World Library and Information Congress: 71th IFLA General Conference and Council

"Libraries - A voyage of discovery"

August 14th - 18th 2005, Oslo, Norway

Conference Programme:
http://www.ifla.org/IV/ifla71/Programme.htm

juillet 27, 2005

Code Number: **154-F**
Meeting: **97 Newspapers**

10 milliards de mots : Le projet « journaux de 1800 à 1900 » de la British Library Quelques règles pour la numérisation de journaux à grande échelle

Jane Shaw
The British Library
London, UK

Traduction *: Annie MILHAUD*
Bayard-Presse, France
Annie.milhaud@bayard-presse.com

Résumé

La British Library a décidé que la clé du projet était une couverture chronologique exhaustive de la totalité du dix-neuvième siècle, et, en tant que détenteur de la « collection maître », le défi réel serait de convertir en une ressource en ligne recherchable un grand volume de texte, sachant que « il y a très peu de chose en termes de masse de contenu ».[1]
A mi-chemin, les enseignements acquis grâce au projet « journaux de 1800 à 1900 » (BN) montrent que, pour numériser un grand volume de journaux historiques avec la meilleure qualité possible, il est nécessaire de prendre le temps nécessaire pour connaître les caractéristiques du matériau de base et de fournir les ressources convenables à l'équipe.
Il existe des standards définis et reconnus pour la numérisation de microfilms, mais ils ne sont pas toujours suivis[2]*. La British Library a donc entrepris d'établir quelques standards de microfilmage pour la numérisation de journaux à grande échelle et des règles de bonne pratique.*
Des problèmes avec le matériau de base ainsi qu'avec le processus de numérisation ont conduits à certaines décisions. Ces décisions sont les suivantes :

- *Mettre de côté les volumes en mauvais état ;*

- *Prendre les conditions de contrôle et d'appréciation du matériau de base comme repère ;*
- *Refilmer comme point de départ à la numérisation, filmer une page par cadre pour assurer une visualisation cohérente ;*
- *Ne numériser qu'à partir de films pour des raisons de rapidité, de cohérence et de coûts ;*
- *Intervention humaine pour aider les vérifications d'état, le collationnement page par page et le balisage simplifié des articles ;*
- *Solutions logicielles Open Source qui pourront être réutilisées.*

Les autres points importants considérés sont la création d'une interface d'accès aux journaux, combien de métadonnées pourraient être ajoutées, et quels types de recherches pourraient être proposées.

Introduction

Cet article est le résultat d'une série de rapports commandités par notre investisseur le JISC (comité commun des systèmes d'information britanniques) et de l'expérience du développement du **«Projet Journaux britanniques 1800-1900 » (BN).** L'article tente de répondre à quelques questions de base sur les processus techniques et les modalités pratiques impliqués dans la création d'un volume important de journaux recherchables en ligne à partir d'images de microfilm. Devra-t-on numériser la totalité du contenu du journal, y compris les publicités, les images, où seulement une sélection d'articles ? Quels outils de navigation seront-ils disponibles ? ; (les lecteurs pourront-ils « tourner les pages », y aura-t-il une recherche par mots-clés ?) Quel est l'impact d'utiliser un microfilm produit pour un usage (c'est-à-dire : la conservation) vers un autre usage (la numérisation) ? Comment la sélection préalable, les tests et les comparaisons des matériaux de base permettent-ils un accès en ligne sans obstacle, qui était auparavant difficile ?

Cet article expose la planification réaliste du projet. Il décrit notre prise de décision et les composantes économiques du microfilmage par ~~à~~ des standards de haute qualité techniques de façon à produire des images de haute qualité numérique et un OCR amélioré.

Contexte

Dès 2004, la British Library s'est assurée le soutien du JISC [3]. Sous le Programme de Numérisation, fondé avec une dotation de 10 millions de £ de la Comprehensive Spending Review, le JISC a autorisé un petit nombre de projets de numérisation à grande échelle qui permettront d'apporter des bénéfices significatifs aux communautés britanniques pour l'enseignement supérieur et la formation continue, l'un d'entre eux est le projet journaux britanniques 1800-1900 (BN).
Les critères de sélection du JISC pour accorder des fonds sous le Programme Numérisation furent :

- Les matériaux devront être d'un intérêt pluridisciplinaire large et former un ou des ensembles cohérents.
- Un petit nombre de projets à grande échelle seront financés qui n'auraient pas pu exister sans un investissement de cette taille.
- Les matériaux devront être entièrement compatibles avec l'environnement informatique habituel.

- Les matériaux devront remplir les conditions assurance-qualité standards les plus rigoureuses, et représenter une valeur pour la communauté de l'enseignement des 16 ans et plus.

Le projet a été financé pour développer les produits suivants : scanning du contenu total du microfilm, balisage d'article et extraction de page ; OCR des images de la page ; et la production des métadonnées nécessaires. Les objectifs principaux sont de numériser jusqu'à 2 millions de pages de journaux britanniques nationaux, régionaux et locaux, la majorité à partir de microfilm neuf et d'offrir un accès à cette collection via une interface sophistiquée de recherche et de navigation sur le web.[4] Ceci comprendra les noms, les dates, les annonces nécrologiques, les publicités, les mises en perspective régionales et locales des nouvelles nationales.

