Rapport de stage / septembre 2007

Diplôme national de Master

Domaine - sciences humaines et sociales

Mention - sciences de l'information et des bibliothèques

Spécialité - gestion et traitement de l'information spécialisée

Archivage du Web à la British Library

Leïla Medjkoune

Sous la direction de Ms Alison Hill

Curator, Web Archiving

The British Library

et de Monsieur Jean-Paul Metzger

Directeur des Études

ENSSIB

école nationale supérieure des sciences de l'information et des bibliothèques

Remerciements

Je tiens à remercier ici le Dr. Kristian Jensen, Head of British and Early Printed Collections, pour m'avoir offert l'opportunité d'effectuer ce stage et pour la confiance qu'il m'a témoignée.

Je remercie tout le personnel de la British Library pour son accueil chaleureux, et tout particulièrement John Tuck, Head of British Collections et le Dr. Richard Price, Head of Modern British Collections, ainsi que son équipe, pour leur bienveillance et leurs conseils.

Je suis reconnaissante au Dr. Clive D. Field pour ses encouragements et pour avoir pris le temps de me guider dans mes travaux de recherche.

Mes remerciements et toute ma gratitude vont à M. Jean-Paul Metzger pour m'avoir fait confiance et m'avoir soutenue durant ces deux années.

Je souhaite remercier l'équipe WAP et plus particulièrement, M. Ravish Mistry pour son aide précieuse et sa gentillesse et Ms. Alison Hill qui a su partager son expérience, m'a transmis son enthousiasme et est allée bien au-delà de son rôle au cours de ces quatre mois.

Enfin, je remercie mes amis pour leur présence et surtout ma famille pour son soutien inconditionnel.

Résumé

La British Library (BL), à l'instar d'autres institutions nationales, est engagée dans de nombreux projets liés à l'ère digitale, parmi lesquels figure le projet UKWAC. Six institutions majeures du Royaume-Uni y sont unies afin de préserver le patrimoine national en archivant l'Internet. Mon stage dans le service Web Archiving de la BL a consisté en la création de collections, m'offrant la possibilité d'appréhender chaque phase composant le processus d'archivage du Web, à commencer par la sélection.

Descripteurs :

British Library; BL; UKWAC; Archivage Web; Dépôt Légal; Droit d'Auteur; Collection; Blog; Internet

Abstract

As a national institution, The British Library is involved in many digital projects, as the Web Archiving project, UKWAC. UKWAC is a consortium of six UK key institutions which objective is to archive, make accessible and preserve selected websites of the UK domain. My placement at the Web Archiving section of the BL consisted in building two collections from the selection to the archive, enabling me to participate in each phase of this archiving project and to understand each of its issues.

Keywords :

British Library; BL; UKWAC; Web Archiving; Legal Deposit; Intellectual Property Rights; Copyright; Collection; Blog; The Web

Sommaire

Introduction

La British Library, en tant que bibliothèque nationale, participe à différents projets européens et mondiaux. Ces projets concernent de plus en plus le passage des bibliothèques à l'ère du numérique, du virtuel et de l'Internet. Confronté à de nouveaux défis, tant matériels (défis technologiques et financiers) qu'humains (nouveau public, nouvelles attentes), le personnel des bibliothèques se doit de trouver de nouveaux moyens de collecter et de transmettre l'information. Pour cela, La British Library a mis en place une stratégie visant à redéfinir la bibliothèque.[1]

C'est dans cette optique que s'inscrit le projet UKWAC (United Kingdom Web Archiving Consortium). Lancé officiellement en 2004, ce projet réunissant six institutions majeures du Royaume-Uni vise à archiver le domaine UK du Web afin de préserver le patrimoine national et de le rendre accessible aux chercheurs.

Les enjeux et défis inhérents à un tel projet sont nombreux, de la définition du document, à la création de métadonnées jusqu'aux techniques de préservation de ces contenus hétérogènes.

Mon stage au sein de la section Web Archiving de la British Library m'a offert l'opportunité d'appréhender ces questions essentielles tout en participant à l'évolution du projet en son entier, aujourd'hui à l'aube d'une phase clé.

La création de collection permet en effet d'aborder les différentes étapes constituant l'archivage d'un site, la préparation, la sélection, la classification, l'autorisation et le processus de capture et d'archivage du site.

1 The British Library's strategy 2005-2008

Partie 1 L'archivage du Web

1 UK Web Archiving Consortium

1.1 Le contexte

L'archivage du Web est un enjeu majeur du 21ème siècle car un nombre grandissant de ressources sont produites et publiées uniquement pour et sur Internet.
La plupart des pays possèdent aujourd'hui un projet d'archivage de ce contenu, que ce soit au niveau national ou au niveau international. Dans la majorité des cas, ce sont naturellement les bibliothèques ou les archives nationales qui sont à l'origine de telles initiatives. Leur rôle étant de récolter, de préserver et de rendre accessible le patrimoine, c'est tout naturellement que ces institutions se sont intéressées aux ressources présentes sur Internet.
Cependant, toutes les institutions ne sont pas égales devant la réalisation d'un tel projet. Hormis les questions techniques et celles liées aux financement, l'une des différences majeures est celle du dépôt légal. Cette question joue un rôle déterminant en ce qui concerne la définition des projets d'archivage européens et mondiaux. Sans dépôt légal, le moissonnage exhaustif du Web est impossible, à moins de passer outre les questions de respect de la propriété intellectuelle. Ainsi, les projets nationaux se limitent en majorité à un domaine et choisissent une approche sélective ou semi sélective d'archivage du Web.

L'un des projet les plus anciens et les plus controversés est celui de l'organisation à but non lucratif américaine, Internet Archive.
Ce projet a été lancé en 1996 et a pour but d'effectuer des captures (ou « snapshots ») du Web mondial à intervalles réguliers, en plus de fournir un accès libre à un certain nombre de ressources numériques. Si Internet Archive est un projet controversé, c'est parce qu'il ne s'embarrasse pas des questions d'ordre juridique. Si l'on ne peut dénigrer l'ampleur et la visée patrimoniale d'un tel projet, on ne peut que constater que cette archive, conséquente en taille, n'offre que peu de possibilités en ce qui concerne l'accès

(recherche par url). L'approche exhaustive ne permet pas de sélectionner en priorité des sites de qualité ou de faciliter leur accès en les indexant de manière appropriée. La masse d'information existe sans être réellement répertoriée. Cependant, Internet Archive participe à de nombreux projets et est un acteur majeur du développement technologique en matière d'archivage du Web.

L'approche sélective, la plus répandue, est utilisée entre autre par MINERVA[2], le projet d'archivage du Web développé en 2002 par la Library of Congress. Ce projet a par exemple permis la collecte thématique de sites lors des évènements du 11 septembre ou relatifs aux élections nationales depuis 2000. Cependant, toutes les collections ne sont pas accessibles depuis le site Internet. La Library of Congress joue cependant un rôle clé dans la conception d'outils normalisés permettant de décrire et de classifier l'Internet. Comme la BNF et la BL, la Library of Congress est un membre du IIPC et a développé de nombreux partenariats afin de permettre au projet MINERVA d'évoluer.

La Bibliothèque Nationale de France (BNF) est un exemple intéressant parce que son approche est mixte mais aussi parce que les captures exhaustives réalisées du Web (domaine national), ne sont pas encore accessibles au public. Depuis 1999, la BNF multiplie les partenariats et les expériences dans le domaine de l'archivage du Web, en attendant une complète régulation des questions de droit d'auteur et de dépôt légal. Par exemple, une capture du domaine *.fr* a été réalisée dès 2002 et des collectes automatisées et thématiques ont été effectuées lors des élections de 2000, 2002 et 2004.

La BNF est également à la tête du Consortium International pour la Préservation d'Internet (IIPC), créé en 2003, dont la British Library est un membre actif. Ce Consortium joue un rôle majeur dans le développement de techniques et d'outils nécessaires à l'archivage du Web. Il réunit les plus grandes institutions mondiales impliquées dans la préservation du patrimoine.

Le IIPC est donc un consortium central dans le développement des projets d'archivage du Web. Ses objectifs sont énumérés ainsi sur le site de la BNF[3]:

2 Mapping the INternet Electronic Resources Virtual Archive
3 Site BNF

- « Approfondir une collaboration technique centrée sur l'identification, l'élaboration et la mise en œuvre d'instruments et de procédés propres à identifier, collecter, préserver et rendre accessibles les contenus de la Toile »
- « Établir progressivement un inventaire des collections des contenus de l'Internet, dans le respect des législations de chaque nation et en fidélité aux politiques de sélection propres à chacune »
- « Plaider partout en faveur d'initiatives et de décisions gouvernementales qui favorisent cette ambition »
- « Apporter un appui aux pays qui souhaiteront s'engager après nous dans cette voie »

La collaboration avec le IIPC est une des collaborations majeures du projet UKWAC et les outils développés pour la capture et la préservation des sites vont permettre au projet d'évoluer à grands pas.

1.2 Historique du projet UKWAC

En octobre 2003, six institutions (The British Library, The Higher Education Funding Council for England, The National Archives, The National Library of Scotland, The National Library of Wales et Wellcome Trust Limited), signaient un accord, le "UK Web Archiving Consortium Agreement". Ce document légal donnait officiellement naissance au "UK Web Archiving Consortium" (UKWAC).
Cet accord désignait alors La British Library comme leader du projet, ce qui ne signifie pas que les décisions ne sont pas prises de concert entre les différentes institutions.

La durée de vie du projet pilote et donc du Consortium, y était alors fixée à trois ans. Il y était également stipulé le rôle et les responsabilités de chacune des institutions le composant.
En effet, outre les questions de budget commun et celles d'ordre technique, chaque institution se devait de prendre en charge une partie du domaine UK à archiver et de faire part de toute modification dans sa politique de sélection aux autres partenaires. L'objectif évident de cette répartition était d'éviter les doublons dans l'archivage mais

aussi de faire en sorte que cette sélection soit de qualité car effectuée par l'institution employant le personnel le plus spécialisé dans tel ou tel domaine.
Par exemple, les bibliothèques nationales d'Écosse et du Pays de Galles ne sélectionnent que les sites reflétant leur histoire et leur culture. La Wellcome Library ne s'occupe elle, que des sites traitant de la santé en général et plus particulièrement, de la médecine.

Ce projet est donc au départ un projet pilote, destiné, durant une phase estimée à trois ans, à partager les coûts ainsi que l'expérience de chacun dans le but d'archiver le domaine UK de manière sélective et ce afin de préserver des ressources de valeur pour les chercheurs ainsi que du point de vue de l'héritage culturel du Royaume-Uni.

En 2003, The Wellcome Limited Trust et JISC[4], commandent deux rapports[5] concernant la situation en matière d'archivage du Web et les questions légales relatives à cet archivage.

Ces deux rapports démontrent l'importance d'un archivage du Web alors qu'un nombre croissant de documents liés aux institutions sont produits sous format numérique et diffusés sur le Web.
Le constat qui y est fait est que la durée de vie moyenne d'un site Internet au Royaume-Uni est équivalente à environ quarante-quatre jours. Or, un nombre croissant de documents sont produits directement au format numérique ou transformés après production initiale au format analogique. Au Royaume-Uni, on considère que la plupart des documents ayant trait à la publication officielle et officieuse du gouvernement se fait directement au format numérique, il en est de même pour une grande part de la littérature grise du type articles et travaux de recherche. Le risque, si les sites Internet ne sont pas rapidement archivés, serait de perdre nombre de ressources précieuses tant pour les chercheurs actuels que pour les générations futures. Ce projet s'inscrit donc dans une visée patrimoniale de préservation puis d'accès du plus grand nombre à notre histoire, à notre culture.

Les rapports font également le point sur les différents projets passés ou en cours en Europe, aux États- Unis, en Australie ainsi qu'au Royaume-Uni, en prenant soin d'examiner pour chacun, les questions d'ordre légal, technique et de bibliothéconomie. La question du copyright y est notamment mise en exergue. Si les difficultés techniques existent, la question de la propriété intellectuelle soulève en effet de nombreux problèmes et influe également sur

4 Joint Information Systems Committee
5 Études de faisabilité

le choix de la méthode d'archivage. Le respect de la propriété intellectuelle et la protection des données personnelles sont notés comme deux aspects essentiels de ce projet.

Suite à ces rapports, il est décidé de former un consortium réunissant plusieurs institutions impliquées dans la préservation des ressources nationales ainsi que dans leur diffusion, UKWAC. Suite aux risques exposés dans le rapport concernant les questions légales, le Consortium choisit d'adopter une approche sélective du Web et de mettre en place un processus d'acquisition d'une licence permettant de capturer, de préserver et de rendre accessibles les sites retenus, en attendant une évolution souhaitée de la loi relative au dépôt légal.

1.3 Objectifs et évolution

L'objectif du projet est donc de construire l'archive Internet du domaine national afin de le préserver et de le rendre accessible aux chercheurs ainsi qu'aux générations futures.

Ce projet à forte valeur patrimoniale comporte, nous l'avons dit, de nombreuses difficultés d'ordre à la fois technique et intellectuel ainsi que de nombreux défis à relever. Pour ces différentes raisons et parce qu'il s'agissait au départ d'un projet pilote, des champs d'action restreins ont été définis.

Les partenaires du consortium faisant tous partie du Royaume-Uni, il a tout d'abord été décidé que l'archive ciblerait les sites du domaine UK[6], sauf exception, par exemple ceux liés à des évènements signifiants pour la nation, comme le tsunami. La mission première était en effet de construire d'ici à 2006 une archive qui soit représentative de la culture britannique dans sa diversité régionale, culturelle, ethnique, etc.

Cette archive devait dans un deuxième temps être accessible gratuitement par le biais d'un site internet. Les membres du Consortium ont donc créés un site Web commun donnant accès aux sites archivés par leur soin.[7]

Étant donné le nombre de partenaires impliqués dans le projet, il a également été nécessaire de créer un certain nombre de documents détaillant la politique à suivre en matière de sélection des sites Web ou bien encore en matière de choix des outils tels que le logiciel destiné à aspirer les sites. La création de ces documents est spécifiée dans

6 *.uk* ou *.co.uk* par exemple, cependant la distinction n'est pas toujours aussi évidente.
7 http://www.webarchive.org.uk/

l'accord initial et vise à uniformiser le projet. Les logiciels, la classification et les documents légaux permettant de traiter les sites doivent être les mêmes pour chacun des membres du Consortium.

