Le Project DEBORA[1] (Digital accEss to Books of the renAissance)

Richard Bouché
Ihadjadene Majid
Enssib
17/21 bd du 11 nov 1918
F-69100 Villeurbanne
Ihadjade@enssib.fr

Objectifs du Project

Pour des raisons de conservation, les collections d'ouvrages du 16ème siècle ne peuvent actuellement être consultées que par une faible minorité d'experts ou d'érudits. Non seulement DEBORA offrirait à cette minorité des moyens de travail intéressants, mais encore les collections deviendraient accessibles à un plus large public. Coûteux en moyens de conservation et peu utilisé, le fonds ancien peut donc parfois apparaître comme un poids mort dans une bibliothèque. L'accès numérisé, en ouvrant ces collections à un plus nombreux public, permet une valorisation sans nuisance de ces fonds. Mais ceci est vrai pour tout fonds ancien. Le 16ème siècle présente cependant l'intérêt de se situer à un moment où le livre imprimé acquiert ses caractéristiques modernes (apparition et structuration de la page de titre, organisation interne, normalisation de la typographie, etc.) tout en conservant un mode de production artisanal qui durera les deux siècles suivants. Les solutions techniques concernant le traitement du document seront donc valables pour les ouvrages produits jusqu'à la révolution industrielle et qui constitueront un patrimoine culturel européen. C'est également le moment où se développe l'utilisation des langues vernaculaires dans le livre.
D'une façon encore plus générale, l'accès distant à des collections numérisées offre des possibilités de consultation qui ne sont pas encore bien évaluées. On peut penser que la demande devrait s'accroître valorisant en même temps les richesses culturelles des bibliothèques. L'élargissement de l'accès aux collections du 16ème siècle et l'analyse des usages par un nouveau public doit permettre de mieux déterminer les caractéristiques de cette demande. Il s'agira, à terme, d'aboutir à un équivalent européen de "American Memory".

L'objectif de DEBORA (Digital accEss to Books of the RenAissance) est de développer des outils permettant l'accès, à partir de postes de consultation distants, à des collections de documents du 16ème siècle des bibliothèques, par la numérisation des ouvrages. Non seulement DEBORA offrirait à des chercheurs des moyens de travail intéressants, mais encore les collections deviendraient accessibles à un plus large public.
L'objectif de DEBORA est vu dans trois perspectives :

- Une analyse des besoins et une analyse des usages de tels outils,
- Une définition de la chaîne de production de documents numérisés,
- Une analyse économique de la numérisation.

Les résultats attendus sont :

[1] http://www.enssib.fr/debora

D'un point de vue technique, il s'agit de réaliser un outil permettant à la fois de recevoir et de stocker les données issues de la chaîne de numérisation mais aussi de mettre en œuvre des moyens de recherche d'information performants et adaptés aux usages actuels et nouveaux qui auront été définis.

Du point de vue des usages : A partir de la réalisation du système d'accès et de son évaluation, DEBORA cherche à déterminer ce qu'apporte la numérisation qui ne doit pas être vue comme un microfilmage amélioré. En particulier, il s'agira de déterminer les besoins d'un public élargi, qu'il soit savant ou non, et d'identifier les nouveaux usages induits par ces nouveaux moyens d'accès aux collections de bibliothèques.

Du point de vue économique, il s'agit de suivre l'ensemble du projet pour déterminer les différents types de coûts rencontrés dans une chaîne de production et de mettre en évidence les données critiques sur lesquelles un effort d'organisation et de rationalisation devront être entrepris. Les bibliothèques qui ont des projets de numérisation pourront ainsi évaluer les charges qu'ils représentent et disposer d'éléments de décision nécessaires

Description de DEBORA

1. Le choix des métadonnées

Le choix des métadonnées est une phase importante dans tout développement d'une bibliothèque numérique. Il préfigure les modes d'accès et de navigation qu'on offre aux usagers . Nous avons recensées quatre modes de descriptions de livres anciens dans plusieurs projets similaires à DEBORA:

Une description au niveau de la collection. Il s'agit d'un catalogage descriptive de l'ouvrage en utilisant l'une des versions du format MARC. C'est l'option choisie dans Gallica[2] et la bibliothèque numérique du Vatican[3]. Le choix d'une numérisation en mode image implique souvent ce choix.