Buts du projet et comment ils ont guidé la sélection des titres de journaux

Le but général du projet :

- Fournir un contenu massif de journaux historiques sur le Web pour des recherches en plein texte par les communautés académiques ;

Ainsi que le but principal,

- Numériser jusqu'à 2 millions de pages de matériaux libres de droits en Grande Bretagne, journaux régionaux et locaux, la majorité à partir de microfilms neufs et d'offrir un accès libre à cette collection via une interface sophistiquée de recherche et de navigation sur le Web ;

N'ont pas changé l'année dernière. Le plan du projet diffère néanmoins du dossier initial dans les domaines principaux suivants ;

1. La proportion de nouvelles prises de vues a augmenté de 50% à 90%, pour permettre la cohérence des images.
2. Microfilmer une page par cadre pour optimiser la numérisation.[5]
3. Introduction d'une équipe assurance-qualité interne pour préparer et réparer les volumes, recueillir les métadonnées de publication et d'état, filtrer les doublons, les variantes, identifier les pages, les éditions manquantes et la dernière édition datée au début du projet.
4. Placer un échantillon d'utilisateurs académiques au cœur du projet pour diriger la sélection des journaux et conseiller la conception du site web.[6]
5. Introduction de deux pilotes pour surveiller les caractéristiques physiques des journaux du dix-neuvième siècle, pour approuver une méthodologie pour la fourniture de microfilm et pour confirmer que les spécifications produisent le produit final désiré, y compris la qualité des images et les résultats de l'OCR.[7]

Contraintes de sélection

Le dossier d'origine incluait une liste préliminaire de nombreux titres, au moins 160 ; répartis entre des quotidiens et des hebdomadaires nationaux londoniens ; quotidiens et hebdomadaires régionaux anglais; journaux des pays membres du Royaume Uni (écossais nationaux et régionaux, gallois, nord-irlandais) et des « sous ensembles spécialisés ». Pour des raisons de droits d'auteurs et pour rester dans le cadre du projet initial, seules les dates entre 1800 et 1900 ont été retenues. Néanmoins, au début de la sélection, des contraintes supplémentaires sont apparues. Les propriétaires de titres incorporés[8] dans des titres encore vivants auraient pu s'y opposer même si les publications d'avant 1900 sont clairement libres de droits. Les propriétaires de titres encore vivants pourraient numériser ou avoir un plan de

numérisation de leurs archives (par exemple : Le Guardian, Le Daily Telegraph) et il serait anti-économique de reproduire leurs efforts.

Etonnamment, il y avait très peu d'information sur le nombre de pages disponibles par titre, et pour tenir le planning du projet, il fut décidé de démarrer avec un pilote d'un sous ensemble spécialisé séparé comme les Chartistes suivi par un premier lot de travail incluant des titres évidents (par exemple : L'Examiner, Le Morning Chronicle, Le Graphic). En même temps, commença un audit de la pagination et de l'état sur les futurs candidats potentiels pour une sélection à partir de la liste préliminaire.

Nonobstant les contraintes précédemment citées, l'échantillon d'utilisateurs décida quand même d'estimer la totalité des titres de la liste d'origine (la longue liste) dans la perspective d'un usage potentiel par la communauté de l'enseignement supérieur.

A partir de la liste priorisée de l'échantillon d'utilisateurs, une « liste souhaitée » se dégagea et elle fut subdivisée en lots cohérents ou lots de travail.

On sélectionna quatre lots de travail pour arriver approximativement à 2 millions de pages au total dans les dates requises. Les lots de travail correspondent à un mélange logique de couverture large du Royaume Uni, de nationaux et de régionaux. Le lot N° 1 comprends le lot pilote (le sous ensemble Chartist), plus trois titres nationaux –un quotidien, un dominical et un hebdomadaire. Le lot N°2 étend la couverture en incluant 3 titres régionaux, du nord, du grand sud ouest et du centre de l'Angleterre. Le lot N°3 continue la presse nationale avec un dominical, un quotidien, introduit l'Ecosse et l'Irlande et continue avec la presse régionale anglaise, le lot N°4 continue à étendre la couverture du Royaume Uni avec l'Irlande, l'Ecosse et le Pays de Galles et met en valeur les régionaux anglais.[9]

La consultation en ligne et les besoins des utilisateurs

La pertinence par rapport aux besoins des utilisateurs actuels ou potentiels a été déterminée non seulement par le choix d'un échantillon académique pour effectuer la sélection, mais aussi validée par un questionnaire en ligne.

En février 2005, eût lieu une consultation en ligne auprès de la vaste communauté académique, plus spécialement sur les titres qui seraient inclus dans le projet BN, et aussi pour s'assurer que certains titres seraient inclus si plus de fonds étaient alloués, soit pour étendre le projet BN, soit pour poursuivre de nouveaux projets.

195 personnes ont répondu, la majorité d'entre eux étaient des bibliothécaires et des moniteurs travaillant principalement dans des universités et des collèges FE, quelques uns étaient chercheurs, étudiants, dirigeants ou enseignants. Etonnamment, 13% des réponses vinrent des USA.

Le questionnaire demandait aux utilisateurs de classer par ordre de priorité de numérisation (un=désapprouve fortement, cinq=approuve fortement), pour les titres de la longue liste du dossier. De plus, nous leur demandions de faire des commentaires ou d'ajouter tout autre titre qu'ils auraient voulu et qui ne figurait pas sur la liste.

D'une façon générale, il apparût clairement que les réponses approuvaient l'approche de large couverture pour le Royaume Uni et la méthodologie adoptée (c'est à dire un cadre de titres

nationaux et une couverture de tout le pays avec la largeur et la profondeur formant une clé virtuelle d'accès aux journaux provinciaux tous medias confondus). Il apparût aussi clairement que l'omission de journaux d'Irlande (Eire) posait problème.[10] Le JISC reconsidère cette situation.

Quelques portraits de titres de journaux sélectionnés

Morning Chronicle : Quotidien londonien, publié par John Black, le jeune Charles Dickens y fût reporter, et Thackeray y travailla en tant que critique d'art.