En 2003, il existait peu de projets similaires à ce que les membres d'UKWAC voulaient entreprendre. Le Consortium se tourne donc vers la National Library of Australia qui a débuté un projet d'archivage du Web sélectif en sa qualité de Bibliothèque Nationale dès 1996, PANDORA.[8] Ce projet a nécessité le développement d'une politique de sélection des sites, puisque l'approche choisie est sélective, le développement d'un outil technologique permettant de gérer, d'archiver et de diffuser les sites sélectionnés, ainsi que la création d'un site Web.

C'est ainsi qu'est développé PANDAS (PANDORA Digital Archiving System)[9], un logiciel permettant en autre d'aspirer les sites Web (en utilisant le logiciel libre HTTrack), d'archiver et de donner accès aux sites sélectionnés.

Ses fonctions sont alors définies comme tel:

- Gérer les données et métadonnées des sites entrés dans la base.
- Commander et planifier la capture des sites sélectionnés pour être archivés.
- Gérer le contrôle qualité des sites et résoudre les éventuels problèmes techniques.
- Préparer l'accès des fichiers à la diffusion au public et générer un titre pour la page d'entrée ainsi qu'un identifiant pour chaque titre.
- Gérer les profils (niveaux d'accès autorisés)
- Produire des rapports statistiques

UKWAC choisit donc d'acquérir une licence d'utilisation de cet outil, qui devient commun à tous les membres du Consortium, permettant ainsi une unicité au niveau de l'archivage et de la diffusion, un contrat est également signé avec l'entreprise MAGUS afin de garantir le support technique.

8 http://pandora.nla.gov.au/
9 Rapports internes

PANDAS permet en effet de donner un accès commun à des sélections effectuées par les différents membres opérant depuis différentes locations géographiques et possédant chacun une base de données propre.
Il permet également d'assurer à ses membres le contrôle des titres rendus publics car les niveaux d'accès sont limités selon qu'un titre appartient à telle ou telle institution. Les informations concernant un titre sont donc à la fois descriptives (url, sujet, etc.), administratives (institution propriétaire, Web archiviste en charge, etc.) et techniques (fréquences d'archivage, dates, etc.).
Ce système permet, en plus de l'aspect « sécurité » lié au travail collaboratif, d'assurer un suivi du projet. Cet aspect est en effet majeur lors de l'élaboration d'un projet pilote puisqu'il permet de prendre en compte les difficultés rencontrées et ainsi, d'améliorer le projet sur le long terme, grâce à des techniques et outils de gestion des connaissances.
En 2005, l'archive est accessible au public par le biais du site Internet nouvellement créé.

Un rapport interne de 2006[10] sur l'évolution du projet montre qu'en septembre 2006, l'archive accessible au public contenait 1603 titres et 5344 captures de sites Internet. A la fin de la période de contrat les liant, les membres du Consortium décident de poursuivre le projet UKWAC.

Les défis évidents à relever dans la réalisation d'un tel projet sont la participation active des six institutions à ce projet ainsi que la collaboration au niveau décisionnel. La répartition des différentes équipes du point de vue géographique de même que la différence éventuelle d'implication (moyens humains et financiers). Les divergences de point de vue relatives aux objectifs et priorités du projet selon les intérêts des différentes institutions.

Il est également techniquement parfois difficile de s'assurer de l'appartenance d'un site au domaine *.uk*, comme de s'assurer de l'uniformité des données même si les institutions utilisent le même outil de classification et de capture des sites. En effet, les bases de données sont différentes selon l'institution, de même que la méthode de classification (plus ou moins subjective), etc.

Enfin les questions hautement techniques telles que le moissonnage des sites, leur stockage afin d'être préservés ainsi que leur diffusion au public sont extrêmement

10 TUCK, John, *Evaluation report*, UKWAC, April 2006

compliquées à résoudre étant donné la rapidité d'évolution des nouvelles technologies (par exemple, PANDAS ne peut pas copier les sites utilisant un contenu dynamique). Les choix effectués sont donc lourds de conséquences et parfois rendus difficiles par le nombre de partenaires participant au projet.

Mais le défi majeur ou frein à ce projet reste la question du copyright ainsi que celle du dépôt légal, pour les institutions habilitées à en bénéficier comme la British Library.

Il faut rappeler ici qu'il est à ce jour impossible pour UKWAC d'effectuer une capture du domaine national en son entier car il est nécessaire de posséder un document écrit autorisant la capture, l'archivage ainsi que la diffusion de chaque site Web par son propriétaire. Les seuls membres n'ayant pas à se soucier du copyright étant The National Archives qui collecte les sites gouvernementaux et JISC, puisque cette institution finance la plupart des projets dont elle archive les sites. La première a d'ailleurs créé des partenariats avec d'autres projets d'archivage du Web, comme par exemple Internet Archive[11].

Le processus est donc extrêmement lent et l'approche nécessairement sélective. A titre d'exemple, le taux de réponses positives actuellement enregistré par la BL est d'environ un tiers, parfois moins.

En 2007, le projet UKWAC entre dans une phase de transition importante à plusieurs niveaux.

PANDAS, l'outil développé par la National Library of Australia est en passe d'être remplacé par le Web Curator Tool (WCT)[12]. qui fonctionne avec le logiciel open source d'Internet Archive, Heritrix. Ces outils développés en collaboration avec le IIPC sont d'ores et déjà utilisés par d'autres projets d'archivage du Web. Le WCT est en phase de test à la BL depuis plusieurs mois et l'équipe technique, conseillée par les archivistes, travaille à son adaptation aux besoins et évolutions du projet. Si PANDAS est limité techniquement, l'expérience tirée de son utilisation durant les premières phases du projet est donc aujourd'hui utile.

Le passage au WCT devrait apporter une gestion plus globale et automatisée des informations ainsi que des étapes de l'archivage, telles que les demandes d'autorisation par email.

11 Collections IA
12 http://webcurator.sourceforge.net/

Le robot, Heritrix, est celui utilisé par des projets pratiquant une approche exhaustive du Web, parfois à très grande échelle. Son utilisation permettra de multiplier la vitesse d'archivage des sites. Les « bugs » sont également moindres qu'avec PANDAS ce qui devrait permettre d'archiver plus vite et mieux en réduisant les problèmes liés à la phase de capture.

UKWAC se préparant au passage d'une loi permettant le dépôt légal des sites Internet en 2008, cette migration vers un nouveau système est absolument nécessaire dans le cadre d'une évolution du projet vers un archivage mixte du Web.

2 L'archivage du Web à la British Library

La British Library est donc la partie leader dans le projet d'archivage du Web et joue un rôle majeur en ce qui concerne les questions d'évolution du projet tant du point de vue technologique que de celui des politiques à suivre.
Il faut rappeler que le Consortium tente cependant de répondre aux demandes de chacune des institutions.

2.1 Web Archiving Programme (WAP)

La British Library (BL) a mis en place un programme s'inscrivant dans celui plus vaste d'UKWAC afin de mener à la réalisation des objectifs fixés.

2.1.1 Les ressources humaines

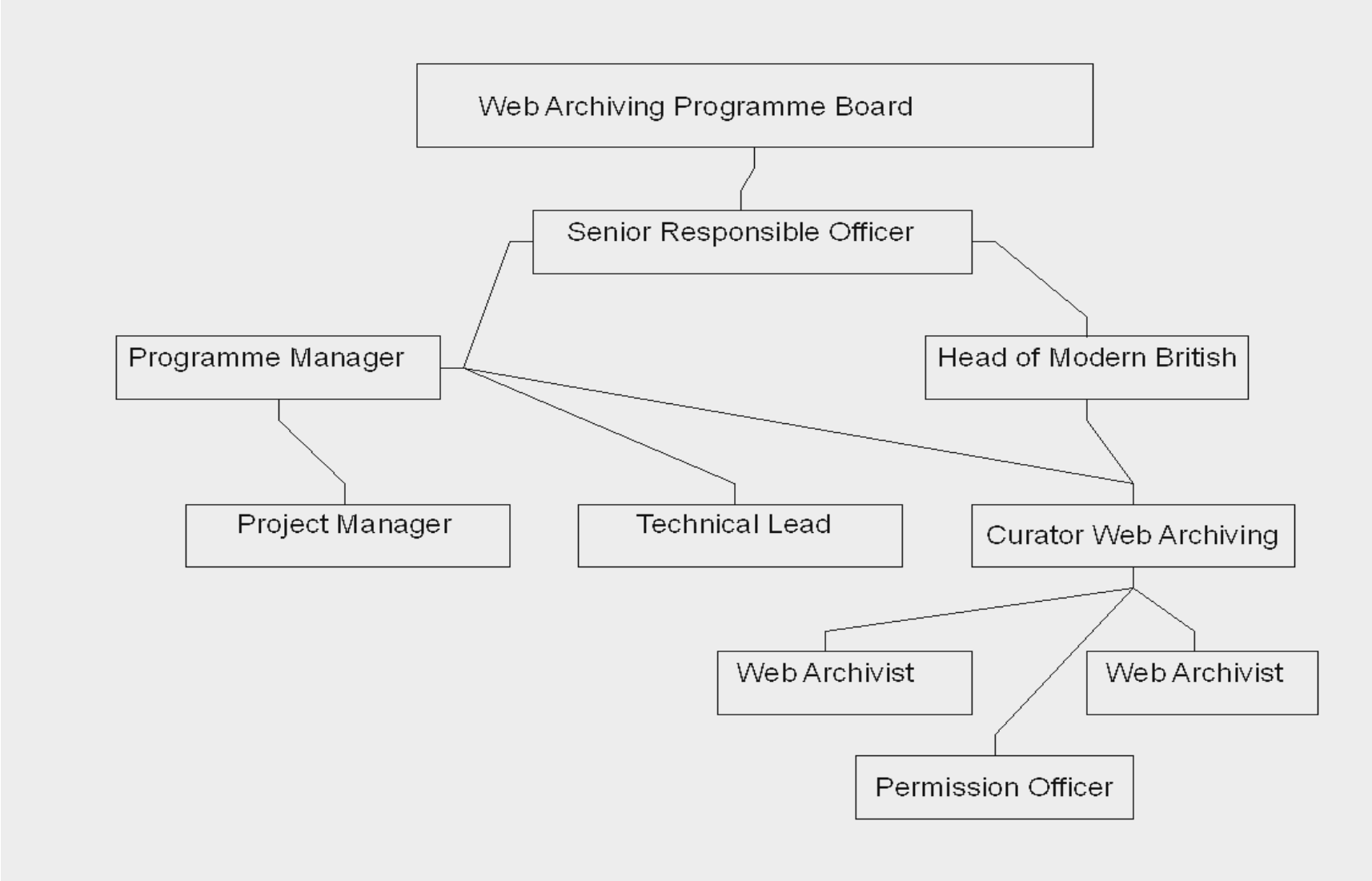
Le programme WAP de la BL est dirigé par un certain nombre de directeurs de Département. Les décisions prises par le manager de l'équipe doivent donc être validées par ce comité décisionnel. Il est composé aujourd'hui de huit personnes, comme indiqué ci-dessous:

Web Archiving Programme Board:

- ET Champion for the Programme
- Head, Collection Acquisitions & Description
- Head, European & American Collections
- Head, British and Early Printed Collections
- Digital Preservation Manager
- Head, Architecture and Development
- Head, Modern British Collections
- Head of British Collections - Senior Responsible Officer (SRO) for the Programme

Le schéma suivant montre la hiérarchie de ce projet et le fonctionnement de l'équipe WAP. On peut constater que cette équipe dépend à la fois du Consortium et de la British Library.

Le service dans lequel j'ai travaillé durant ces quatre mois est le Département Modern British Collections de la bibliothèque.



L'équipe WAP fonctionne sous la responsabilité du comité décisionnel décrit ci-dessus. Elle se compose de la manière suivante:

- **Senior Responsible Officer**

Il siège au comité décisionnel et dirige les réunions du consortium UKWAC. Il est chargé de valider toute décision prise pour le projet de la BL.

- **Programme manager**

Il est chargé de gérer et d'organiser le projet à tous les points de vue (techniques, budgétaires, etc.). Il est présent aux réunions du Consortium et propose une vision et une évolution du projet. Il dirige également les réunions mensuelles de l'équipe de la BL.

- **Project manager**

Son rôle est d'assister le chef de programme et de gérer le projet en suivant les recommandations du comité. Il participe également aux réunions du Consortium.

- **Technical Lead**

Les questions techniques sont prises en charge par le directeur technique. Il gère par exemple les accès aux outils de communication utilisés par l'équipe ou les évolutions technologiques du projet en collaboration avec le programme manager.

- **Curator Web Archiving**

Les questions relevant de la collection sont gérées par la conservatrice en chef de ce projet. Elle décide des questions relatives à l'évolution de la collection et dirige les réunions de sélecteurs. Elle gère enfin l'équipe composée des archivistes et du permission officer.

- **Web archivists**

La conservatrice est assistée par deux Web archivistes, l'un possède un profil scientifique et l'autre un profil d'historien. Ils s'occupent de répondre aux questions des propriétaires de sites sélectionnés, sélectionnent certains sites, gèrent l'aspect collection du site Internet et archivent les sites (technique). Ce sont également eux qui s'occupent de collecter les sites lors la création de collections thématiques liées à un événement.

- **Permission officier**

Il est chargé de gérer la correspondance, la BDD ainsi que les statistiques et l'administratif en général. Toutefois, durant ma période de stage, cette personne était en cours de recrutement.

Les décisions concernant la politique de développement des collections est sous la responsabilité du Head of Modern British Collections, le Dr. Richard Price et d'Alison Hill, la Conservatrice en chef. Il est à noter que le Dr. Richard Price siège également dans le comité de direction du programme, tout comme Alison Hill.

De plus, il est à signaler que l'équipe est scindée en deux, à savoir, quatre personnes dans le Yorkshire (Programme et Project Manager, Technical Lead et l'un des archivistes), les trois autres membres de l'équipe se trouvant à Londres.

Le travail de sélection est effectué par les Web archivistes de même que par un groupe de spécialistes, désignés par la conservatrice. Cette équipe de spécialistes a été créée de manière à couvrir tous les sujets représentés au sein de la BL de manière professionnelle. Le fait que la sélection soit effectuée par des spécialistes garantit la qualité et le haut niveau de sélection nécessaires à la crédibilité et donc à la réussite du projet. En effet, l'argument de cette politique est de mettre en avant la qualité de l'archive ainsi obtenue et sa valeur pour les futurs utilisateurs.

