion peuvent suffire. De plus, ce travail peut être mené par des personnes moins qualifiées. L'inconvénient est la durée de la période de formation (d'un

minimum de deux mois). Une simple saisie induisant souvent trop d'erreurs, il faut procéder à une double ou à une triple saisie.

Le coût dépend entièrement du salaire. Les dactylos sont généralement payées environ 150$ par mois dans les pays en voie de développement. La productivité est de 20 à 30 pages par jour, pour un total de 400 pages par mois, en comptant les images. Avec une double saisie, cela donne un coût de revient salarial d'environ 300$ par mois, plus les à côtés.

3.3.2 Fichiers image

Une solution de remplacement à la ROC très peu onéreuse est de se contenter d'utiliser une version image PDF des pages du document. Le coût est négligeable devant celui de la ROC — environ 0.1$ par page.

À l'issue du processus de numérisation et une fois que les fichiers TIFF sont disponibles, un convertisseur automatique (on utilise en général Adobe Acrobat ou Adobe Photoshop) transforme tous les fichiers TIFF des pages du livre en fichiers PDF.

Le revers de la médaille est l'impossibilité de mener des recherches textuelles dans de tels fichiers. De plus, ils sont très lourds : environ 50 kilo-octets par page, plus ou moins 20% selon la qualité du fichier TIFF original.

Les fichiers image PDF sont lents à télécharger (parfois, dans les pays en voie de développement, cette opération est impossible ou a un coût prohibitif). Ils tiennent rarement sur une disquette, et il est impossible de manipuler leur texte, ne serait-ce que pour effectuer des copier-coller.

Il faut réserver cette méthode aux situations où aucun budget ne peut être débloqué pour la ROC, et pour les documents susceptibles d'être utilisés par un petit nombre de personnes, disposant d'une connexion Internet à haut débit.

3.4 Mettre bout à bout numérisation et ROC

Si un scanner est directement relié à l'ordinateur qui héberge le logiciel de ROC, la plupart des programmes de ROC peuvent numériser une page et effectuer immédiatement la reconnaissance de caractères. Procéder une page à la fois, en enchaînant numérisation puis ROC, est une stratégie raisonnable pour des petits volumes, mais s'avérera coûteux en temps pour des tâches plus importantes et plus continues.

Cette solution peut suffire pour 100 à 150 pages par mois. Pour des volumes plus

élevés il est plus rapide et plus efficace de commencer par numériser le document, puis de démarrer en deuxième lieu une opération de ROC sur toutes ses pages à la fois.

4 Trois exemples : de 1000 à 100 000 pages

4.1 Cas d'une petite collection : de 500 à 1000 pages

La plupart des ONG ont un volume de 500 à 1000 pages à numériser. Un tel volume peut être traité sur place si on trouve des volontaires motivés.

Numérisation

La première étape consiste à numériser les publications pour produire un fichier TIFF de bonne qualité de chaque page, et une image séparée pour chaque illustration (croquis, image en niveaux de gris ou en couleurs). Si on fait l'hypothèse que 1000 pages doivent être numérisées, ceci peut représenter un travail à temps partiel d'environ un mois — pour la numérisation seule. Les fichiers TIFF occuperont de 60 à 80 méga-octets d'espace disque, et c'est une bonne idée que de graver un cédérom réinscriptible contenant ces fichiers. Un scanner à plat de bas de gamme de 100 à 300$ suffira à mener cette tâche de numérisation à bien. La numérisation peut être prise en charge par un volontaire, après les heures de bureau ou les jours non ouvrés, au bureau ou à la maison.

ROC

La ROC, menée par un autre volontaire ou par une équipe de volontaires, doués en langue et en correction, vient ensuite. Les fichiers TIFF peuvent être partagés entre ordinateurs, ou bien on peut utiliser un seul ordinateur pour l'ensemble du travail. Il faudra en moyenne de 5 à 6 mois à temps partiel (par exemple, environ 20 heures par semaine) pour convertir 1000 pages en bons documents Word ou HTML.

Sous-traiter

On peut aussi envisager de sous-traiter les opérations de numérisation et de ROC. Il en coûtera probablement de 1500 à 2000$ pour tout convertir en bons fichiers Word ou HTML.