Reynolds Newspaper : vendu à plus de 350 000 exemplaires au début des années 1870. A l'origine un journal radical, il resta sous le contrôle des frères Reynolds jusqu'en 1894.

Poor Man's Guardian : Fondé par henry Hetherington en 1831 pour encourager la cause du suffrage universel et le mouvement syndical. Les bureaux furent perquisitionnés en 1835, les presses saisies et détruites.

Corbett's Weekly Political register : William Corbett fonda ce journal en 1802 pour soutenir sa carrière parlementaire, initialement d'apparence Tory, mais devint progressivement plus radical.

Birmingham Daily Post : Le Birmingham Daily Post fut lancé en 1857 par l'irlandais John Frederick Feeney comme journal de 4 pages du lundi au vendredi au prix d'un penny. Il existe encore.

Belfast newsletter : fondé en 1737, pendant presque 200 ans la famille Henderson fut étroitement associée au journal. Il est toujours publié aujourd'hui.

Droits d'auteurs

La British Library est en négociations permanentes avec les éditeurs, y compris les éditeurs de journaux, sur bon nombre de sujets portant sur le droit de la propriété intellectuelle couvrant le cycle du traitement de l'information, de l'acquisition aux accès et à la conservation.

La politique de la Bibliothèque est de procéder avec l'accord des détenteurs des droits et de leurs représentants. Dans le cas des journaux, des discussions récentes avec les éditeurs de journaux et le conseil légal mis à jour de la bibliothèque signifie que, pour ce projet, le point de départ est qu'aucun journal de moins de cent ans ne sera numérisé pour un accès donné à l'enseignement supérieur et à la formation continue.[11]

Comment le projet fut modelé par la résolution de problèmes

Les livrables

Le projet délivrera 2 millions de pages, totalisant approximativement 10 milliards de mots à partir de journaux britanniques de 1800 à 1900 ; cela revient à 40 titres à peu près.

Le processus de numérisation délivrera une matrice d'archivage pour chaque page, au format TIFF, version 6.0. Ces fichiers seront scannés effectivement à la résolution de 300 dpi, 8 bits à niveaux de gris.[12]

Les images seront créées après le balisage des articles et l'OCR ; les copies de service seront délivrées en hybrides à niveaux de gris, en TIFF version 6.0, et en JPEG.

Beaucoup de titres de journaux ont été filmés sur des films acétate et ce, avant que les standards de conservation nationaux soient adoptés systématiquement (1990). Afin d'assurer des images de haute qualité et économiser à long terme, la meilleure méthode fut de contrôler la qualité du microfilm d'origine et d'aider à l'élimination de beaucoup de questions de postproduction. Réduisant ainsi la charge de travail de Questions /Réponses et donnant au fournisseur un comparatif uniforme et une apparence cohérente.[13] On estima comme étant le plus efficace et de meilleure valeur à long terme pour le scannage du microfilm le travail additionnel de collation des doublons et des pages manquantes, le travail de contrôle de l'état et la collecte des métadonnées au niveau de la publication.
Si les organismes ne désirent pas refilmer et décident d'utiliser leur stock existant, soit par choix politique, soit pour gagner du temps, il faut tout de même envisager un coût, à la fin du processus et davantage sous le contrôle des fournisseurs.

Il n'était pas question de débrocher les journaux et lorsqu' arrivèrent au département microfilm de la BL les caméras microfilm Zeutschel de dernière génération, le projet bénéficia de la nouvelle technologie en refilmant les volumes reliés utilisant des barrettes de reliure, et non sous verre. La bibliothèque n'essaie pas de corriger les erreurs de l'imprimeur ou les zones de texte manquant. Les images en fac-similé stockées en tant que fichier maître à niveaux de gris peuvent aisément être réutilisées et /ou désassemblées pour de futurs projets.

L'équipe du projet BN décida qu'il serait intéressant que l'image soit vue en contexte, comme dans la publication originale, avec une image pleine page.[14] De plus, les trois critères de sélection : séries complètes de chaque journal, vaste couverture du Royaume Uni et découpage de la totalité du 19è siècle assureraient un accès substantiel aux nouvelles par une recherche « simple ».

Le projet ne porte pas sur la numérisation de publications du 18è siècle ou d'éditions variantes, non plus que sur les riches ressources de la presse coloniale britannique. Seule la dernière édition en date de chaque publication est filmée plus quelques suppléments occasionnels par exemple :The Graphic Stanley number, 30 avril 1890. La BL a pris ces décisions pour éviter de doublonner au maximum et pour inclure le plus possible de pages originales dans notre total de 2 millions. L'échantillon d'utilisateurs créé par le projet décida aussi qu'il souhaitait avoir le plus possible de journaux différents. Ceci de préférence aux variantes d'éditions.

D'autres projets de numérisation à la BL, ont permit d'être conscient des enjeux de la conservation et de la numérisation de journaux et de connaître les stratégies pour les numériser. Cet article décrit comment l'équipe du projet BN prit une décision bien informée sur la stratégie adéquate pour ce projet particulier. La densité des textes du 19è siècle rend l'identification automatique des coupures entre les articles plus difficile et requiert un équilibre entre la génération automatique de métadonnées et une intervention humaine. Pour avoir travaillé pendant des années avec un fournisseur, la BL réalisa que l'intelligence humaine donnerait le meilleur résultat sur le plan de la qualité, et conçut donc le projet autour d'un système informatique assisté par l'intelligence humaine tout au long du cycle de traitement.

Compte tenu de la taille du projet, le coût de numérisation par page est estimé à :

100% dupliqué de film existant = 75p la page
100% nouveau film = 98p la page

D'ailleurs, cela correspond au budget d'origine, (1£ la page).