Cette archive étant constituée par la BL, le public visé est en priorité celui des chercheurs et universitaires. Il me semble que c'est également un moyen de faire évoluer le rôle du conservateur et du bibliothécaire de l'intérieur. La section Web Archiving est en effet intégrée à la politique de développement de la BL. Les conférences et présentations du projet UKWAC en interne sont une excellente opportunité de faire évoluer les consciences.

Le rôle des Web archivistes n'est donc pas principalement la sélection. L'archivage d'un site Internet est aujourd'hui encore très « artisanal », c'est pourquoi la présence des archivistes est essentielle à ce processus.

Il leur faut donc entrer les données dans PANDAS, planifier la capture du site, vérifier la copie obtenue et enfin archiver le titre sur le site Web d'UKWAC. Cependant, PANDAS, le logiciel utilisé, comporte de nombreux « bugs », notamment dans la capture des CSS[13] et bien souvent, il est nécessaire de récupérer certains fichiers manuellement, à l'aide d'HTTrack et d'un programme Perl créé en interne, qui permet de retrouver les erreurs de connexion lors de l'aspiration du site et de les corriger une à une. Ce processus d'archivage sera développé dans la deuxième partie de ce document.

13 Cascading Style Sheet

2.1.2 Ressources matérielles

Étant donnée la séparation géographique, les membres de l'équipe ont mis en place un Wiki ainsi qu'un espace de travail virtuel spécifique au projet en développant le logiciel libre Track[14] selon leurs besoins.

Ces outils facilitent d'une part la communication et permettent d'autre part de conserver une archive précise et détaillée du projet lui-même. L'équipe conserve ainsi les documents produits dans leurs différentes versions, les problèmes rencontrés par les uns ou les autres et les solutions adoptées, les comptes-rendus des réunions, etc.

Il s'agit de conserver une bonne cohésion tout en facilitant le partage des connaissances. De plus, cela permet de pallier aux éventuels changements dans la structure de l'équipe. Le projet étant un projet à long terme et l'équipe ayant déjà eu à souffrir du départ de certains de ses membres, il est important de garder une trace de ce qui les constitue.

Cet espace de travail m'a été très utile puisqu'il m'a permis d'accéder à l'historique du projet de l'intérieur ainsi que de répondre à de nombreuses questions, notamment d'ordre technique ou lexical. L'équipe étant réduite et éclatée du point de vue géographique, la gestion des connaissances est indispensable à son bon fonctionnement comme à l'intégration rapide et efficace de nouveaux membres.

Le logiciel Track permet quant à lui de gérer et de planifier le travail de chacun par le biais d'une interface commune.

Chaque membre de l'équipe possède des « tickets » c'est-à-dire des actions à accomplir et chacun d'entre eux peut assigner une action au membre de son choix. Cela est extrêmement utile dans le suivi des correspondances par exemple. Avant la mise en place de ce système, les courriers, relatifs bien souvent aux questions de propriété intellectuelle, étaient conservés au format papier. L'équipe étant réduite et éclatée au niveau géographique, cela pouvait être source de retard dans le traitement de ces questions. Grâce à ce système, une limite de temps est assignée à chaque action, ce qui permet non seulement une plus grande efficacité mais aussi la création automatique d'un historique concernant chaque dossier.

Cependant, la majorité des informations concernant les sites sélectionnés sont conservées dans la base de données Access ainsi que dans PANDAS, le logiciel commun à toutes les institutions du projet. La base de données est classique et comprend les champs suivants:

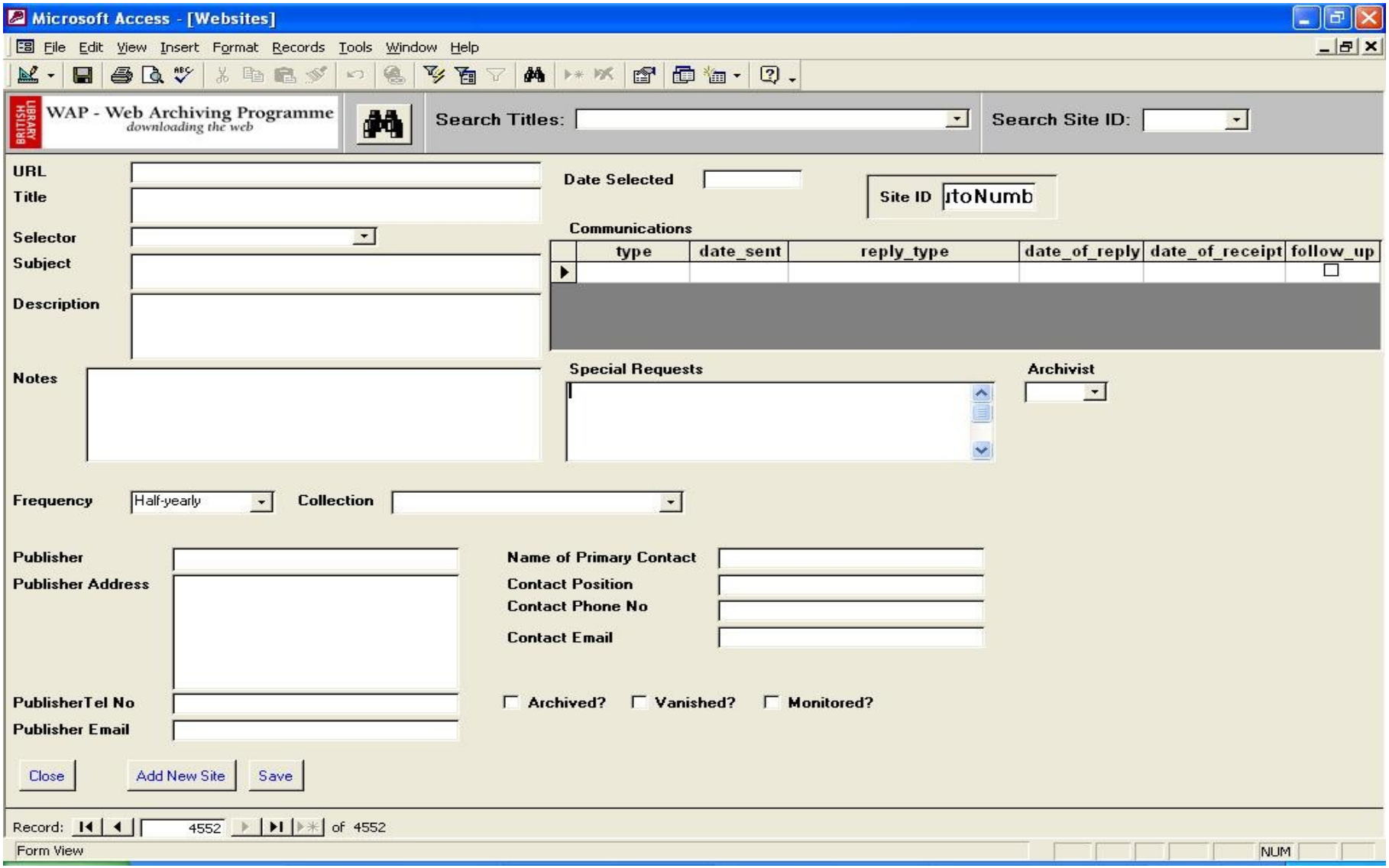
14 Track

- Url
- Titre
- Sélecteur
- Sujet (classification normalisée)
- Description
- Note (problèmes techniques éventuels, licence Creative Comons, etc.)
- Fréquence (de capture du site)
- Collection (le cas échéant)
- Éditeur (Nom, adresse, contact)
- Contact (Nom, position, adresse, etc.)
- Date de sélection / relance / réponse
- ID unique
- Communication (email, courriers, dates, réponses, etc.)
- Demandes particulières (du propriétaire du site, par exemple « ne pas capturer les photos », etc.)
- Nom de l'archiviste
- État du site (archivé, disparu, prioritaire)

Les champs essentiels sont le « titre », « l'url », les descripteurs, les informations concernant le propriétaire, les dates de correspondance et le résultat obtenu. Ces champs sont en effet ceux utilisés pour la création de la majorité des statistiques. Le champ « follow-up » permet par exemple de créer des statistiques concernant l'impact des relances lors des demandes d'autorisation. Le rôle du « Permission Officer », qui gère ces champs, est donc essentiel pour le suivi du projet.
Chacune de ces statistiques mensuelles et annuelles permet de contrôler l'évolution du projet et de valider ou non certaines stratégies d'approche des propriétaires de sites Internet ou encore, des méthodes de sélection.

Certains champs comme le champ « Note » ou « Description » sont optionnels.
Il est à noter qu'il n'y a pas de champ décrivant le format du document, cette information est donc insérée soit dans la description, soit dans les notes.

Il est possible de faire une recherche simple dans n'importe quel champ ou de créer des requêtes pour les recherches plus complexes (concernant plusieurs champs) ainsi que pour celles concernant de nombreux sites. C'est ainsi que sont créées des requêtes pour suivre l'état d'une collection par exemple ou pour récupérer des statistiques mensuelles.
Le système de classification utilisé pour indexer les données sera développé dans la deuxième partie de ce document.



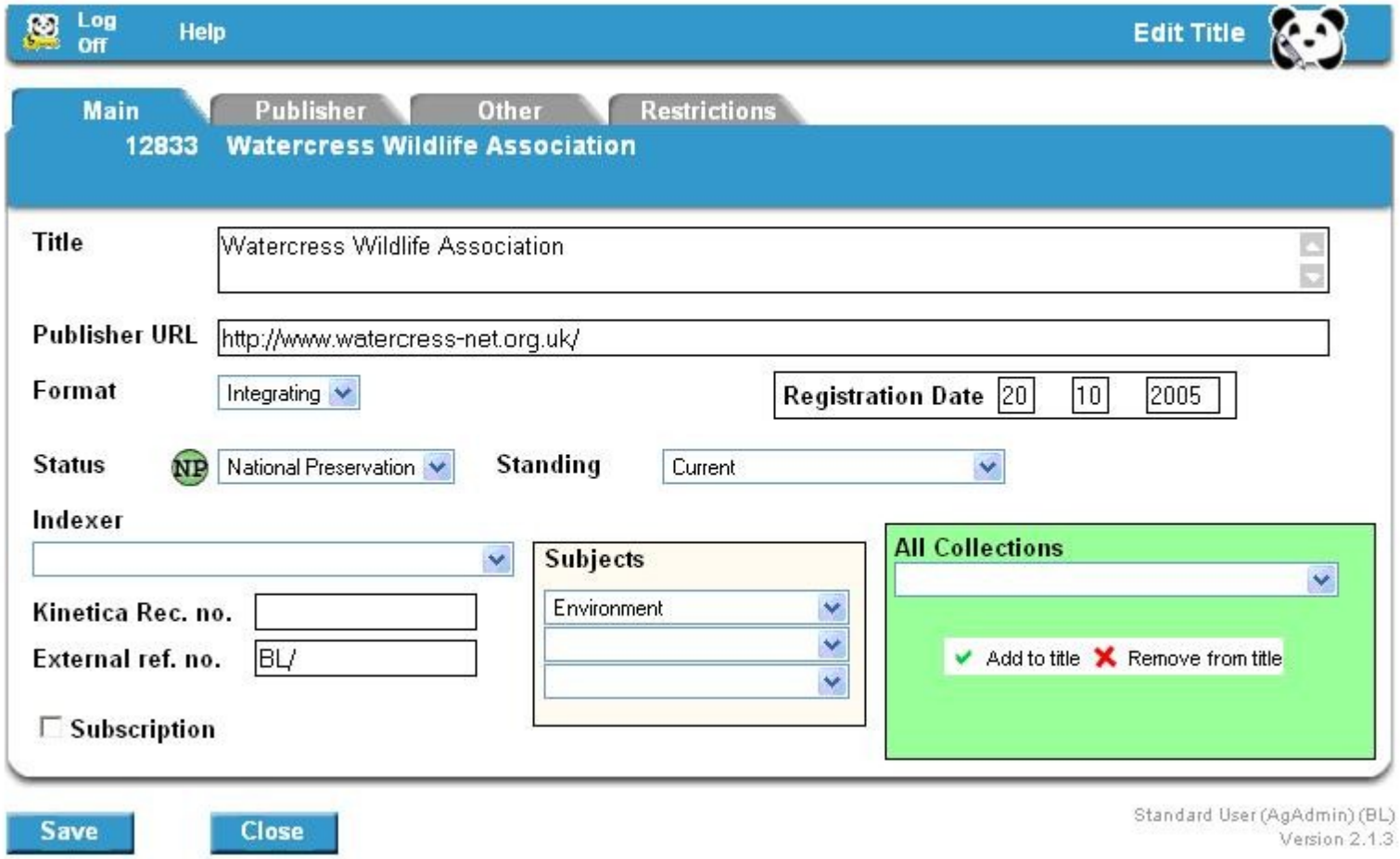
La capture des sites ainsi que l'archivage sont gérés, comme expliqué plus haut, par PANDAS depuis le début du projet, qui est, comme expliqué précédemment, en passe d'être remplacé par le Web Curator Tool, utilisant Heritix.

PANDAS est un outil développé par des bibliothécaires aux débuts de l'archivage du Web. Les caractéristiques techniques ne sont donc aujourd'hui plus adaptées aux besoins des projets d'archivage du Web. Il provoque de nombreux « bugs » lors de la capture des sites et est assez vite dépassé. En effet, il n'est possible de moissonner que quelque sites à la fois (environ cinq), au-delà, le logiciel est considérablement ralenti, ce qui paraît peu propice à l'évolution souhaitée du projet vers un archivage exhaustif ponctuel du domaine national.

Pour palier aux problèmes de capture rencontrés lors de l'utilisation de PANDAS, l'équipe de la British Library utilise en parallèle le logiciel open source HTTrack (en externe à PANDAS) ainsi qu'un programme développé en Perl qui permet de repérer les erreurs de capture des sites et de les corriger manuellement (CSS non localisées ou localisées au mauvais emplacement le plus souvent, etc.) sans avoir à relancer toute la procédure sous Pandas. Ce dernier étant incapable de gérer un grand nombre de sites à la fois, cela permet de ne pas le freiner davantage tout en évitant les « bugs ».