La description est inexistante ou très limitée: c'est le choix des projets ABU[4] et Guetenberg qui offrent un accès en texte intégral (Ascii) à leurs fonds numérique.

Une description "maison" qui peut être très élaborée mais qui reste éloignée des métadonnées et normes , c'est le cas des projets HELIOS[5] ou Antique Books[6].

Une description élaborée et précise du documents suivant les recommandations de la TEI (Text Encoding Initiative) ou de la métadonnée EAD (Encoding Archival Description) : C'est le choix du système digital scriptorium[7]. Plusieurs textes du 16ème siècle existants dans le Web sont structurés selon ces métadonnées. Ian Lancashire[8] a mis à la disposition internantes une description des livres de la renaissance qui respecte la TEI.

---

[2] http://gallica.bnf.fr/
http://sunsite.berkeley.edu/Scriptorium/
http://www.library.yale.edu/preservation/pobweb.htm
http://www-lis.glis.ucla.edu/DL/UCLA_DL_Report.html
http://lcweb.loc.gov/ead/
http://etext.lib.virginia.edu/TEI.html

[3] http://www.almaden.ibm.com/journal/rd/mintz/mintzer.html

[4] http://cedric.cnam.fr/ABU

[5] http://info.lib.uh.edu/pr/v6/n4/gall6n4.html

[6] http://www.dlib.org/dlib/september97/thibadeau/09thibadeau.html

[7] http://sunsite.berkeley.edu/Scriptorium/

[8] http://www.library.utoronto.ca/utel/ret/guidelines/guidelines1.html

Dans DEBORA, nous avons choisi deux niveaux de descriptions: un catalogage au niveau de la collection en format MARC et respectant la norme ISBD (A); une seule notice permet d'accéder à plusieurs pages-images. Une deuxième description plus fine à l'unité pour exprimer les particularités de chaque page-image (lettrines, collofons,…etc). Une relation est établi en stockant le nom du fichier dans un champ de la notice descriptive MARC. Cette relations sera gérée par encodage de la zone 856 du format MARC. Ce choix de description s'explique par l'intérêt du mode image en numérisation. L'analyse des usages et des usagers nous éclairera certainement sur le choix de description de livres anciens.

2. La chaîne de numérisation

Il s'agira de réunir toutes les spécifications, provenant du démonstrateur, du choix des collections ou des usages qui ont une incidence sur les choix des formats et des normes de numérisation des images et de faire un choix raisonnable parmi ces normes.
Le système Digibook 5600, codéveloppé par Xerox et par la PME I2S, nous apportera des garanties en ce qui concerne le respect des ouvrages du 16ème siècle, lors de la numérisation. Compte tenu du fait qu'il vient à peine d'être construit et de notre volonté de travailler en niveau de gris nous allons devoir le tester comme un matériel expérimental et créer ou adapter des logiciels pour le prétraitement des images de pages de livres. En particulier, nous ferons un logiciel pour le nettoyage des images scannées provenant de textes où l'encre est passée à travers le papier(confusion recto et verso) et un logiciel pour la restauration des images après ce nettoyage; nous réaliserons des logiciels de redressement des caractères d'une part, des lignes d'autre part.
Le mode image présente plusieurs avantages: la mise en page, la disposition des différentes parties, la police de caractère, les imperfections et même ce qui apparaître comme des fautes d'orthographe ou de syntaxe; tous ces éléments donnent une certaine personnalité au document et sont utiles pour identifier des conditions de production. Tous ceux qui s'intéressent au livre, à son évolution, au langage etc… ne peuvent accepter l'appauvrissement que représenterait le passage en mode caractère. Seules une partie de l'ouvrage (tables de matières et index) sera en mode texte (OCR). La numérisation proprement dite comporte :

La détermination des paramètres de numérisation en fonction des ouvrages choisis et des usages prévus.