Caractéristiques physiques du matériau de base, problèmes sous-jacents et conséquences de ceux-ci sur le processus de numérisation.

Stocks de journaux de la BL
Comme la plupart des grandes collections de journaux historiques, les stocks de la BL sont majoritairement reliés en volumes. On considérait généralement que c'était le meilleur moyen d'assurer une conservation à long terme. A cause de la nature des reliures, certains volumes sont reliés très serrés et d'autres ont commencés à se désintégrer, présentant des dégâts sur les bordures des pages, et encore plus sur les derniers journaux du volume.

Deux problèmes viennent de la méthode de reliure des journaux, le texte étant relié dans la charnière et la courbure du papier vers la charnière. Le texte peut être perdu pendant la prise de vue à cause de l'ombre du petit fonds, le texte peut être distordu, et ceci peut affecter le pourcentage de la dernière colonne pouvant être OCRisé. D'autres exemples de problèmes devant être résolus pour l'OCR comprennent ; problèmes de « transparence » (on peut voir le texte imprimé au verso de la page à travers le papier), souvent dus à la faible qualité du papier, encre trop forte, ou un mélange des deux, erreurs de l'imprimeur dues à un glissement du papier ou un papier plié / rainé provoquant des cassures dans la police de caractères durant l'impression. Les moteurs d'OCR pourraient avoir des problèmes à surmonter ces erreurs et à reconnaître des mots entiers.

Une complication supplémentaire vient du fait que les éditions variantes sont généralement reliées ensemble et différents titres peuvent être mélangés dans le même volume. Pour ce projet, utilisant seulement la dernière édition datée, le fonds devait être épuré des éditions variantes et des doublons. Nous avons décidé d'épurer dans la phase de préparation initiale, plutôt que plus tard avant ou après le scannage.

Les journaux du début du 19è siècle révèlent une quantité significative d'erreurs d'imprimeur (par exemple une page rainée pendant l'impression provoque une perte de texte lorsque la rainure est supprimée). Il y a aussi l'apparence du mot coupé, qu'on trouve principalement à la fin de textes de journaux sur 2 colonnes.[15] Les formats et la structure changent fréquemment et de façon inattendue, d'une publication de 4 pages à une de 6 pages, l'ordre du contenu souvent inversé, la position de l'éditorial comportant la signature de l'éditeur, des changements brusques de style d'une presse bon marché à un grand format. Ils représentent une ressource très riche pour l'histoire de l'imprimerie et le développement des opinions et des discussions radicales apparaissent à travers les innovations de typographie et de mise en page. Ils représentent aussi un sérieux défi à l'OCR, qui préfère un texte mis en page simplement, une tonalité neutre et une impression plus large. Nous n'avons pas accepté de perte substantielle de texte venant de l'ombre du petit fonds ou un film de densité inégale venant de vieilles pages laminées.

Stocks de microfilms de la BL

La politique de la BL est d'utiliser les films existants et de compléter les manques en négatifs maîtres ou duplicatas, pour éviter une double manipulation des collections. Néanmoins, une grande proportion des titres sélectionnés par le projet BN est sur ancien film acétate. Pire est l'état du film, plus long est le processus de copie, ce qui réduit la productivité et augmente les coûts unitaires [16]. De plus, le microfilm existant est trop variable pour une production constante, comprenant les film acétate, le polyester datant d'avant et après le standard National de conservation et les prises de vues avec les nouveaux appareils depuis 2004. Il pourrait donc y avoir des problèmes supplémentaires de qualité et des coûts de post production dus à ce mélange de films non-conformes. Enfin, la BL a décidé que de diriger des processus de travail complexes dus aux différentes vitesses de réalisation nécessaires à l'adaptation au mélange de films, pouvait conduire à des délais inacceptables et à compromettre la qualité. Le stock de films existants pouvait inclure des copies et des variantes de publications pour n'importe quel titre relié, ce qui était historiquement laissé au hasard et que la politique habituelle est de filmer la dernière édition datée, en accord avec le projet ; Un opérateur de scan devrait apprendre quelles parties du film scanner et quelles parties laisser de côté. Pour surmonter ces problèmes, la BL a décidé, que, pour le projet de numérisation, lorsque l'état et la reliure des matériaux le permettent le service Microfilm de la BL travaillant avec les nouveaux appareils de prise de vues filmerait la plupart des journaux en interne.

La plupart des pages ont été pincées pour les maintenir à plat pendant la prise de vue. Cela peut sembler bizarre et apparaître comme du noir si on présente un fac-simile de la page aux utilisateurs. Presque chaque bobine a une collure et cela, à cause des pages manquantes identifiées lors du stade de contrôle ou des reprises dues à divers problèmes techniques. Voila quels ont été les problèmes.

Source d'acquisition des données

ETAT du microfilm	ETAT du journal	DECISIONS				
Microfilm de mauvaise qualité	Journal papier en bon état	Refilmer	Dupliquer	Scanner	Mettre en évidence	
Microfilm de bonne qualité	Journal papier en mauvais état	Dupliquer	Scanner	Mettre en évidence		
Microfilm de mauvaise qualité	Journal papier en mauvais état	Laisser de côté	Sélectionner un autre titre			
Microfilm de bonne qualité	Journal papier en bon état	Dupliquer	Scanner	Rejeter	Refilmer	Rescanner
Microfilm de bonne qualité	Journal papier en bon état	Dupliquer	Scanner	Mettre en évidence		