Cependant, cela constitue un temps de travail supplémentaire pour les Web archivistes qui, après avoir effectué le contrôle qualité du site, doivent examiner les erreurs et les traiter une à une. Le contrôle qualité est en effet essentiel au projet UKWAC et représente une forte valeur ajoutée pour les partenaires comme pour les propriétaires de sites. La qualité de l'archive construite doit être à la fois visible dans la sélection et dans les copies disponibles à la consultation.

Les étapes techniques de l'archivage d'un site ainsi que les difficultés liées à l'utilisation de PANDAS seront développées dans la deuxième partie de ce document.



Le fonds conservé dans la base de données et dans PANDAS comprend des sites sélectionnés du point de vue de la recherche, pour leur intérêt culturel, pour leur caractère innovant ou parce qu'ils sont représentatifs du Royaume-Uni. Ces sites sont sélectionnés individuellement ou dans le cadre de la création d'une collection.
Il existe des collections thématiques:

- liées à un événement de l'actualité ayant une répercussion sur la vie des britanniques:
  - Avian & Pandemic Influenza
  - UK General Election 2005
  - London Terrorist Attack 7th July 2005
  - Indian Ocean Tsunami December 2004

- plus générales, initiées par l'un des membres du Consortium:
  - Women's Issues
  - Latin America UK
  - British Countryside
  - Digital Lives
  - Travel
  - Blogs
  - etc.

PANDAS permet à chaque membre de décider si les collections qu'il a créées seront publiées sur le site Internet ou non. L'accès à ces collections peut être effectué à partir de la page d'accueil du projet ou à l'intérieur de chaque catégorie, le cas échéant.
Toutes les collections ne possèdent pas une entrée par l'onglet de la page d'accueil mais celles qui y sont répertoriées possèdent également un descriptif.
En plus de cet accès par collection, le site Internet créé pour le projet UKWAC permet une recherche par ordre alphabétique, par catégorie et par sujet. Une recherche simple en texte intégral est également disponible sur la page d'accueil, cependant le moteur de recherche ne donne pas toujours de résultats satisfaisants. Le site d'UKWAC est d'ailleurs en passe d'être modifié après la migration du projet vers WCT et Heritrix.

Au 31 juillet 2007, le nombre de sites sélectionnés par la BL était de 4503 dont 3244 copies accessibles par le biais du site public UKWAC. La taille de l'archive UKWAC était de 1671 GB au 31 juillet 2007, dont 769 GB occupés par la BL.

3 Projet de stage à la section Web Archiving

Mon projet de stage à la BL a consisté en la création d'une collection de blogs, en la sélection de sites de religion Non Conformiste, c'est-à-dire pour simplifier, non conforme à la religion anglicane et catholique, ainsi qu'en un travail visant à contourner les problèmes de droit dans le cadre de l'archivage des sites d'art.
Ces derniers posant le plus de problèmes en ce qui concerne l'obtention des droits de conservation et de diffusion des sites, cette partie du projet était au départ conçue à la manière d'un projet pilote. Nous avons donc mis en place et testé différentes stratégies d'approche, notamment des grandes institutions.
Le travail concernant les sites d'art a malheureusement été ralenti puis suspendu au cours de mon stage, notamment en raison des problèmes de droit.
Il est également à noter que ma période de stage a coïncidé avec la période de négociation entre les membres du Consortium en vue de son évolution prochaine. Le contrat liant les membres du Consortium arrivait en effet à son terme en septembre 2007 (date repoussée lors du dernier rapport d'évaluation de 2006).

3.1 La demande

Les collections créées suite à l'approche sélective d'UKWAC constituent la valeur ajoutée de l'archive. La sélection de la majorité des sites est donc effectuée par des spécialistes, tous conservateurs à la British Library.
Cependant, la taille réduite de l'équipe ne permet pas de développer ces collections de manière satisfaisante. C'est pourquoi l'équipe a choisi de créer des partenariats avec des spécialistes internes ou externes à la BL. Cela a été le cas par exemple pour la collection concernant les femmes, réalisée en collaboration avec The Women's Library.[15]

15 http://www.londonmet.ac.uk/thewomenslibrary/

Mon rôle durant ce stage a consisté à créer une collection destinée à la diffusion, celle des blogs, ainsi qu'une collection répondant à la demande du Dr. Clive D. Field, les sites de religion Non Conformiste.
J'ai également été chargée de faire l'analyse de l'existant en matière de sites d'art et, le cas échéant, de contacter les institutions propriétaires afin d'obtenir l'autorisation de les copier, de les archiver et de les diffuser.
Il est à noter que la demande de départ était large, elle consistait en la création d'une collection de blogs, en la sélection de sites d'art ainsi qu'en un travail sur leur classification et en un certain nombre de tâches destinées à palier les temps d'attente. Cependant, il a été nécessaire de revoir la demande initiale pour plusieurs raisons.
Tout d'abord, une question de temps, les quatre mois de stage nous ont finalement paru insuffisant pour réaliser chacun des objectifs de manière satisfaisante. Le fait que la personne chargée d'envoyer les demandes d'autorisation d'archivage ait été en cours de recrutement a modifié mon planning puisqu'il a fallu compter un temps assez long de travail administratif.
Ensuite, la demande de sélectionner quelques sites de religion Non Conformiste est devenue une priorité du point de vue de la hiérarchie, ce qui a transformé un travail de sélection en la création d'une seconde collection.
Enfin, le travail de classification est passé au second plan parce que le Consortium migre d'un système à un autre. Ce nouveau système utilisera un logiciel permettant une recherche par url (ou au mieux en texte intégral) et qui plus est, une étude a été commandée afin d'analyser les besoins d'un nouveau site Web à venir. Mon travail a donc consisté pour cette part en l'uniformisation des mots clés choisis pour la base de données interne à la BL.
Pour ce qui est de la collection de sites d'art, elle a été reléguée en troisième projet, consistant surtout à essayer de récréer ce qui sera plus tard une collection en retrouvant les sites d'art dans la bases de données. La deuxième partie de mon travail a été de reprendre contact avec tous ces sites afin d'obtenir l'autorisation de les archiver et ce en collaboration avec Chris Michaelides, le conservateur chargé de créer la collection Art. Ce dernier n'a d'ailleurs pas pu être aussi présent qu'il le souhaitait car mon stage a coïncidé avec la préparation d'une exposition concernant l'Avant-garde, ce qui a monopolisé toutes les personnes intéressées et susceptibles de m'appuyer, à commencer par lui.

A cela se sont ajoutées des tâches quotidiennes, comme la vérification de la base de données et des sites archivés pour réparer les liens morts ou la participation active aux réunions de l'équipe.
J'ai également participé à un certain nombre de projets ponctuels, tels que le choix d'un site Web pour une opération publicitaire concernant le projet, la refonte du site Internet, des présentations lors de réunions, etc. J'ai aussi eu l'opportunité d'assister à de nombreuses conférences.
Tout cela n'était évidemment pas planifié au départ et a beaucoup influé sur la nécessaire redéfinition de ma planification des différents projets.

3.2 Enjeux

Après la période de familiarisation avec le lieu, l'équipe et le service, j'ai tenté de définir une méthode de sélection de sites Web dans le cadre d'une création de collection.

Mon premier réflexe a été de lire les documents concernant les collections existantes, ainsi que les collections elles-même. Mon premier constat a été l'accent mis sur la diversité de forme et de contenu des sites. J'ai très vite réalisé en discutant avec la conservatrice, que son souci majeur était de respecter cette représentativité qui est liée autant à la qualité de la collection pour les futurs chercheurs qu'à la politique de neutralité de la BL qui se doit d'être représentative du Royaume-Uni. Il m'a semblé que cette question était d'autant plus importante que mes collections allaient être très sensibles puisque liée à la religion pour l'une et représentative de la population et en pleine évolution (format, technologie, groupe,...) pour l'autre.

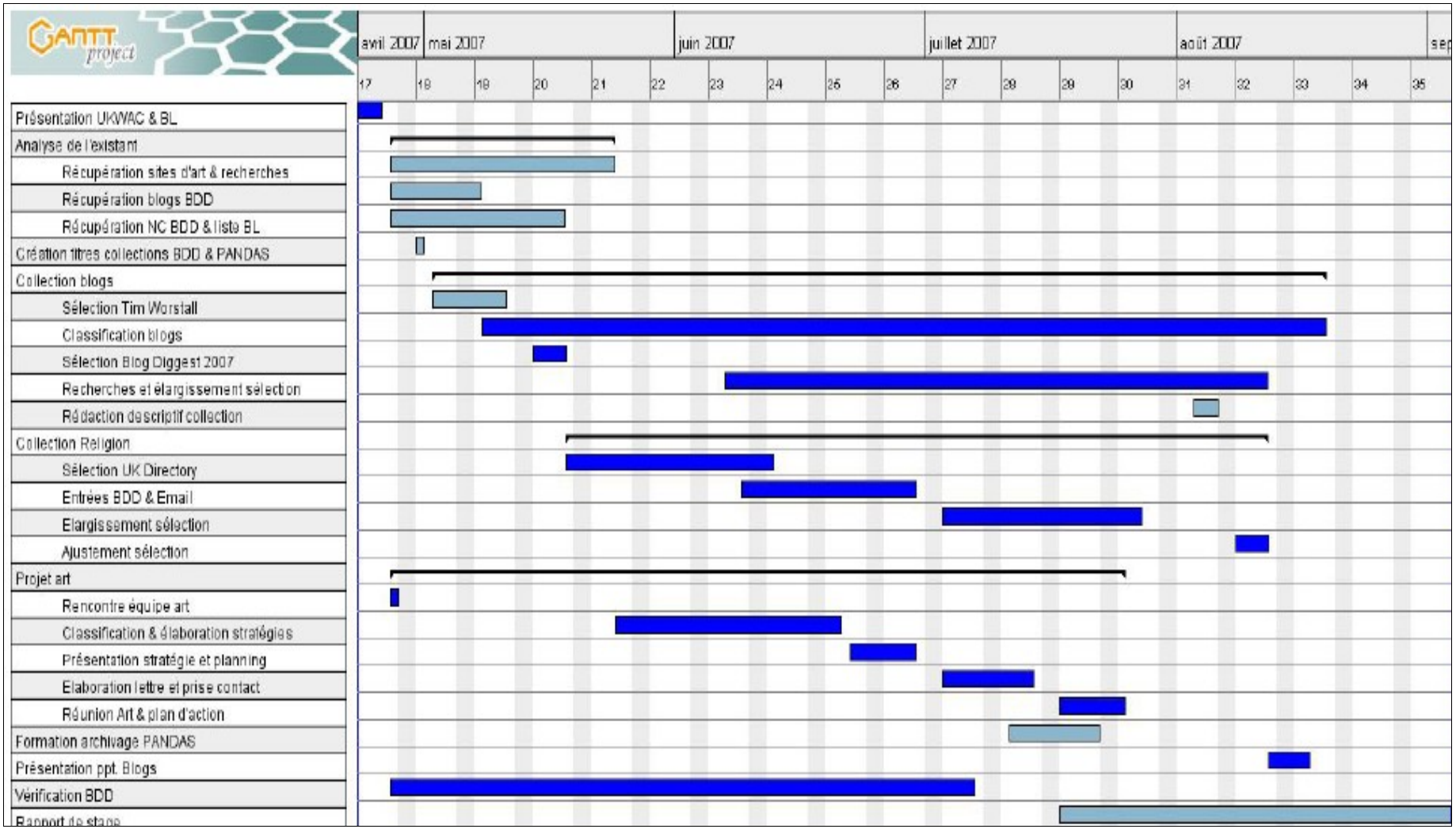
Le deuxième enjeu est lié à l'absence de dépôt légal concernant les sites Internet. Cet aspect m'a conduit à me poser énormément de questions concernant la sélection de sites dans le cadre de la création de ces deux collections car je devais en quelque sorte anticiper quels propriétaires donneraient leur accord.

Enfin la question de l'indexation du site Web en tant que document m'a énormément intrigué et je me suis intéressée à la classification utilisée par le projet.

3.3 Planification du projet

La planification des projets a constitué l'étape sans doute la plus délicate. Le projet étant « pilote », les données étaient approximatives en la matière et ce pour plusieurs raisons. D'une part, les collections sont d'habitude créées par plusieurs spécialistes et rarement par une seule personne. D'autre part, les spécialistes travaillant à l'élaboration de collections occupent en général un poste à temps plein. La BL a aujourd'hui décidé de palier à ce problème en allouant un temps de sélection de sites à chacun des spécialistes de la BL. Cependant les chiffres relatifs aux collections créées dans le passé ont été peu significatifs car mon cas était différent des précédents.
Le temps moyen de création d'une collection est de six mois. Il a donc été décidé que la collection de blogs était une priorité. Le nombre de blogs à livrer a été fixé à 150 (avec licence).
En ce qui concerne les sites religieux, le nombre demandé était de 150 sites sélectionnés (avec ou sans autorisation).
Pour les sites d'art, le résultat n'a pas été chiffré au départ car l'approche était complètement nouvelle. Le planning du projet a d'ailleurs été modifié à plusieurs reprises parce que le personnel compétent était peu disponible à cette période, jusqu'à l'annulation du projet pour des raisons de politique de développement de collection à revoir.

Le diagramme de Gantt présenté ci-dessous montre l'évolution de mon travail dans le temps.



Partie 2 Archivage du Web

1 Création d'une collection

La création d'une collection destinée à une archive telle que celle de ce projet amène de nombreuses interrogations, à commencer par la méthode à adopter en matière de sélection. Il m'a semblé que détailler mes projets un à un serait peu représentatif de mon travail durant ces quatre mois. Je suis en effet passée d'un projet à un autre de manière permanente et il me semble que l'expérience acquise pour les uns m'a été très utile pour les autres. Finalement, j'ai considéré ce projet dans sa globalité en essayant de comprendre chaque défi à relever lors d'une entreprise aussi massive que celle de l'archivage du domaine national.