4.2 Toutes les publications d'une organisation : 5000 pages

De nombreuses organisations plus importantes disposent d'environ 5000 pages d'archives de livres, journaux, revues, et autres documents, actuels ou épuisés.

Numérisation

Voilà un volume trop important pour un scanner à plat. La numérisation doit donc être sous-traitée (ce qui coûtera environ 400$ pour 5000 pages) ou confiée à un scanner à chargeur (qui coûte environ 900$). On peut aussi envisager l'achat groupé d'un scanner plus performant, avec d'autres institutions ou ONG (il en coûtera 6000$, à diviser par le nombre de participants). Les 5000 pages converties au format TIFF occuperont 300 à 400 méga-octets d'espace disque. Là encore, c'est une bonne idée de graver un cédérom réinscriptible contenant ces fichiers.

ROC

La ROC, menée par un volontaire ou par une équipe de volontaires, doués en langue et en correction, vient ensuite. Ici encore, les fichiers TIFF peuvent être partagés entre ordinateurs, ou bien on peut utiliser un seul ordinateur pour l'ensemble du travail. Il faudra en moyenne de 25 à 30 mois à temps partiel (par exemple, environ 20 heures par semaine) pour convertir 5000 pages en bons documents Word ou HTML. En pratique, c'est là une tâche trop longue et gourmande en ressources informatiques pour pouvoir fonctionner sur la base du volontariat. Il faudrait rémunérer les volontaires, surveiller leurs performances et la qualité de leur travail, fournir l'espace adéquat, etc., pour obtenir un travail finalisé de bonne qualité dans des délais raisonnables.

On peut aussi créer des fichiers image PDF, ce qui occupera de 300 à 400 méga-octets d'espace disque, et sera plus difficile à télécharger depuis l'Internet.

Sous-traiter

On peut aussi envisager de sous-traiter les opérations de numérisation et de ROC. Il en coûtera probablement de 7500 à 10 000$ pour tout convertir en bons fichiers Word ou HTML.

4.3 Une petite bibliothèques : 100 000 pages

Des organisations plus importantes, des universités, des gouvernements, et des bibliothèques spécialisées disposeront peut-être de toute une bibliothèque à numériser — disons, 100 000 pages. La première question à se poser est celle des droits attachés aux publications : si elles ne sont pas placées ou tombées dans le domaine public, il faut obtenir des détenteurs des droits l'autorisation explicite de les numériser. Pensez aussi à vérifier si les fichiers ne sont pas déjà disponibles sous forme numérique.

Numérisation

Le volume est trop important pour un scanner à chargeur. Il faut donc sous-traiter la numérisation (il en coûtera 8000$ pour 100 000 pages) ou acheter un scanner plus performant en commun avec quelques autres institutions ou ONG (il en coûtera 6000$, à diviser par le nombre de participants). Les 100 000 pages converties au format TIFF occuperont 6 à 8 giga-octets d'espace disque. La meilleure idée est de graver une série de copies de ces fichiers sur cédérom réinscriptible.

ROC

La ROC vient ensuite (on peut aussi penser à créer des fichiers PDF pour des documents moins largement utilisés). Il faudra en moyenne de 500 à 700 mois à temps partiel pour convertir 5000 pages en documents Word ou HTML. C'est une opération impossible à mener sur la base du volontariat, et il faut avoir une approche professionnelle.

On peut réduire les coûts en transformant les pages les moins fréquemment utilisées (par exemple, les 80% les moins utilisés, ce qui représente 80 000 pages) en PDF, et ne transformer que les 20 000 pages restantes en Word et en HTML. Les fichiers PDF occuperont 6 à 8 giga-octets d'espace disque et seront plus difficiles à télécharger sur l'Internet, mais ils ne coûteront que 0.2$ par page à faire produire par des professionnels (pour un coût total de 16 000$). S'il fallait faire créer 80 000 fichiers PDF à partir de fichiers TIFF par des volontaires utilisant des programmes de conversion vers PDF tels qu'Adobe Acrobat, il faudrait prévoir 10 à 20 mois de travail à temps partiel sur un ordinateur puissant.

Sous-traiter

On peut aussi envisager de sous-traiter le travail. Si on reste sur l'hypothèse de 80% des pages converties en PDF et les 20% les plus fréquents en HTML, le coût du PDF s'élévera à environ 16 000$ et celui du HTML de 30 à 40 000$, pour un budget global d'environ 50 000$. Si toutes les pages passaient à la ROC, il en coûterait de 150 à 200 000$ pour convertir toute la collection en fichiers Word et HTML.