Réduction et résolution

Lors des prises de vues, nous avons utilisé la plus faible réduction possible, par exemple Le *Pall Mall Gazette* a été filmé à une réduction de 12. De plus, il est important de conserver la même réduction sur toute une bobine. Néanmoins, cela n'a pas toujours été possible car certains titres comportaient des illustrations en dépliant incluses dans les pages.
Le large spectre de titres sélectionnés implique que quelques uns des volumes les plus grands ont été filmé à une réduction plus élevée, jusqu'à 20 dans certains cas. Il est intéressant de noter que nous n'avons utilisé que des réductions totales (c'est-à-dire pas de demi réductions). Malgré les soucis causés par les taux élevé de réduction, nous obtenons d'assez bons chiffres de résolution. Dans certains cas, nous obtenons des écrans de lecture de 140 LPM (ligne par millimètre), ceci est souligné plus avant par de stricts contrôles de densité des écrans de lecture.[17]

Adjudication

Etant donné que le contrat de numérisation pouvait dépasser 153 376 £, nous étions soumis aux règles d'adjudication du JOUE, processus lent impliquant beaucoup de choses à apprendre.[18] Le processus entier prit approximativement un an, de la demande de propositions jusqu'à la signature du contrat.

Tous les fournisseurs potentiels proposèrent d'effectuer plus tard une conversion de haut niveau. Ils pouvaient tous scanner et OCRiser les pages, scanner l'OCR et surlignant les termes recherchables, scanner, OCRiser les surlignages et la visualisation de la page, mais un seul pouvait scanner, baliser, OCRiser, vérifier et corriger le balisage, réindexer les zones de titre si nécessaire dans un budget raisonnable. Aucun ne présentait la possibilité d'un fichier OCR complètement « propre ».

Enseignements de la première année de l'opération

Que ferions nous de la même façon ou différemment si nous devions recommencer.

- La clé, c'est la préparation : l'inspection en profondeur et le constat de l'état des caractéristiques physiques des matériaux de base pour mettre en place un comparatif pour les questions à venir.
- Le nécessaire décompte des pages et le désherbage des doublons et des éditions variantes pour produire des lots de travail et progresser vers le total final.
- Le format d'une publication étalée dans le temps change souvent sur un siècle, le décompte des pages est donc vital pour en comprendre la structure.
- Placer un échantillon d'utilisateurs au cœur du projet, pour servir d'ambassadeurs et s'approprier le projet.
- Définir très tôt des critères de sélection pour guider les délibérations de l'échantillon d'utilisateurs.
- Prendre en compte très tôt les enjeux de la propriété intellectuelle, choisir une approche solide et constante, maintenir le dialogue avec l'industrie des journaux.
- Conduire une consultation en ligne avec des communautés d'utilisateurs durant le processus de financement. Une grande liste de titres potentiels signifiait que nous ne pouvions pas nous recentrer.

- L'échantillon d'utilisateurs a présélectionné nos 2 millions de pages potentielles, puis la consultation en ligne ne proposa pas de nouveautés ou de titres non testés, à part les journaux d'Irlande , le travail pu alors commencer.[19]
- Considérer les titres « laissés de côté » trop fragiles à filmer, comme comparatifs pour l'avenir. Tolérer et accepter des intervalles dans les publications à long terme, et ne pas chercher à combler ces intervalles avant longtemps.[20]
- Considérer la qualité de votre microfilm [refilmer une partie de votre contenu ajoute de la valeur à long terme] de façon à faire face à beaucoup de risques liés à l'hétérogénéité des originaux et à la qualité variable du microfilm, ce qui ensuite aidera à l'acquisition de l'image.[21]
- La collection de journaux du 19è siècle de la BL est en meilleur état que ce que l'on avait prédit, moins de 2% inapte au microfilmage. Néanmoins, à cause du processus mécanique de scannage (trop lent et trop éprouvant pour un matériau de base fragile et vulnérable), on décida de numériser à partir du microfilm.
- Organiser les attentes futures par l'approbation en ligne de listes de titres par les communautés d'utilisateurs.
- Imaginer des cibles de production ne fonctionne pas, des cibles concrètes basées sur du travail réel pour une date précise doivent être régulièrement revues et redéfinies, elles sont nécessaires à la prévision de tendances.
- Utiliser des « doubles » ou des deuxièmes prises de vues intentionnelles comme technique d'assurance qualité, après avoir mesuré ceci en relation avec des chaînes de travail effectives.[22]

Quelques points sur la numérisation

Quels standards ?

Nous respectons certains standards très rigides par exemple : METS, un standard d'encodage et de transmission de métadonnées et le standard BSI pour le microfilmage.[23]Nous croyons obtenir des métadonnées bien formées en utilisant des personnes travaillant sur des pages originales et à partir de microfilms de bonne qualité, ce qui fournit l'authentification.

La BL exige les métadonnées descriptives du Dublin Core conformes au British Library Application Profile (BLAP) avec les éléments encodés en XML et finalement un fournisseur de données OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) comme moyens de remplir nos exigences d'interopérabilité.

Niveaux de métadonnées

Il y aura 4 niveaux structurels, le titre, la publication, la page et l'article. Au niveau du titre, tout changement dans le titre, date de publication, type et sous collection, sera capturé par l'équipe Questions / Réponses de la BL. Les métadonnées de la publication comprennent, le numéro de publication, la date normalisée, le nombre de pages et le numéro d'identification de la bobine. Les métadonnées additionnelles fournies par le fichier XML incluent :

- Date de publication au format ISO standard
- Taux de qualité, (A, B, C) sur l'état du matériau d'origine
- L'énoncé de Copyright de la BL
- L'année du Copyright
- Les crédits de conversion

- Balises pour les crédits photo et infographies
- Balises pour les noms d'auteurs

Hybrides à niveaux de gris contre image bitonale

Les pages scannées en niveaux de gris 8-bits semblent plus douces, avec des détails plus fins et des gris plus subtils, ce qui est particulièrement bénéfique pour le texte et les illustrations et donne des images d'archives plus conformes à l'original. L'apparition de points noirs est réduite, la correction de désalignement et la précision de l'OCR est supérieure avec une résolution de 300dpi.