1.1 Analyse de l'existant

Je me suis tout d'abord intéressée aux autres projets d'archivage sélectif ou semi automatisé du Web afin de savoir s'ils possédaient des collections.
PANDORA, le projet d'archivage du Web de la Bibliothèque Nationale d'Australie, possède des collections, cependant celles-ci sont le plus souvent événementielles (tsunami, élections, etc.) et ne sont pas mises en avant par un accès direct sur la page d'accueil.
Pour les atteindre et découvrir leur existence, il faut entrer dans l'une des catégories qui contient des sites appartenant également à une collection. L'onglet collection apparaît alors et il est possible d'accéder à la totalité de la collection. Dans cette archive, les blogs forment une sous catégorie et non pas une collection.
Je me suis également intéressée au projet MINERVA de la Library of Congress. Ce site m'a été utile parce qu'il propose un certain nombre de documents tels que la politique de

développement des collections ou encore les métadonnées utilisées. Cependant, les collections sont similaires à celles de PANDORA.
Il m'a semblé que les autres initiatives étaient davantage axées sur l'archivage automatisé du Web, la question du classement par catégories ou de regroupement par collections, venant en second.
Internet Archive est par exemple axé sur l'archivage du Web mondial. L'interface utilise la « Wayback machine » comme accès à la recherche. Si les ressources sont accessibles par format et si quelques collections existent[16], la recherche se fait avant tout par url. Les collections sont soit réalisées de manière automatisée, en programmant Heritrix, par exemple pour les évènements liés au tsunami,. soit réalisées par d'autres organismes, comme par exemple The National Archives. Les objectifs du projet sont donc différents de l'approche sélective de la BL.

L'enjeu était donc de différencier le travail de la BL par rapport aux autres projets. Comme cela est précisé plus haut, la BL entend développer les collections dans ce projet d'archivage car celui-ci est destiné avant tout aux chercheurs et universitaires. Les collections doivent en ce sens être composées de manière réfléchie et doivent couvrir au mieux la politique de développement spécialement conçue pour UKWAC[17].
UKWAC fonctionne de telle manière que l'archive doit être commune à toutes les institutions et toute information, concernant par exemple la création d'une collection, doit être partagée avec les partenaires. Dans les faits, il est impossible d'imposer à tous les membres de trouver de l'intérêt et le temps nécessaire à la création de telle ou telle collection. Je me suis donc limitée dans ce travail aux sites sélectionnés par la BL. Cependant, j'ai informé les bibliothèques nationales du Pays de Galles (NLW) et d'Écosse (NLS) de la création prochaine par la BL d'une collection de blogs. La NLS a souhaité contribuer à ce projet.

La première étape concernant la création de ces collections a été de m'informer des collections précédentes. Il m'a été fourni des documents internes rédigés, l'un par Berverly Kemp de la Women's Library à Londres et l'autre par Joanne Harwood, qui a occupé le poste de Conservatrice de la Collection Amérique Latine à la BL. Ces

16 Page collections IA
17 http://www.bl.uk/collections/britirish/pdf/modbritcdpwebsites.pdf

documents décrivent les étapes de création de collections relatives aux femmes et aux sites latino-américains aux Royaume-Uni ainsi que les objectifs d'un tel travail.

Une dizaine de collections initiées par la BL sont accessibles sur le site d'UKWAC et leur aspect, contenu ainsi que la manière dont elles ont été créées varient d'une collection à l'autre.[18] Certaines collections sont destinées à croître parce qu'elles couvrent un sujet susceptible d'évoluer, comme par exemple la collection « Women's Issues ». D'autres, comme « UK General Election 2005 », sont figées dans le temps et visent à préserver les sites relatifs à un fait particulier de l'actualité.

Il m'a donc fallu, après m'être familiarisée avec le fonctionnement du projet et les outils employés, faire l'analyse de l'existant des sites déjà sélectionnés dans les domaines qui m'intéressaient, à savoir, les blogs, les sites religieux et les sites d'art.

Cette tâche a pris un temps assez long, d'une part parce que l'indexation dans PANDAS est assez limitée et ne facilite pas la recherche, mais surtout parce que le classement de ces sites n'est pas normalisé. Les documents destinés à aider les personnes qui sélectionnent existent, cependant, les descripteurs décris dans l'aide à la recherche du site Internet ne sont pas toujours suivis et repris. De plus la classification n'est pas toujours adaptée aux sites Internet.

C'est pourquoi, il m'a été difficile de les récupérer dans la BDD Access et j'ai dû croiser les équations de recherche à plusieurs reprises pour estimer avoir récupéré la majorité des sites que je recherchais.

Par exemple, le terme « blog » n'avait pas été utilisé de manière systématique pour décrire ce type de site, il m'a donc fallu chercher « log » dans les descripteurs, les sujets et les notes. J'ai ensuite procédé à une recherche par url en utilisant une partie du nom des CMS (Content Management System) spécifiques à la publication de blogs, comme par exemple *wordpress.com* ou *typepad.com*.

Parallèlement à ce travail, j'ai entamé des recherches concernant les domaines qui m'intéressaient, à savoir la religion au Royaume-Uni, les blogs, les questions de copyright en général et plus spécifiquement, concernant les sites d'art. En effet, la politique de la BL est d'être neutre et représentative de la Nation. Il m'a donc semblé qu'étudier l'état de la pratique religieuse au Royaume-Uni, par exemple, était un travail de préparation nécessaire.

18 http://info.webarchive.org.uk/col.html

J'ai ainsi déterminé pour les blogs un certain nombre de sujets et de formats à sélectionner, tel que l'Art, le Web, les blogs utilisant la vidéo ou la photo.
De même pour la religion, j'ai déterminé grâce à un certain nombre d'ouvrages de référence des pourcentages représentatifs de la pratique religieuse au Royaume-Uni afin de les respecter au mieux au moment de ma sélection de sites. Par exemple, j'ai accordé plus de place aux sites méthodistes et baptistes.
Cependant, créer une collection alors que des sites ont déjà été sélectionnés m'a posé quelques problèmes.
D'abord, les sites sélectionnés l'avaient été pour d'autres raisons que les miennes. Par exemple, un blog avait pu être sélectionné pour faire partie de la collection « Women's issues » parce qu'il était représentatif de telle ou telle question concernant les femmes. Cependant, dans une collection nouvellement créée concernant les blogs, son intérêt devenait moindre. D'autres sites avaient été sélectionnés de manière indépendante sans idée d'intégration dans l'une des collections. Il m'a pourtant semblé que je devais inclure les blogs déjà sélectionnés parce que du point de vue du public, il semblait difficile d'expliquer que la collection ne comprenait pas tous les blogs archivés sur le site.
J'ai également rencontré une difficulté concernant les sites généraux comprenant des blogs. En effet, certains sites sélectionnés étaient décris comme contenant un ou plusieurs liens vers un blog. Il m'a donc fallu vérifier que chaque lien existait toujours puis visiter ce lien afin de déterminer la qualité ou l'intérêt du blog par rapport au site. Cette question en a ensuite soulevé d'autres parce que techniquement il est impossible avec HTTrack et PANDAS de créer deux entrées distinctes, l'une pour le blog et l'autre pour le site, sans doubler la capture du site et donc la taille de fichier archivé. Or, UKWAC a prévu un accroissement de l'archivage de sites dans les années à venir, ce qui inclut l'idée qu'il faut sauver le plus d'espace possible dans l'archive pour permettre au projet de grandir.
J'ai dressé une liste des sites de ce type et j'ai proposé différentes approches selon l'intérêt du blog.
Quelques rares blogs ont été proposés à l'archivage avec une entrée séparée parce qu'ils semblaient particulièrement intéressants, par exemple Concealed, Discovered, Revealed du Victoria & Albert Museum.
D'autres ont été classés dans la catégories blogs, parce qu'il me semblait que le blog prévalait sur le site, c'est le cas de The Tinbasher.

D'autres encore ont bénéficié de la création de deux entrées distinctes parce que l'url du blog différait de l'url du site à la racine. Par exemple, *Guide2care Blog* (http://www.blog.guide2care.com/blogs/) qui appartient au site *Guide2care* (http://www.guide2care.com/).

Enfin, pour la dernière catégorie, je n'ai effectué aucun changement, hormis celui de noter de manière claire et normalisée dans la BDD que ces sites possédaient un blog, afin qu'ils puissent être retrouvés facilement.

Si nous avons décidé d'utiliser ces solutions, c'est parce que UKWAC migrera bientôt ses données de PANDAS à WCT Tool qui utilise Heritrix et non HTTrack. Or, Heritrix permettra de pointer directement sur le lien du blog, on pourra donc créer deux entrées sans devoir pour autant créer deux archives.

Ce travail m'a fait entrevoir la difficulté de bâtir une collection alors qu'un certain nombre de sites ont déjà été sélectionnés. Cela implique le fait que cette collection a déjà pris une direction. Tout en la concevant comme un nouveau projet, il m'a tout de même fallu tenir compte du travail déjà effectué.

Un dernier aspect difficile concerne la question du dépôt légal, qui sera développée plus loin. En effet, créer une collection en ne possédant aucun contrôle sur les sites qui seront finalement archivés dans cette collection parce que l'accord aura été donné ou non par leur propriétaire, est extrêmement déstabilisant. J'ai tenté de sélectionner plusieurs exemple de chaque type de site, chaque fois que cela était possible tout en maintenant un niveau de qualité dans mes sélections. Cependant, il m'a été impossible de m'assurer à cent pour cent que chacune de ces catégories serait représentée de manière équitable.

1.2 Sélection

En ce qui concerne la sélection, il m'a été demandé de suivre la politique de développement de collection de la BL tout en l'exploitant au maximum selon le type de collection.

La conservatrice, Alison Hill, m'a fourni les documents utilisés par les sélectionneurs de la BL afin que je m'en imprègne et que, tout en les respectant, je puisse souligner certains aspects spécifiques à chaque collection.

Je me suis donc appuyée, tout comme pour ma réflexion préalable à la constitution d'une collection, sur la politique de développement[19] conçue par le Dr. Richard Price ainsi que par Ms Alison Hill, dont le résumé contient les principales règles à suivre:

> *« The British Library will collect sites selectively **from the UK web space by prioritising** the archiving of sites of **research value** across the spectrum of knowledge. In addition, the Library will archive a selection of sites which are representative **of British cultural heritage** in all its **diversity** across the regions. It will also archive a small number of sites which demonstrate **web innovation. »***

J'ai utilisé ce document comme un ensemble de critères à suivre pour élaborer mes deux collections. En contrôlant de manière régulière la base de données et en essayant de couvrir un à un tous les domaines qui y sont notés, je me suis efforcée de rendre ces collections conformes à la politique de la BL.

- **« UK Web space**. »

La sélection se devait donc de collecter en majorité, et sauf exception, des sites:

- ➔ Ayant un nom de domaine « .uk ».
- ➔ Hébergés au Royaume-Uni, de type « .org », « .com », etc.
- ➔ Hébergés hors du Royaume-Uni mais possédant un contenu relatif au Royaume-Uni.

Afin d'être en adéquation avec ce critère, j'ai utilisé différents outils de recherche et de contrôle.

- ➔ Pour les blogs, il suffisait parfois de visiter le profil de l'auteur afin de savoir s'il répondait à ce critère, de même pour la sélection d'églises ou d'organisations religieuses, les coordonnées permettaient le plus souvent de s'assurer que ce point était respecté.
- ➔ J'ai également utilisé des outils tels que les sites de référencement de profils ou les annuaires de sites spécialisés. Ces outils, tels que le British Blog Directory ou

19 Collection Development Policy for UK Websites

Vineyard Churches UK Website, m'ont permis de me focaliser sur les sites Internet appartenant à ce domaine et donc de ne pas perdre de temps.

➔ En cas de doute, j'ai toutefois eu recours à un logiciel libre délivrant le lieu d'enregistrement du nom de domaine ainsi que les coordonnées du propriétaire: DNS Stuff.com.

- **« Research value »**

La politique de la BL prévoit de collecter en priorité des sites Internet pouvant être utilisés dans un futur proche ou lointain par des chercheurs. Ces sites sont parfois difficiles à trouver et sont considérés comme « fragiles » par la BL car ils risquent de disparaître rapidement, on peut associer la plupart de ces sites à la catégorie littérature grise.
Les sites à sélectionner devaient donc répondre à certains critères de qualité et leur contenu devait soit inviter à la réflexion, soit illustrer une idée ou un fait. Par exemple, certains sites contenus dans la collection sur les femmes illustrent le statut de la femme au Royaume-Uni sans pour autant l'analyser.

Ces sites sont en priorité:

➔ Les sites hébergés par des universités (projets de recherche, sites d'enseignants, etc.).
➔ Les sites liés au gouvernement (organisations, campagnes, partis, etc.).
➔ Les associations indépendantes et les organisations caritatives.
➔ Tout site relatif à des projets de recherche indépendants et diffusé par le biais de l'Internet.

La sélection de ce type de sites peut être facilitée en se focalisant sur les institutions citées ci-dessus, à savoir, le gouvernement, les universités, les associations ou organisations (indépendantes ou non). Pour ce qui concerne les blogs, j'ai également utilisé:

➔ Les annuaires spécialisés, comme Blogscholar
➔ Les forums et blogs universitaires, comme Warwick Blogs
➔ Les réseaux et projets en ligne de chercheurs, comme NYLON

- **« British cultural heritage »**

Cet aspect a été majeur dans l'élaboration de mes deux collections parce que les sites religieux comme les blogs sont extrêmement représentatifs de ce critère. Par héritage culturel britannique, il faut comprendre:

- ➔ Représentatif de la diversité culturelle au Royaume-Uni.
- ➔ Représentatif de la diversité régionale.
- ➔ Toute institution ou entreprise représentative de la nation.
- ➔ Reprenant des évènements marquants de la vie des britanniques (tsunami, attaques terroristes de Londres, J.O, etc.).

Ce critère semble de prime abord simple à satisfaire, cependant, parce que l'héritage culturel est une acception large, le critère de sélection doit être plus précis que pour les autres. Je me suis aidé pour cette sélection de:

- ➔ Moteurs de recherche ou annuaires spécialisés, notamment pour effectuer des recherches par région, comme Manchester and the Northwest Region, Britblog.com
- ➔ Ouvrages spécialisés, comme le *UK Christian Handbook*[20], ou l'ouvrage de Tim Worstall[21] concernant les blogs.
- ➔ Les « Awards » sur Internet, comme Irish Blog Awards
- ➔ Les classements publiés par les journaux et magazines, comme Faith Central, Times Online
- ➔ Les aggrégateurs de sites, comme Blogdom of God
- ➔ Le personnel spécialisé de la British Library. J'ai travaillé en particulier avec le Dr. Clive D. Field pour la collection religion, et ai bénéficié des conseils avisés de Mme Helen Robins.