5 Créer une collection électronique

Quand on décide de créer une collection, il faut garder à l'esprit trois aspects importants. D'abord, il faut organiser la collection. Plus dense et complet sera le contenu, plus le besoin d'indexation et de systèmes de recherche puissants se fera sentir. De tels outils sont indispensables pour des collections de 3000 à 5000 pages ou plus. Ensuite, il faut donner la priorité aux besoins des utilisateurs finals. Identifiez le public ciblé par la collection, et consultez-le régulièrement. Enfin, le budget disponible tranchera tous les choix en matière de développement ou raffinement.

5.1 Méthodes de construction de collections

On trouve de nombreux exemples d'excellents cédéroms créés sur le modèle de la page web. Les documents HTML, PDF ou Word peuvent être ajoutés et reliés au reste de la collection grâce à des liens hypertexte. La navigation est simple et attrayante grâce aux liens hypertexte, aux cadres, aux mots-clefs, aux index, etc. De tels systèmes peuvent convenir pour des volumes de quelques milliers de pages, mais à partir de 3000 ou 5000 pages ils atteindront leurs limites et il deviendra important de disposer d'une collection bien structurée et de fonctionnalités de recherche puissantes. C'est là que Greenstone peut rendre service.

Le logiciel de bibliothèque numérique Greenstone crée une bibliothèque numérique structurée disposant d'un moteur de recherche très puissant. On peut indexer jusqu'à 150 000 pages sur un simple cédérom, et chaque cédérom peut se transformer en serveur web. Greenstone est un logiciel libre, disponible selon les termes de la licence publique générale de GNU.

Les manuels fournis décrivent la manière de confectionner des collections Greenstone. Il y a essentiellement trois façons de créer des collections.

– Le "Librarian Interface" (l'Interface Bibliothécaire)
– Le "Collector" (Le collectionneur)
– La création à partir de la ligne de commande.

La première est le "Librarian Interface" (l'Interface Bibliothécaire) décrit dans le "Guide de l'utilisateur de la bibliothèque numérique Greenstone" (Chapitre 3, "Réaliser des collections Greenstone"). C'est un outil interactif fonctionnel dans la création des collections. Avec lui, on peut collecter des groupes de documents, importer ou assigner des méta-données, et confectionner une collection Greenstone.

La deuxième méthode est le sous-système "collectionneur", décrit au chapitre 4 du guide de l'utilisateur. C'est un outil plus ancien qui fournit une manière alternative de créer des collections de pages Web et autres documents. Il vous guide à travers une séquence de pages Web conventionnelles qui exigent l'information requise. Il ne fournit cependant aucun moyen d'ajouter une métadata aux documents et, parce que c'est une interface Web, il n'est vraiment pas adapté aux collections qui prennent plus que quelques minutes pour être créées.

La troisième méthode consiste à exécuter directement, à partir de la ligne de commande, les programmes de création de collections ; elle figure dans le "Guide du développeur de la collection Greenstone" (Chapitre 1). Elle offre plus de flexibilité dans l'exécution individuelle des programmes et dans la sauvegarde des résultats intermédiaires, hautement souhaitable pour les collections qui prennent des heures à être créées. La lecture du chapitre 2 du Developer's Guide s'avère aussi nécessaire pour exploiter de façon optimale la puissance de Greenstone dans la confection des collections de pointe.

Une quatrième méthode de création et d'édition du matériel associé à la collection existe ; c'est un programme appelé "The Collection Organizer" (Organisateur de la Collection). Cependant, sa fonctionnalité a été dépassée par le "Librarian Interface" (l'Interface Bibliothécaire) mentionné ci-dessus. Ce document est décrit comme un leg sous le titre "Using the Organizer" (Utiliser l'organisateur).

5.2 Commencer en sept étapes et en 15 minutes

La meilleure façon d'appréhender et de sentir l'Interface Bibliothécaire est en fait de créer une bibliothèque par un petit test. Si vous disposez de 15 minutes on vous conseille de suivre ces étapes pour une meilleure compréhension de ces programmes.