L'intégration des images de pages dans SGBI avec le traitement nécessaire : L'amélioration globale ou locale de l'image (contraste, luminosité) et, en particulier, l'atténuation ou même la suppression des différents défauts de l'image (taches, traits de pliure, etc.).

La détermination de la structure logique de l'ouvrage (identification des tables des matières et d'index, des premières pages de chapitre, etc.) et la mise en place de liens permettant l'accès aux éléments logiques identifiés.

Il est nécessaire de développer des outils fondés sur la reconnaissance des formes et permettant d'identifier des parties de documents utiles aux usagers (formes particulières (lettrines, bandeaux, illustration, etc.)) et la structure logique du document.

3. L'utilisation des techniques de reconnaissance des formes

L'un des objectifs de DEBORA est la création de méthodes et la réalisation des logiciels correspondants pour extraire les informations nécessaires au fonctionnement de la banque de

données d'images, Ces données seront extraites des informations stockées (entités numérisées en mode image et informations complémentaires). Il ne s'agit pas ici de créer des logiciels pour l'OCR ou pour la dématérialisation mais des outils permettant une sorte de premier accès aux textes. Ils pourront requérir l'assistance de l'homme pour des tâches non triviales et non répétitives, en particulier il s'agit de réaliser les fonctions suivantes:

Adaptation et modification de logiciels existant pour faire une analyse de bas niveau des images permettant : 1) de séparer les composants des documents (texte, illustration, lettrines, etc.), 2) de segmenter les textes en lignes, 3) de segmenter les lignes en graphèmes (mots)

Création de logiciels d'analyse des formes, des images de mots (et de groupes de lettres) permettant de : 1) d'élaborer des distances et des ressemblances entre les mots dans le contexte des documents du 16ème siècle, 2) de retrouver un mot par son image, 3) de déterminer une liste d'index.

Création de logiciels permettant d'accéder partiellement au texte à partir de la table des matières (ou de la table des index)

Réalisation de logiciels pour l'analyse et la comparaison des composantes non textuelles : lettrines, bandeaux, ornements, etc. à l'aide des techniques actuellement utilisées en indexation d'images

Intégration de ces logiciels dans la banque de données d'images.

Avec des algorithmes très sophistiquées de reconnaissance des formes, on pourra faire des analyses du textes comme la possibilité de calculer les fréquences des mots (en fait des formes de mots) dans un texte.

4. La recherche et la navigation dans DEBORA

Il s'agit de la réalisation d'un serveur de documents numérisés installé dans la bibliothèque possédant les collections concernées, et d'un client destiné aux postes de consultation distants. Cet ensemble client-serveur doit être doté de possibilités d'accès aux documents eux-mêmes ou à des parties de documents. Grâce aux techniques de reconnaissance de formes , le traitement des documents en mode image est possible ainsi que le repérage des parties non textuelles de ces documents. Pour les ouvrages possédant des tables des matières et/ou d'index, l'obtention de ces données en mode caractère doit également permettre de définir des accès à des parties précises du document concerné, dans la mesure où le procédé de numérisation a permis d'établir les liens nécessaires. L'accès à des parties du document repose sur des procédures de reconnaissance de formes, notamment en ce qui concerne les éléments graphiques.

DEBORA repose essentiellement sur le système SGBI[9], le Système de Gestion de Base d'Images réalisé par la Maison de l'Orient Méditerranéen de l'université Lumière – Lyon 2 pour gérer son fonds d'images en archéologie.

Les caractéristiques de SGBI sont les suivantes :

Les informations stockées sont des images accompagnées d'informations complémentaires (description, indexation - éventuellement sous différents points de vue, caractéristiques de l'image et notamment les conditions de production).

Des collections d'images peuvent être constituées autour de caractéristiques communes (même ouvrage, même thème, même caractéristiques, etc.).

La navigation dans la banque d'images est arborescente.