- **« Web innovation »**

La collection de blogs démontre l'innovation du Web du point de vue technologique parce que de nombreux logiciels sont développés pour l'utilisation des bloggers. Il existe également des outils simplifiés de publication sur Internet, de création de statistiques,

20 Uk Christian Handbook
21 2005 Blogged, Dispatches from the blogosphere

etc. Mais l'innovation Web apparaît aussi dans les nouveaux modes de communication (techniques, communautés, vocabulaires, codes, etc.) et cet aspect est valable pour tous les types de sites.
Pour ce critère, j'ai donc insisté sur:

- ➔ L'innovation technologique dans le domaine de l'art, de l'enseignement, du commerce, des logiciels, etc.
- ➔ Les nouveaux modes de diffusion de l'information.
- ➔ Les nouveaux modes de communication.

L'innovation est partout sur Internet, cependant certains sites m'ont particulièrement aidés:

- ➔ Les sites d'universités, pour ce qui concerne les projets innovants d'étudiants ou de professeurs ainsi que les projets d'e-learning, par exemple: Media @ LSE
- ➔ Les Awards spécialisés, par exemple: Photoblog Awards
- ➔ Les sites axés sur l'innovation, par exemple: Virtual Christianity
- ➔ Les sites personnels de professionnels en Sciences de l'Information, par exemple: Karen Blakeman's Blog

En plus de cette politique de développement, la British Library sélectionne en tenant compte des autres membres du Consortium, des limitations techniques et des impératifs liés à la politique générale de la bibliothèque.

- **Membres du Consortium**

Comme cela a été spécifié dans la première partie de ce document, chaque membre du consortium sélectionne dans un domaine qui lui est propre et qui a été défini au début du projet:

- ➔ **Joint Information Systems Committee (JISC)**: Collecte les sites en rapport avec des projets financés par JISC ou le Higher Education Funding Council for England (HEFCE). En général ces projets sont liés au domaine universitaire et à l'éducation en général.

- ➔ **The National Archives (TNA):** Les Archives Nationales collectent des sites gouvernementaux en rapport avec la défense, la politique étrangère, la justice, la sécurité, les différentes administrations et les différents services de l'État.

- ➔ **The National Library of Scotland (NLS):** La Bibliothèque Nationale d'Écosse est chargée de sélectionner des sites relatifs à l'histoire et à la culture écossaise en général.

- ➔ **The National Library of Wales (NLW):** La Bibliothèque Nationale du Pays de Galles collecte des sites représentatifs de la culture et de l'histoire du Pays de Galles dans un objectif de recherche.

- ➔ **The Wellcome Library:** Collecte des sites du domaine national, relatifs au domaine de la santé et plus particulièrement de la médecine.

Il existe une politique de développement des collections pour chacune des institutions ainsi que des moyens de vérifier que l'on respecte ces critères:

- ➔ En consultant les politiques de développement transmises par chaque institution à la BL.
- ➔ En consultant la base de données de PANDAS et les listes de sites échangées entre les membres.
- ➔ En vérifiant sur chaque site sélectionné si l'un des membres le finance ou y est impliqué d'une manière ou d'une autre.

- **Critères techniques**

Il existe un certain nombre de contraintes techniques qui peuvent limiter ou rendre difficile l'archivage et sont donc à prendre en compte lors de la sélection.

- ➔ Flash
- ➔ Streaming (non téléchargeable)
- ➔ Javascript
- ➔ BDD dynamiques
- ➔ Taille du site
- ➔ CSS

➔ Calendrier, commentaires, etc.

Toutes ces limitations seront développées dans la partie suivante. Certaines posent problème mais ne bloquent pas totalement l'archivage. D'autres ne peuvent être gérées par PANDAS.

- **Contenu sensible**

En tant que bibliothèque nationale, la BL se doit d'être neutre et représentative de chacun. En ce qui concerne la sélection de sites politiques et religieux, ce statut m'a parfois posé problème.
Le principe d'archivage du Web et de création de collection impliquait de mon point de vue, un minimum de censure quant à la sélection. En effet, chaque site collecté doit être représentatif d'un état de la société et il est extrêmement délicat d'exclure des sites parce qu'ils pourraient être jugés inappropriés dans le cadre d'un archivage par la BL.
Il m'a semblé très important de respecter une équité dans le nombre de sites de tel ou tel parti politique ou mouvement religieux, cependant, certains sites un peu hors norme ou plus tranchés m'ont parfois semblé dignes d'intérêt. Lorsque cela a été le cas, je les ai présentés à Alison Hill en lui exposant les raisons de mon choix afin qu'elle détermine si oui ou non ces sites devaient être archivés.
D'autres sites, au contraire me semblaient parfois de qualité moindre parce que peu innovant et originaux. J'ai pourtant décidé d'en inclure quelques exemples afin de représenter au mieux l'état de la publication sur Internet, qu'il s'agisse de blogs ou de sites religieux.
Le journal intime tant décrié[22] pour ce qui concerne l'exemple des blogs est ainsi présent sur le Web comme d'autres formes de blogs parfois plus intéressantes. La collection se devait de respecter les impératifs de sélection de la BL certes mais aussi de montrer le Web tel qu'il est. Inclure du bon et du moins bon a donc été un parti pris et un gage de représentativité. Ce choix a d'ailleurs été validé et soutenu par l'équipe.

- **Méthode de sélection**

22 Andrew Keen

Avant toute sélection, il m'a fallu répertorier les sites déjà archivés par la BL ou entrés dans la BDD et susceptibles d'intégrer l'une de ces deux collections. La création de ces collections ne partait pas d'une collection vide et répertorier les sites déjà archivés ou sélectionnés et en attente de l'être, permettait de respecter les autres impératifs décris dans cette politique, tels que la diversité par exemple.
J'ai donc listé (à l'aide d'un fichier Excel) et visité les sites récupérés dans la BDD et dans PANDAS. Ce type de fichier m'a permis de faire l'état de ce qui était possédé et de l'intérêt de ces sites. Les sites non archivés ont été par la suite recontactés et les sites ayant disparu ont été notés comme « vanished », dans la BDD.
Grâce à ce premier travail de classement, j'ai pu dégager quelles catégories de sites étaient déjà sélectionnées tout en me familiarisant avec la base Access ainsi qu'avec le fonctionnement de PANDAS.

Deux constats en sont ressortis:

- ➔ les blogs étaient en majorités des blogs liés à des collections, telles que « UK General Election 2005 » ou « Women's Issues ».
- ➔ La catégorie de sites religieux ne contenait quant à elle que quelques exemples de sites Non Conformistes (moins de dix), ces sites étant ceux des principales Églises.

La deuxième étape de ce travail de sélection a consisté à trouver des sources à la fois abondantes et sûres afin de sélectionner un maximum de sites dans un temps court. Les sources utilisées pour ce travail proviennent soit des recommandations des spécialistes de la BL, soit de mes recherches personnelles.
Je me suis par exemple appuyée sur deux ouvrages répertoriant les blogs les plus intéressants ou connus au Royaume-Uni[23], fournis à mon arrivée par la conservatrice, Alison Hill.
Après cette sélection de départ, tant pour les blogs que pour les sites de religion Non Conformiste, je me suis appliquée à élargir ma sélection et à la diversifier.
En effet, les blogs répertoriés dans les ouvrages cités précédemment traitaient pour la plupart de politique, ils étaient publiés en majorité par des hommes dans leur trentaine et habitant Londres.

23 Tim Worstall & Justin McKeating

Cette centaine de sites orientaient donc la collection de manière notable et m'éloignait de l'objectif de représentativité de la culture britannique. De plus, les formats et technologies utilisés dans la réalité n'étaient que peu représentés.
Pour les sites religieux, ma sélection de départ partait de l'une des sections d'un ouvrage de référence, le *UK Christian Handbook,* qui répertorie les églises des mouvements religieux, cette première sélection, si elle m'a permis de me familiariser avec ces mouvements m'a apporté une sélection plutôt homogène, nécessitant d'être diversifiée par la suite.

La troisième étape a donc été de confronter cette première sélection aux objectifs de représentativité, de recherche et d'innovation que je m'étais fixés pour chacune des collections.
J'ai donc organisé les sites sélectionnés de la manière suivante:
Religion:

- Nom / url
- Religion
- Type de site
- Type d'organisation
- Lieu géographique
- Sujet
- Public
- Notes éventuelles

Blogs:

- Nom / url
- Format
- Outils techniques
- Sujet 1
- Sujet 2
- Sujet 3
- Auteur(s) / organisation
- Lieu géographique
- Notes éventuelles (Creative Commons, intérêt particulier, etc.)

Ce type de classement a évolué dans le temps lorsque je découvrais de nouvelles catégories utiles.

La base de données, si elle m'a facilité la tâche, n'a pas tout automatisé parce que tous les champs n'y étaient pas contenus. Pour les sites religieux, l'aspect finalement prédominant a été de respecter le pourcentage de personnes pratiquant ces religions aux Royaume-Uni et donc le nombre de sites.
Pour les blogs en revanche, il m'a fallu créer une autre bases de données plus spécifique et contenant les champs nécessaires. Par exemple, il était important que les descripteurs sujets apparaissent en tant que sujet 1, sujet 2 et sujet 3, ceci afin de bien comptabiliser chaque sujet traité dans un même blog. Cela m'a aidé à avoir une idée plus précise de la collection que je créais.
De même, l'utilisation des bases de données m'a fourni une comptabilité précise des propriétaires de sites ayant autorisés l'archivage. J'ai pu ainsi rapidement constater les disparités entre les sites sélectionnés et ceux effectivement présents dans la collection. Par exemple, le taux de réponses positives des sites de quakers était très élevé au contraire des sites protestants, je me suis donc adaptée en essayant d'en sélectionner davantage et de les relancer par courrier.

Après avoir défini de nouveaux angles de recherche pour ces deux collections, j'ai utilisé différentes techniques, en plus de celles déjà citées.
L'une des techniques de sélection qui m'a été utile dans les deux cas a été l'utilisation des liens externes présents dans pratiquement tous les sites religieux et dans les blogs, sous forme de liens recommandés par l'auteur (ou blogroll).
Cependant, cette méthode de sélection comporte le désavantage de l'homogénéité des sites récoltés. En effet, pour la plupart, les liens recommandés par les auteurs renvoient vers des blogs traitant d'un sujet similaire. Les auteurs de blogs (ou bloggers) sont en effet souvent organisés en « communautés virtuelles » afin d'être plus visibles mais surtout afin de partager des informations et d'échanger sur les sujets qui les intéressent.
Il en est de même pour les sites religieux, la plupart renvoient à des organisations appartenant au même mouvement.

Dans les deux cas ces organisations en « communautés virtuelles » ou non, m'ont à la fois été utile car j'ai pu retrouver facilement les communautés que je recherchais et à la fois desservie, lorsque je tentais d'élargir ma sélection. Ceci était d'autant plus vrai pour les blogs puisque les blogs les plus en vue dans la communauté de bloggers (ou blogosphère), ont un indice de citation très élevé.

Enfin, comme je l'ai expliqué plus haut, j'ai souvent eu recours aux moteurs de recherche, annuaires spécialisés et sites d'Awards, tout simplement parce qu'ils condensent en général le meilleur de tel ou tel domaine.

Par exemple, afin de trouver des blogs édités par des femmes, j'ai utilisé des moteurs de recherche et des annuaires spécialisés du type de Girls Blog UK.

Pour trouver davantage de sites méthodistes, j'ai eu entre autre recours au Methodist Recorder Online.

Toutes ces méthodes m'ont permis de m'adapter et de réorienter constamment mes recherches.

Au terme de mon stage, ma pré-sélection de sites Non Conformistes s'élevait à environ 250 sites dont 159 ont été sélectionnés, soit environ 63%.

En ce qui concerne les blogs, environ 500 sites ont été collectés au départ et 364 sont effectivement sélectionnés, soit environ 72 %.

Il me semble donc que le travail de préparation et l'analyse constante de l'évolution des deux collections m'ont permis de gagner en efficacité lors du processus de sélection.

1.3 Classification

Comme expliqué précédemment, la sélection est en général effectuée par un certain nombre de spécialistes de la BL, tous bibliothécaires ou conservateurs. Ces personnes dédient quelques heures par semaine à la sélection de sites Internet mais cela ne constitue évidemment pas la majorité de leur temps de travail.

Une réunion mensuelle de ces spécialistes, dirigée par la conservatrice, permet de faire le point sur les objectifs par mois de sites à sélectionner ainsi que sur les méthodes à employer pour la sélection et la classification.

Il leur est également fourni un modèle de fiche[24] à remplir contenant les champs de la BDD Access, de cette manière, chacun, selon sa disponibilité essaie de rendre une sélection organisée. Cependant, tous ne fournissent pas tous les détails relatifs aux sites qu'ils ont sélectionnés.
De plus, certains sites sont proposés à l'archivage par d'autres personnes, soit faisant partie d'une des institutions du Consortium, soit par le biais du formulaire d'auto-sélection mis en place sur le site d'UKWAC. Comme des livres sont proposés à la BL dans le cadre du dépôt légal, des sites sont proposés ou recommandés à l'archivage[25].

Le classement dans la base de données de ces fiches plus ou moins bien complétées est logiquement effectué par le « Permission officer ». Or, son principal travail est de gérer la BDD et de s'occuper de la correspondance. La classification des sites ne constitue donc pas une priorité. Les descripteurs liés aux sites sont donc le plus souvent corrigés ou repris au moment de l'insertion dans PANDAS.

Après avoir effectué un travail de recherche dans la base de données et dans PANDAS afin de repérer et de comptabiliser les sites pouvant être intégrés aux collections de blogs et de sites religieux, je me suis aperçue que cette multiplicité d'acteurs dans le processus de classification rendait difficile tout travail de recherche.
La question s'est alors posée de savoir comment décrire et classer les sites de ces deux collections.