Avant tout, vous installez Greenstone (Voir Le Guide d'installation de Greenstone) "The Greenstone intaller's Guide" qui comprend la collection de démonstration "Demo Collection" dans un format DLS et ses fichiers sources. **Noter que si vous souhaitez ajouter à la collection l'un quelconque des 140 documents de la collection DLS (au lieu seulement des 11 de la collection Demo de Greenstone), vous devez installer DLS comme un des modèles des bibliothèques Greenstone.** Dans `C:\Program Files\gsdl\collect`, comme d'ailleurs Demo, dans les sous-repertoires respectifs DLS et Demo. Si Greenstone avait déjà été installé sans la collection DLS et que vous souhaitez installer celle-ci, vous devez réinsérer le Cd-rom de Greenstone et procéder à l'ajout de DLS. La désinstallation de Greenstone n'est en aucun cas nécessaire.

Nous vous conseillons d'imprimer les instructions ci-dessous et de les suivre pas à pas :

1. Démarrer le "Librarian Interface" (l'Interface Bibliothécaire) sous Windows en sélectionnant Greenstone Digital Library à partir de la section Programmes du menu de Démarrage et en sélectionnant "Librarian interface". Si vous utilisez Unix à la place, il faut taper :

   ```
   cd ~/gsdl
   cd gli
   ./gli.sh
   ```

   où /gsdl est le répertoire contenant votre système Greenstone.
2. Sélectionner "Nouveau" à partir du menu "Fichier" dans la barre de menu horizontale placée en haut de la fenêtre. Lui donner un titre, par exemple "Ma première collection" et mettre votre adresse électronique et une description sommaire de la collection. Dans le menu "Baser cette collection sur", choisir "greenstone demo" (démonstration de greenstone) ou "Developement Library Subset" (le résultat est le même parce que ces deux collections ont la même structure).
3. Ajouter des documents à partir de la collection Demo (ou la collection DLS si elle est installée) à votre nouvelle collection. Pour ce faire, double-cliquer sur le répertoire des Collections Greenstone à gauche du panneau, et double-cliquer sur la collection désirée. Les documents qui s'y trouvent sont affichés en dessous. Sélectionnez-en un, le faire glisser et le déposer dans le panneau de droite ("glisser-coller"). Celle-ci représente la collection en cours de création. Choisir plusieurs documents, les faire glisser un à un à l'intérieur, ou utiliser une sélection multiple de façon standard.
4. Ajouter vos propres documents qui ne sont pas dans la collection Demo ou DLS. Fermer le répertoire des collections Greenstone du panneau de gauche et double-cliquer sur le répertoire "Local filespace". Naviguer vers le répertoire qui contient des documents(c'est à dire de petits fichiers Word et HTML). En faire glisser quelques-uns dans le panneau de droite pour les inclure dans votre collection.
5. Ajouter des méta-données aux documents de votre collection. Jusqu'à présent, l'exécution se déroulait sous le panneau "Gather", indiqué par l'onglet "Gather" en-dessous de la barre de menu horizontale en haut de Windows. Cliquer sur l'onglet "Enrich" qui se trouve à côté. Les documents dans votre collection apparaissent maintenant dans le panneau de gauche : cliquer sur l'un et examiner la méta-donnée qui lui est associée dans la liste "Elément

... Valeurs" en haut à droite. Utiliser le panneau en-dessous pour changer les valeurs individuelles en sélectionnant l'élément désiré et, soit choisir une valeur existante sur la liste, ou taper une nouvelle valeur dans la boîte située en bas. Ajouter les meta-données Titre, Organisation et Mot-clé à chacun de vos propres documents figurant sur la collection. Après la saisie de chaque valeur, il est nécessaire de cliquer sur "Ajouter" pour ajouter cette valeur à la méta-donnée.

6. Cliquez l'onglet "Créer" pour quitter le mode "Enrich" et créer votre nouvelle collection. Cliquer sur l'onglet "Créer la Collection" en bas. Pendant que l'ordinateur élabore la collection, vous recevrez en feedback un compte rendu sur ce qui est en cours d'exécution.
7. En fin d'exécution, cliquer sur l'onglet Preview pour visualiser la collection à partir du "Librarian Interface" (l'Interface Bibliothécaire). Vérifier les titres a-z, organisations et comment lister pour vous assurer que vos documents ont été inclus dans la collection. En visitant votre page Web Greenstone, le constat est fait aussi que la collection a été installée parmi les collections de Greenstone.