Le projet développe un nouveau produit, une image hybride, pour optimiser à la fois la qualité des images et du texte. Les zones de texte sont converties en images bitonales, affinées en utilisant une version améliorée de IZ-image, et sauvegardées en niveaux de gris en même temps que les images d'illustrations en niveaux de gris. Ces hybrides reconstruits sont les copies de consultation qui seront accessibles sur le site Internet. Le principal désavantage de cette méthode est un accroissement par 10 de la taille de l'image, ce qui pourrait augmenter les coûts de stockage et poser des problèmes aux utilisateurs de modems à 56K. Dans la mesure ou les futurs utilisateurs auront accès au haut débit et ou les coûts sont acceptables, l'équipe projet accepta le concept d'hybrides à niveaux de gris.[24]

Il faut, bien évidemment, une exacte concordance entre le fichier matrice et les copies de consultation.

Recadrage

Nous avons ressenti la nécessité d'équilibrer la lisibilité et l'exhaustivité et que les fichiers matrice devaient représenter aussi précisément que possible le contenu visuel des pages originales.

Une image objet ou un fichier matrice d'archive sera généré pendant le scannage initial à partir du microfilm en éliminant les bordures noires de toute l'image objet cadre par cadre. L'image de la page, devant être utilisée comme copie de consultation pleine page, est générée en recadrant l'image objet vers une bordure uniforme autour du contenu de chaque page, ôtant ainsi tout clip ou témoin. Toute impression d'échelle ou de changement de taille de page sera perdue à cause de cette uniformisation, mais d'un autre côté, cette méthode devrait produire des images optimisées pour le web.

Surlignage de mot-clés

Le projet ne veut pas de surlignage non pertinent de mot qui entraînerait une mauvaise expérience pour l'utilisateur. Cela sera reparti sur la page et est en cours de développement.

Articles sans titre

Cela est fréquent dans les catégories telles que publicité, notices, nécrologies, etc. On a décidé que quand la catégorie ne pouvait pas servir de titre, un titre serait construit à partir des deux premières lignes de texte.

Correction de désalignement

L'œil humain corrige simplement. Les images sont corrigées jusqu'à ce qu'elles apparaissent alignées au superviseur, et comme toute les images sont inspectées avant l'OCR, c'est acceptable. De plus, les images en niveaux de gris peuvent être corrigées jusqu'à 10 degrés de désalignement, alors que les images bitonales, avec seulement 1 degré peuvent affecter la qualité de l'OCR.

Niveau article contre niveau page

Les images des pages seront constituées des articles avec les termes de recherche surlignés, une fois complète et en format pleine page. Les utilisateurs pourront visualiser les images de pages associées à n'importe laquelle de leurs recherches.

Recommandations pour les catégories

Avec une telle masse de mots recherchables, et un degré sans précédent de référencement croisé, le projet fournit aussi aux utilisateurs une catégorisation simple, permettant un accès aux documents pour enrichir la recherche par mots-clés.

Nos recommandations pour les catégories sont,

1. **Actualités (domestiques)** ; articles relatifs au Royaume Uni
2. **Actualités (étranger)** ; articles relatifs au reste du monde
3. **Publicités et Notices** : des petites annonces de vente, d'emploi, annonces légales. Différenciées de actualités par le format.
4. **Art et Culture Populaire** ; critique de livres, musique, théâtre. Poèmes et feuilletons. Articles relatifs aux foires et spectacles itinérants.
5. **Naissances, Décès et Mariage** ; annonces de chaque événement
6. **Nécrologie** ; article spécifique relatant la mort de quelqu'un avec sa biographie.
7. **Cour et Société** ; en relation avec la famille royale et l'aristocratie.
8. **Crime et Punition** ; reportages de différentes cours de justice, sessions et assises.
9. **Commerce** ; actualité des affaires et des transports, cotation des marchés et des bourses
10. **Lettres** ; habituellement écrites à l'éditeur et imprimées en parties ou en totalité.
11. **Sports** ; articles couvrant une large gamme de domaines sportifs
12. **Editorial / Commentaires** ; articles habituellement, mais pas exclusivement écrits par l'éditeur sur un ou des sujets particuliers.
13. **Divers** ; articles qui ne correspondent pas aux catégories précédentes
14. **Aucun** ; catégorie par défaut
15. **Illustration** ; sans légende

Les étapes du processus de production

Les simples étapes du projet suivi par la BL et qui, elle l'espère, améliorera l'accès aux journaux historiques sont les suivantes,

1. Examen de l'état : des stocks de matériau source pour comprendre les structures et identifier les problèmes sous-jacents. Les reliures serrées causant des ombres de petit fonds, comment manipuler les encarts et suppléments, dupliquer les publications

incluses dans les films, titres répartis dans plusieurs bobines et mélangés avec d'autres titres.