Par exemple, la catégorie sujet « blog » n'existant pas, les sélecteurs avaient parfois classés ces derniers dans la sous-catégorie « journals » dont le sens premier est « périodiques ». D'autres avaient utilisés comme descripteur le sujet dominant du blog, en ajoutant dans le champ « description » ou « note », qu'il s'agissait d'un blog ou d'un site personnel.
En concertation avec Alison Hill, j'ai décidé d'utiliser le descripteur « journals » uniquement pour les journaux intimes en ligne, j'entends par là, les blogs qui s'intéressent en majorité à des évènements intimes de la vie quotidienne.
Cette décision n'est pas idéale parce que le classement utilisé par UKWAC fait que les « journals » font davantage référence aux journaux électroniques et sont de ce fait

24 Fiche Sélecteur
25 http://info.webarchive.org.uk/cgi-bin/submission.cgi

classés dans la rubrique: « Reference Works » ou Travaux de référence. Cependant, le site et la classification utilisés étant tous deux en passe d'être complètement modifiés, nous avons considéré que modifier la classification dans le site actuel n'était pas une priorité.
De plus, chaque nouvelle collection créée est accessible depuis la page d'accueil et sur chaque page de résultats ce qui rend les sites composant les collections très visibles et facilement accessibles. Chaque collection créée par la BL bénéficie d'ailleurs d'un traitement particulier puisqu'elles sont intégrées au catalogue de la bibliothèque.

Toutefois, j'ai proposé un certain nombre d'améliorations de cette classification lorsque cela concernait directement les collections sur lesquelles je travaillais.
En effet, la classification utilisée par les membres d'UKWAC a été créée au début du projet. Les conservateurs l'ayant créée se sont appuyés sur Dewey, tout en essayant de spécifier au mieux les domaines qu'ils imaginaient avoir à décrire.

Une page d'aide détaille cette classification[26]:

Arts & Humanities

- ➔ Architecture
- ➔ Dance
- ➔ Fine and applied arts
- ➔ Geography
- ➔ History
- ➔ Languages
- ➔ Literature
- ➔ Music
- ➔ Philosophy
- ➔ Religion
- ➔ Theatre

Business & Economy

- ➔ Economics

26 Copyright UKWAC

- includes theory, economic systems, econometrics

➔ Trade & commerce
- includes international economic relations, integration, globalisation

➔ Labour economics
- includes employment, unemployment, trade unions

➔ Financial economics
- includes public and private finance, banking, stock market, financial services

➔ Industries
- includes agriculture and transport

➔ Company web sites A-Z

➔ Economic development
- includes sites covering all aspects of Third World development

➔ Management
- includes HRM

➔ Marketing; Market Research

Education & Research

➔ Pre-school education; early years

➔ Schools
- includes primary and secondary

➔ Special needs education

➔ Lifelong learning; adult education

➔ Vocational education and training [post-16]

➔ Higher education

➔ Museums & libraries
- for art galleries see Arts

Government & Politics

➔ Central government Whitehall

➔ Central government agencies

➔ Devolved administrations

➔ Local government

➔ Politics

- includes political theory, political systems, elections

➔ Civil and political rights

➔ Political parties & politicians

➔ International relations

- includes terrorism

Health

➔ Medicine

- includes medical research, evidence-based movement

➔ Allied & complementary medicine

➔ Conditions & diseases

➔ Regulatory & advisory bodies

➔ Consumer/patient support & advocacy sites

➔ Healthcare professionals

- includes professional bodies

➔ Healthcare services; NHS

➔ Public health and safety

News & Media

➔ Newspapers; newswires

➔ Broadcast media

➔ Film

Science & Technology

➔ Life sciences

➔ Environment

- includes earth sciences, meteorology, oceanography, pollution, climate change, etc.

➔ Physical sciences

- includes physics and chemistry

➔ Engineering

➔ Technology

➔ Mathematics

➔ Computing & IT; the Web

- ➔ Popular science/public engagement

Society & Culture

- ➔ Sociology & anthropology
  - ○ theoretical aspects, includes social groups and classes, culture
- ➔ Social problems & welfare
  - ○ for health services see Health
- ➔ Communities
  - ○ includes urban studies, rural communities, community regeneration, town and country planning; housing
- ➔ Demography
  - ○ includes immigration & emigration
- ➔ Law; legal system
  - ○ includes the courts
- ➔ Crime & criminology
  - ○ includes policing, penology
- ➔ Psychology
- ➔ Sports & recreation; Festivals & Event

Reference Works

- ➔ General encyclopaedias & fact books
- ➔ Dictionaries
- ➔ Statistics
- ➔ Directories
- ➔ Journals
- ➔ Bibliographies & bibliographic databases; indexes
- ➔ Biographical sites

Cependant, cette classification n'a pas évolué depuis le début du projet et elle contient peu de descripteurs spécifiques ou relatifs à Internet (en rouge).

Concernant les sites sélectionnés pour ces deux collections, j'ai pourtant systématiquement utilisé les descripteurs contenus dans cette aide afin de respecter le

système de classification actuel. J'ai également effectué un travail de normalisation afin que les sites puissent être retrouvés aisément lors d'une prochaine recherche.

Je me suis aussi employée à proposer un certain nombre de descripteurs davantage liés à Internet, et notamment pour les blogs.
En accord avec l'équipe, j'ai fourni de nouveaux descripteurs, tels que « Web communities » pour les sites communautaires sur le Web ou encore en reprenant la terminologie spécifique aux blogs, « photoblog » pour les blogs composés de photos ou « disseration blog » pour ceux utilisés par les universitaires lors de la rédaction de leur thèse afin de partager leur travaux de recherche et de dialoguer. Cependant, ces ajouts ont été fait uniquement dans la base de données de la BL et marqués entre parenthèses de manière à ce qu'ils ne soient pas repris dans PANDAS.
Pour les sites de religion Non Conformiste, la collection créée ne sera pour l'instant pas publiée, l'onglet collection ne pouvant être utilisé, j'ai ajouté « Nonconformist » de manière systématique dans les descripteurs. Ainsi, au moment de la publication, il sera facile à l'équipe de récupérer ces sites et de les regrouper en une collection.
J'ai également inclus dans cette collection, conformément à la politique de la BL, un certain nombre de sites montrant les nouveaux modes de communication des groupes religieux ainsi que l'utilisation des nouvelles technologies. Les églises virtuelles en sont un bon exemple et ont été classées dans la catégorie religion avec l'ajout du descripteur « virtual churches » entre parenthèses.

L'intérêt de cette archive réside dans la préservation de matériaux utiles aux futurs chercheurs. Il m'a donc semblé que montrer les caractéristiques propres aux sites Web ainsi que l'utilisation sociale des technologies mises à disposition des internautes constituait l'un des points essentiels.
L'accès à ces informations se doit d'être facilité par l'élaboration prochaine, en plus des collections, d'une classification permettant d'accéder aux documents à la fois par sujet et à la fois par type de matériaux. Malheureusement, ce projet de classification est pour l'instant suspendu. Cela est dû à l'évolution de l'archive vers le Web Curator Tool.
Dans un premier temps et avant toute personnalisation par l'équipe d'UKWAC du nouveau site Internet, la recherche d'archive se fera par url uniquement ou au mieux en texte intégral.

Cependant, une nouvelle classification se devra d'accompagner l'élaboration prochaine d'un nouveau site Internet contenant des modes de recherches plus élaborés. Si la BL est la partie leader de ce projet, il est évident que toute modification concernant le site ou le système de classification devra être décidée d'un commun accord entre tous les membres d'UKWAC.

Mon soucis premier n'a donc pas été de modifier la classification utilisée par la BL même si autant lors de mes recherches de sites dans la base de données que lorsque j'ai dû y entrer les données relatives aux nouveaux titres, il m'a semblé que cette classification était peu appropriée aux matériaux qu'elle traitait.

Les autres projets archivant le Web utilisent bien souvent une classification par format puis par sujet ou collection et les thésaurus applicables à Internet sont un sujet de préoccupation pour de nombreuses bibliothèques. Or, il m'a semblé lors de l'archivage de blogs par exemple, que ne pas les classer selon leur format constituait une perte importante d'information.

Gardant en mémoire que cette archive est destinée à des fins de recherche, j'ai tenté de montrer dans les descriptifs de ces sites leur aspect technique autant que le sujet qu'ils traitaient. J'ai aussi transmis à l'équipe les différents glossaires, thésaurus et annuaires trouvés sur Internet durant mes recherches.

2 Le dépôt légal

2.1 Situation

La question du dépôt légal est centrale en ce qui concerne l'archivage du Web au Royaume-Uni comme dans de nombreux pays européens.

Au Royaume-Uni, le dépôt légal concernant les contenus numériques n'est pas encore légalisé. En 2004, alors que le projet UKWAC débutait, les membres du Consortium et en particulier la BL, espéraient que la loi concernant le dépôt légal évoluerait rapidement. En effet, le « Copyright and Related Act » de 2000 suivi du « Legal Deposit Libraries Act 2003 » montraient une évolution favorable de la loi.

Le « Legal Deposit Libraries Act 2003 » a pris effet de manière officielle en 2004. Il constitue une proposition de loi qui permettrait à la BL ainsi qu'aux autres bibliothèques en droit de réclamer le dépôt légal, de récupérer, de préserver et de permettre l'accès aux matériaux non analogiques, en ligne ou non, de type CD, DVD, sites Internet, etc.
Les autres bibliothèques pouvant prétendre au dépôt légal de documents selon le Copyright Act 1911 sont:

- Bodleian Library, Oxford
- University Library, Cambridge
- **National Library of Scotland**
- Library of Trinity College, Dublin
- **National Library of Wales**

Cependant, seule la British Library peut prétendre à un dépôt légal de document automatique par les éditeurs. Le délai de ce dépôt est fixé à un mois par la loi et est obligatoire.
En donnant systématiquement une copie de chacune de leur publication à la BL, les éditeurs permettent que les ouvrages soient accessibles au public dans les salles de lecture et préservés par la BL dans une visée patrimoniale. Ces documents sont intégrés au catalogue en ligne de la BL et la plupart sont également référencés par la BNB (British National Bibliography) au format MARC, ce qui rend ces ressources bibliographiques disponibles au niveau international.
En 2005, un comité indépendant a été mis en place afin de décider d'une possible évolution de cette loi vers les matériaux numériques.
Le « Legal Deposit Advisory Panel » (LDAP) s'est donc réunit pour la première fois en septembre 2005. Son rôle est de conseiller le gouvernement dans l'élaboration de cette loi en considérant tous les aspects liés à ces nouveaux matériaux, l'un des aspects majeurs étant de les définir de manière à créer une loi utile et performante pour les années à venir. En effet, l'hétérogénéité de ces matériaux rend difficile l'élaboration d'une loi unique et viable dans le temps. Le ministère de la culture ainsi que les comités indépendants, aidés des institutions telles que la BL et des éditeurs, travaillent à cerner

au plus vite les questions inhérentes à ces problèmes. Les enjeux sont importants puisqu'une quantité non négligeable de documents notamment produits par le gouvernement est perdue chaque jour.

Au Royaume-Uni, toute oeuvre originale est protégée par un certain nombre de copyrights. Selon l'Intellectual Property Office[27], le copyright donne au créateur d'une oeuvre un droit de contrôle sur les matériaux créés. Cependant, ce droit n'existe qu'à partir du moment où l'oeuvre est enregistrée sur un support que ce soit par écrit ou par tout autre moyen. L'idée de l'oeuvre n'engendre pas de droits sur cette oeuvre, par contre, il n'est pas nécessaire de s'inscrire, une fois l'oeuvre originale créée, pour bénéficier du copyright. Il n'est pas non plus obligatoire de mentionner que l'oeuvre est protégée par le copyright sur le support.

Les droits couverts par le copyright sont:

- La reproduction
- L'adaptation
- La distribution
- La communication au public par voie électronique ou numérique
- La location ou le prêt de copies au public
- La représentation devant un public

Il est également spécifié que dans la plupart des cas, le nom de l'auteur doit être cité lors de toute reproduction. Ce dernier a également le droit de s'opposer légalement à toute mutilation ou modification, même légère, de son oeuvre.
Le copyright s'applique à toute oeuvre originale qu'elle appartienne à la littérature, au théâtre, à la musique et aux travaux artistiques en général, à la publication de travaux de recherche, aux enregistrements sonores (en incluant les CD), aux films (en incluant les DVD et les vidéos) ou aux formes de diffusion des données comme la télévision, la radio, Internet, etc.
A titre d'exemple, sont compris dans cette liste, les programmes informatiques, les bases de données et les sites Internet.