L'interrogation est multicritère.

La gestion des droits d'accès à l'information est directement assurée par les utilisateurs.

[9] http://web.univ-lyon2.fr/SGBI/SGBImenu.html

L'accès est possible à travers un certain nombre d'interfaces normalisées et notamment à travers les navigateurs WWW. Par cette diffusion sur le Web, elle assure un véritable élargissement du public ainsi qu'un augmentation des leur de consultation.

Pour exploiter pleinement la richesse de DEBORA, l'usager doit avoir à sa disposition plusieurs méthodes complémentaire de recherche d'information et de navigation: requêtes à base de mots-clés (sur les champs définies de la métadonnée), navigation suivant les liens hypertextes, recherche sur des parties du documents. Toutes ces méthodes doivent être combinées dans une interface utilisateur restant simple et conviviale. Le démonstrateur de DEBORA sera construit selon une architecture client/serveur et respectant certains standards tels que Z39.50, le protocole http ou l'usage de java offrant ainsi une solution ouverte et une meilleure "**interoparab**" avec d'autres logiciels.

5. Le Poste de lecture assisté par ordinateur

L'interface de DEBORA serait développée en Java, elle devra comporter des fonctions de recherche reposant sur la reconnaissance des formes (recherche effectuée à partir d'une forme (et non d'une chaîne de caractères) identifiée dans le document, par exemple dans l'index, la table des matières ou quelque part ailleurs), des fonctions de recherche personnelle (telles que prise de notes, téléchargement d'extraits, établissement de liens) et de travail collaboratif (Technologie CSCW – programme ARIADNE[10]). Les usagers peuvent facilement annoter, copier, partager en groupe ces textes numériques. Ils permettront ainsi de nouvelles méthodes de recherche d'information et de travail entre les usagers.

L'objectif de cette partie de DEBORA est de donner à l'usager des outils performants permettent une lecture à la fois savante et collaborative en respectant les caractéristiques suivantes:

Performance : Bien qu'accessible par le Web, DEBORA doit permettre l'usage de questions complexes faisant appel aux résultats de questions précédentes et facilitant la définition d'une certaine stratégie d'interrogation.

Lecture savante : L'usager doit pouvoir intervenir sur les documents en associant à ceux-ci des annotations ou d'autres documents obtenus par ailleurs, pour se constituer une sorte de fonds personnel

Outils coopératifs : Partage d'information entre usagers, Echange d'information avec le bibliothécaire,

Définition et implémentation d'outils d'aide à la formulation et à la reformulation des requêtes

6. Une analyse des usages

En concertation avec l'équipe qui développera le Démonstrateur DEBORA, une équipe de chercheurs développera des études d'usage du logiciel DEBORA prenant en compte les besoins, attentes et demandes des usagers ainsi qu'une connaissance des pratiques.

Les usagers sont ici soit des producteurs (bibliothécaires, documentalistes, conservateurs, etc. qui créent ou enrichissent, avec l'outil DEBORA, un fonds d'ouvrages Renaissance), soit des utilisateurs (chercheurs, graphistes, imprimeurs, etc. qui utilisent DEBORA pour accéder à un fonds numérisé). Un autre objectif de cette phase est d'évaluer l'utilité réelle du système pour des groupes d'usagers dans leur activités professionnelles.

---

[10] http://www.comp.lancs.ac.uk/computing/staff/dmn.html

7. Analyse des Coûts

L'objectif de cette partie de DEBORA est d'assurer un suivi périodique de la numérisation dans toutes ses étapes dans le but d'évaluer les coûts mis en jeu :

préparation des collections,
repérage et traitement des éléments d'accès à des parties d'ouvrages (table des matières, tables d'index ou autres),
numérisation elle-même, validation, établissements des liens.

L'objectif final est de proposer des scénarios de passage du système prototype à une plus grande échelle. Les bibliothèques qui ont des projets de numérisation pourront ainsi évaluer les charges qu'ils représentent et disposer d'éléments de décision nécessaires.