2. Collationnement : Les résultats de l'examen ont été collationnés et utilisés pour aider au processus de sélection et au choix d'un fournisseur. Ces résultats comprenaient le nombre de pages par publication, les différences entre nationaux et hebdomadaires, les changements de structure au long d'une publication entière, la mise en page, les polices de caractères tout au long du siècle et l'état général du stock de microfilm.
3. Sélection : Une liste de titres a été sélectionnée selon des critères simples approuvés par les communautés d'utilisateurs.
4. Microfilm : Les volumes reliés serrés, les volumes en mauvais état et les microfilms en dessous du standard ont été traités au cas par cas soit pour mettre de côté un volume ou un microfilm en dessous du standard, soit pour travailler dessus.
5. Microfilm : Un système d'assurance qualité a été construit en utilisant les résultats des examens de démarrage du projet et des seuils de tolérance établis pour les « laissés de côté », la clarté et la lisibilité.
6. Microfilm : On prit une décision sur la quantité à refilmer si on suivait pour de bon les niveaux de qualité retenus.
7. Définition des tâches de fond : Les hypothèses furent testées sur deux pilotes, une pour chaque flux de travail, toutes deux assez vastes pour donner des résultats valables.
8. Adjudication : De façon à permettre la croissance d'une masse critique de contenu de journaux historiques dans l'avenir, au niveau national et international, nous savions que nous devions fournir une solution technique Open Source.
9. Numérisation : Scanner à partir de microfilms ; génération d'images à niveaux de gris, contrôle qualité, nettoyage manuel ou pas, contrôle qualité, fichiers matrice d'archivage.
10. Numérisation : Utiliser l'affinage de caractères IZ-Image, charger la publication complète, baliser et catégoriser les articles, contrôle qualité.
11. Numérisation : lier les articles multi-pages, passer l'OCR et nettoyer les métadonnées de chaque champ, générer des livrables.

Coûts

Combien l'utilisation coûtera-t-elle à une institution ou à un particulier ? La proposition couramment prise en compte est que le projet vise un accès gratuit pour tous après le printemps 2007, l'utilisation sera donc gratuite.

L'avenir et le patrimoine

Le stock en ligne sera-t-il augmenté dans l'avenir ? En raison du Copyright et des droits de la propriété intellectuelle, il est vraisemblable que les projets futurs sélectionneront des matériaux de base plus anciens et libres de droits. Il y a un besoin urgent d'un registre ou d'un réseau d'information sur ce qui a été sélectionné pour la numérisation et sur quoi on a travaillé pour pouvoir comparer avant que d'autres bibliothèques entreprennent de numériser leurs collections.

Avec les données créées à partir de ce projet, la BL sera en bonne position dans les projets futurs pour prévoir quels seront les coûts par page puisque tous les coûts de notre projet ont été analysés.

Résumé

L'équipe du projet BN a passé l'année précédente à estimer chacun des titres de journaux sélectionnés et leurs microfilms associés pour la numérisation ; sélectionnant une liste finale de 40 titres, approuvés par une consultation en ligne des communautés académiques, et a choisi un fournisseur de numérisation. Les équipes internes ont préparé et filmé 650 000 pages depuis le mois d'octobre 2004 et sont en passe d'avoir atteint 1 millions d'images en septembre de cette année. Les équipes Unité microfilm (MU) et assurance qualité (QA) travaillent toutes deux sur des objectifs journaliers et hebdomadaires.[25]

L'année à venir (jusqu'à septembre 2006) verra un changement de cible, avec le démarrage à la fois des chaînes de travail de la numérisation et du développement du site Internet en parallèle avec le travail interne de préparation et de prise de vue.

Le projet continuera à suivre une méthodologie déroulante[26] à l'intérieur d'un sous ensemble de PRINCE 2, [27] selon laquelle nous apprenons en progressant sur le matériau de base et la technologie. Un échantillon d'utilisateurs ayant terminé sa tâche de sélection, a changé de rôle pour conduire et définir les besoins des utilisateurs potentiels, commenter le design du site Internet, avant la livraison d'une maquette de test au début 2006.

Une livraison par étapes du site Internet devrait commencer en septembre 2006.

La BL a aussi placé l'intervention humaine au cœur du projet en introduisant une équipe d'assurance qualité pour passer au crible les doublons et les éditions variantes, récolter les métadonnées de publication et d'état avant la prise de vue. De plus, en collaborant avec un échantillon d'utilisateurs académiques pour choisir les 2 millions de pages sur une possibilité de 200 millions qui seraient sélectionnés pour être numérisés.

La cohérence de cette approche va jusqu'au choix du fournisseur, dont la méthodologie d'extraction de contenu utilise l'intelligence humaine / assistée par l'ordinateur et qui a innové dans la catégorisation d'article et l'indexation par mot-clés. Ce que les utilisateurs auront :

- Recherche plein texte sur 2 millions de pages de texte de journaux.
- La possibilité de recherche sur un journal particulier par date.
- Comparaisons à l'identique du traitement d'un sujet par différents titres.
- Navigation avant et arrière dans une publication.
- Images de la publication originale à lire de la manière habituelle.
- Il devrait être possible de conserver et de construire des recherches, pour aider à l'enseignement et au travail collaboratif
- Visualisation des résultats de recherche au niveau de l'article dans le contexte de la page originale.
- La possibilité de rechercher sur les publicités, nécrologies, etc.
- La possibilité de télécharger des versions texte des pages originales.
- Une précision prévue de 80% de la conversion OCR.

On peut ajouter de la valeur à des matériaux de base variés dans un état moins que vierge, et même avec un original mal imprimé, on doit toujours tendre vers la meilleure image possible.

En conclusion, le projet BN est une initiative qui a déjà mis en évidence la compréhension qu'a la bibliothèque des multiples sujets que la numérisation fait apparaître. La BL est

impatiente de voir sa réalisation aboutir avec succès, et plus important encore, que son contenu trouve une vaste audience.

Jane Shaw
Juin 2005

---

[1] Voir OCLC
[2] Pour une description complète, voir : *RLG Guidelines for Microfilming to support Digitisation*, janvier 2003. Voir aussi, IFLANET, Newspapers section, *Microfilming for Digitisation and Optical Character Recognition*, décembre 2002.
[3] Le JISC a accepté de soutenir le projet d'avril 2004 à septembre 2006 pour un coût total de 2 022 131 £. Voir http://jisc.ac.uk

[4] BL Business Plan, 14 octobre 2003
[5] Il existe des densités variables dans l'image et l'exposition sur des films comportant deux pages par cadre
[6] Bien qu'un débutant total eût été un membre utile en tant que conseiller comme « candide » désirant simplement naviguer
[7] Reference : Jane Shaw, JISC *Development Programmes, January 2005 Progress report*
[8] Des journaux publiés au dix-neuvième siècle peuvent être incorporés à des journaux vivants
[9] Voir annexe 1, la liste des titres sélectionnés.
[10] Par exemple *« c'est un projet prévu de longue date, qui a la capacité de transformer la recherche sur cette période. Cela fera une grande différence dans l'enseignement de l'histoire dans les télé-universités. Idéalement, les journaux choisis offriront un bon éventail géographique, politique et chronologique. »*
« Etant donné que des journaux tels que le Limerick Chronicle ou le Cork Examiner ou le Dublin Papers du 19è siècle se trouvaient dans des villes faisant partie du Royaume Uni, pourquoi aucun de tels journaux ne sont inclus dans votre projet ? D'un point de vue d'historien, une telle omission est hautement illogique et produit une sélection non représentative. Y a-t-il une chance pour que les principaux journaux irlandais soient inclus dans ce projet très intéressant et hautement prometteur ? »
« Les journaux numérisés sont une ressource merveilleuse pour les département d'histoire. Ils facilitent la recherche indépendante et l'enseignement, que nous encourageons tous, à travers un format que les étudiants apprécient. C'est très spécialement bienvenu étant donné la pression exercée sur les livres et les périodiques en bibliothèque due à l'accroissement du nombre des étudiants ces dernières années. »
[11] La direction de l'échantillon d'utilisateurs a suggéré que « des trous dans le processus de sélection » pourraient être comblés par un accord de licence à plus long terme, tel que acheter à d'autres sources licenciées et en créant des relations avec des compagnies commerciales.
[12] C'est moins que les 400 dpi que la bibliothèque du Congrès recommande. La British Library a choisi 300 dpi parce qu'une plus haute résolution donne seulement une augmentation significative de la taille des fichiers, particulièrement pour les niveaux de gris et la qualité de l'OCR ne s'améliore pas au dessus de 300 dpi, et que la résolution d'écran est généralement entre 72 et 100dpi.
[13] Cela plaça aussi la responsabilité de la dernière assurance qualité précisément chez le fournisseur.
[14] Cette décision a été réexaminée et est exposée en détails dans les publications Numérisation
[15] C'est particulièrement vrai pour Cobbett's et l'Examiner pour lesquels la mise en page est sur 2 colonnes, les mots étant coupés à la fin de la ligne, plutôt que d'utiliser un blanc plus large pour imprimer le mot entier sur la ligne suivante. Comment le logiciel d'OCR prend-il en compte à la fois la mise en page et la césure des mots ?
[16] Le projet Burney Newspapers a prouvé qu'il était possible de scanner à partir de films acétate anciens, mais à des coûts élevés de production.
[17] La résolution utilisée est un minimum de 115 LPM (lignes par millimètre), et la réduction est déterminée par le document filmé, bien qu'on le filme à la plus faible réduction possible, selon la taille de l'original. Nous avons des variations de taille de page du format A4 au grand format. Le contrôle qualité est fait cadre par cadre pour la qualité de l'image, chaque bobine est testée en résolution et densité avec un microscope et un densitomètre. Nous utilisons aussi une fiche de contrôle qualité pour chaque bobine de film qui documente toute l'information technique sur cette bobine et pour toutes les suivantes.
[18] Le Journal Officiel de l'Union Européenne, on invite par annonce les entreprises intéressées à remplir et renvoyer un questionnaire. Une demande de propositions (RFQ) très approfondie, produisit 26 expressions d'intérêt desquelles on tira une short-list de 6 fournisseurs potentiels, finalement 3 sociétés atteignirent le stade final (BAFO).
[19] On pourrait arguer que la consultation aurait pu commencer plus tôt, avant le démarrage du pilote, mais cela fut impossible à cause des possibles problèmes liés aux droits de la propriété intellectuelle et réciproquement, on

pourrait situer le meilleur moment pour consulter de larges communautés après avoir démarré sur quelques une de nos prises de position.

[20] Jusque là, il y a eu très peu de volumes en mauvais état laissés de côté, moins de 2%, bien moins que ce qui était prévu dans le dossier initial. Nous avons pu redéployer le budget de conservation sur la fondation de l'équipe de Questions/réponses et refilmage.

[21] La BL a bénéficié de nouvelles technologies. De meilleures images venant de nouveaux appareils de prise de vue, l'éclairage est constant, le petit fonds des volumes reliés serrés est maîtrisé avec des socles pouvant être levés, et les pages ne sont pas filmées sous verre, juste clippées si nécessaire.

[22] C'est une exigence des standards de la BL, que la page devant être doublée soit filmée à une exposition normale, puis refilmée à la nouvelle exposition. Le fournisseur de numérisation doit effectuer le choix de la meilleure qualité d'image après le scannage.

[23] BS ISO 4087 (1991)

[24] Une idée couramment en discussion est la possibilité pour les utilisateurs de visualiser le texte OCR sans correction pour le réutiliser selon leur expérience d'apprentissage virtuel (VLE)

[25] Voir annexe 2

[26] C'est la méthodologie préférée du JICS pour leurs projets

[27] PRINCE 2 est une méthodologie de management de projet orientée produit.