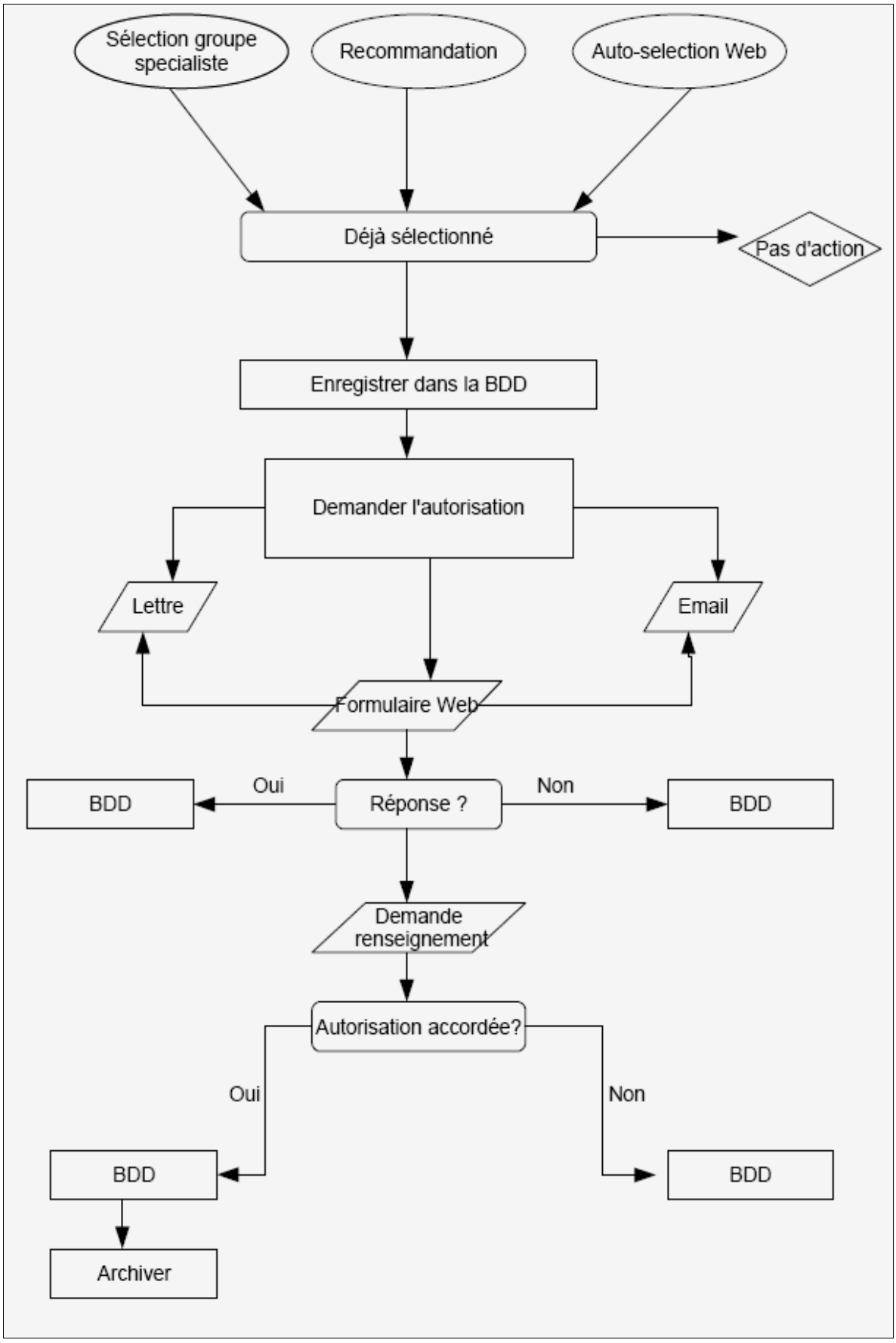
27 http://www.ipo.gov.uk/copy.htm

Dans le cas où la création a été effectuée pour un employeur, par exemple un programme informatique, elle appartient à l'employeur. Le copyright est de l'ordre de la propriété intellectuelle, il peut être cédé pour un temps court ou définitivement, il peut être hérité, acheté, etc. En règle générale, le copyright perdure pendant 70 ans après la mort de l'auteur. Pour les enregistrements sonores ou modes de diffusion de données autres, la protection est en général de 50 ans, sauf exception ou condition spécifique.
Évidemment, les dérogations sont nombreuses tout comme les cas particuliers et le propos n'est pas ici de les détailler mais de comprendre en quoi une entreprise telle que l'archivage du Web se révèle complexe sans l'établissement d'une loi de dépôt légal pour une institution telle que la BL.
En effet, les auteurs et éditeurs de sites Internet bénéficient du copyright et leur création originale, c'est-à-dire leur site Internet, ne peut être copiée par la BL sans leur accord. Être en mesure de copier ces sites tout en ne les diffusant pas jusqu'au passage d'une loi le permettant, serait pourtant le minimum à espérer lors d'une entreprise de préservation du Web comme celle d'UKWAC.

2.2 Processus

Ainsi, en attendant l'élargissement du dépôt légal aux sites Internet, les membres d'UKWAC, hormis The National Archives et JISC dans la plupart des cas, doivent demander par écrit l'autorisation d'archiver chaque site Internet les intéressant.

Le processus d'acquisition de cette autorisation peut se schématiser ainsi:



Lors de la création du Consortium, un document visant à acquérir une licence permettant de copier, d'archiver et de diffuser les sites Internet a donc été élaboré[28]. Ce document est utilisé par chaque membre d'UKWAC chaque fois qu'il souhaite archiver un site Internet. Bien évidemment, si le site est édité par plusieurs auteurs, il est nécessaire d'obtenir leur signature.

Cette question de la multiplicité des auteurs / contributeurs à un site Internet m'a posé problème lors de la création des collections de blogs et de sites Non Conformistes. Elle a également fait l'objet d'un travail sur les sites d'art sélectionnés par la BL.

Comme je l'ai expliqué, sans légalisation du dépôt légal pour les sites Internet, le travail de sélection effectué par la BL reste pour une part importante sans suite. Les sites sélectionnés sont enregistrés dans la base de données et les auteurs sont relancés tous les six mois à un an afin qu'ils signent la licence. Malheureusement, dans de nombreux cas, cette licence n'est jamais retournée à la BL et ce pour plusieurs raisons.

D'abord, les demandes de licence sont envoyées par courrier électronique, par voie postale ou encore en remplissant un formulaire de demande de renseignement sur certains sites. Cependant, pour contacter ces sites, il faut pouvoir trouver les renseignements nécessaires par le biais du site. Certains logiciels libres cités précédemment permettent parfois de retrouver le nom et l'adresse électronique du propriétaire du site. Cependant, la tâche est parfois plus délicate qu'il n'y paraît car de nombreux propriétaires cachent ces informations pour éviter les courriers mal intentionnés ou intempestifs des « spammeurs ».

Mon étude des sites d'art a également montré qu'il était parfois difficile de contacter le bon interlocuteur lorsqu'il s'agissait de sites institutionnels ou simplement d'organisations importantes. L'une des hypothèses concernant le peu de réponses reçues par la BL est que les courriers n'atteignent probablement que très rarement la personne habilitée à accorder cette licence. D'ailleurs, le nombre de réponses négatives en ce qui concerne l'archivage de sites par la BL est minime. D'après les dernières statistiques, sur 5516 demandes de licence envoyées depuis le début du projet (relances comprises), 1419 ont été accordées et 38 refusées.

Cependant, lorsque l'on parvient à atteindre le bon interlocuteur, la question du copyright reste prégnante. La plupart des courriers reçus par la BL demandant des explications à propos du projet concernent en effet des questions de copyright. Les

28 Licence

propriétaires de sites ne comprennent en effet pas toujours ce qu'ils accordent en signant la licence. Certains souhaitent donc savoir s'ils conservent leurs droits d'auteur sur le site et si, par exemple, ils ne seront pas entravés dans leur liberté éditoriale.

Mais le cas de figure qui pose le plus de problèmes est celui des sites qui soit sont co-édités (plusieurs auteurs / éditeurs), soit utilisent régulièrement la citation, soit reprennent les travaux ou oeuvres d'autres auteurs dans leur intégralité.

En ce sens, les sites d'art ont été ma préoccupation première, parce que le taux de réponses pour ces sites était bien en dessous de la moyenne. Le document de licence contient en effet une case à cocher demandant au détenteur du copyright de dire si certains contenus de son site appartiennent à d'autres, du point de vue du droit d'auteur, et si oui, s'il en a acquis les droits. Le document précise aussi que si le propriétaire du site utilise du contenu appartenant à un tiers sans en détenir les droits, son site ne sera pas archivé par le Consortium.

Dans le cas des sites d'art, il est pratiquement impossible de cocher cette case. Les institutions telles que les musées par exemple, possèdent les droits nécessaires à la reproduction des oeuvres sur leur site Internet tout comme ceux permettant de les exposer. Cependant, il leur est impossible de cocher une case spécifiant qu'elles détiennent le copyright pour les oeuvres exposées, car cela est rarement le cas. Étant donné le nombre d'oeuvres exposées dans de telles institutions, il est également impossible de demander une autorisation à chaque auteur ou ayant droit.

Certains propriétaires de sites estiment pouvoir cocher cette case après avoir publié un encart expliquant que le site va être archivé par la BL et que toute personne ne souhaitant pas faire partie de l'archive doit se manifester.

Si ceci n'est pas une procédure légale, il m'a semblé en découvrant de tels exemples que ce procédé était peu risqué dans le cas de blogs auxquels les lecteurs avaient été heureux de participer. D'autant plus que la plupart des auteurs et des courriers reçus durant ces quatre mois me poussent à penser que la plupart sont plutôt flattés de participer à un tel projet et de voir leur travail préservé.

Cette procédure n'est évidemment pas recommandée par UKWAC ou la BL, même si cela a été utilisé lors de la capture de sites relatifs à des collections d'actualité. UKWAC avait alors envoyé un message à chaque propriétaire de site pour les prévenir de la capture prochaine de leur site dans ce cadre. Une majorité de propriétaires de sites ont

accepté bien volontiers ce procédé de « Notice and take down », utilisé entre autre par la Library of Congress.

D'autres sites ont choisi le procédé inverse, à savoir demander une autorisation écrite à chacun des auteurs ayant contribué. C'est le cas du blog *Un-Made-Up*.[29] La licence n'est toujours pas parvenue à la BL ce qui signifie que la totalité des auteurs n'a pas répondu à l'appel de cet éditeur. Ce blog est à l'origine d'un livre ainsi que de réalisations artistiques dans la ville de Brighton et regroupe des nouvelles fondées sur des histoires réelles, rédigées par ceux qui les ont vécues.

Cet exemple est bien la preuve que la plupart des éditeurs de blogs ou de sites personnels se soucient peu des questions de copyright avant que la question de l'appartenance des droits soit posée comme c'est le cas lorsque la BL demande une autorisation.

2.3 Résultats

Le processus d'acquisition d'une licence n'est pas automatisé et ce travail repose sur une seule personne, le « Permission Officier ». Cette personne, je l'ai déjà expliqué, était en cours de recrutement durant mon stage. Or, si envoyer des demandes de licences prend énormément de temps et constitue un travail administratif, il m'a semblé qu'il s'agissait pour moi d'une bonne expérience. En effet, cela m'a permis de me rendre compte non seulement que le taux de réponse peut être plus ou moins faible selon le type de sites sélectionné mais aussi de dialoguer avec un certain nombre de propriétaires de sites et de mieux appréhender leurs questionnements.

Il faut préciser que lors de la demande de licence, le document à compléter et à signer est accompagné d'un document répertoriant les questions les plus fréquemment posées (FAQ) ainsi que d'une lettre type. Lors de mon stage, nous avons décidé de tenter de nouvelles approches et nous avons donc créé une lettre type spécifique à chacune des collections[30].

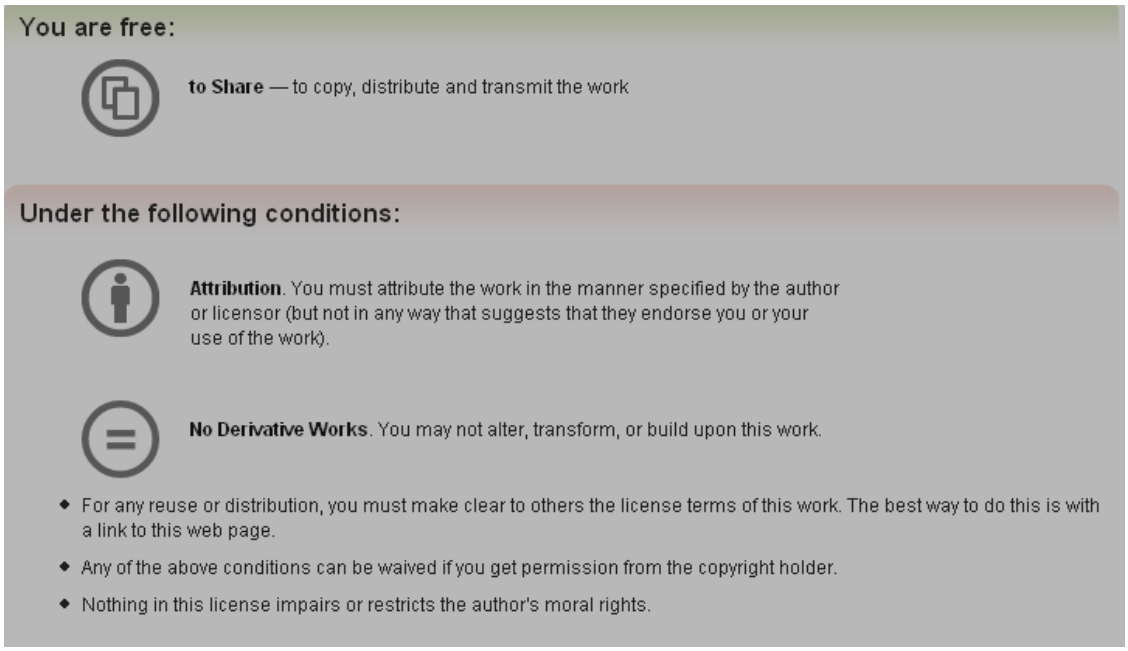
La lettre concernant la collection de blogs a été créée dès mon arrivée car l'expérience avait montré que le fait d'avoir leur site inclus dans une collection était valorisant pour les auteurs.

29 Voir article du 23/05/07
30 Lettres type

En ce qui concerne les sites non conformistes, j'ai d'abord envoyé une lettre type. Le taux de réponses pour ces sites étant extrêmement décevant, j'ai alors fait appel à l'initiateur de cette collection, le Dr. Clive D. Field, qui a proposé de joindre au courrier un paragraphe signé de son nom. J'ai été contactée par des mouvements religieux dès que j'ai ajouté son nom aux courriers.
Pour ce qui concerne les sites d'art, après avoir contacté environ 50 sites, j'ai reçu moins de 5 réponses. J'ai alors décidé de modifier la lettre en incluant une liste des musées ayant acceptés d'être archivés par la BL afin de persuader les institutions que ce projet était important. J'ai également demandé aux personnes susceptibles d'avoir des contacts avec ces institutions à un plus haut niveau hiérarchique de bien vouloir prendre contact. Cependant, durant le mois de juillet, le projet d'entrer en contact avec les musées, entre autre, a dû être abandonné. L'expérience n'ayant pas pu être menée à son terme, il est difficile de dire si ce type d'approche plus active et utilisant d'autres ressources de la BL serait plus efficace ou non.
La question de la « third party » semble tout de même constituer un enjeu de taille, cependant le projet étant mené par des institutions de renom et dans un but purement patrimonial, il me semble que des négociations à un niveau hiérarchique élevé auraient pu faire évoluer la situation favorablement. L'une des solutions également avancée par l'un des musées contactés était de régler le problème à la base. Si les grandes institutions et musées décidaient de modifier leur contrat de licence et si ces oeuvres pouvaient être reproduites et diffusées dans un but patrimonial de façon légale, cette question ne poserait plus de problème.
Il faut préciser ici que la BL et UKWAC en général, ne s'intéressent qu'aux sites accessibles gratuitement et librement par le biais d'Internet. Les sites commerciaux ne sont pour l'instant pas concernés par le projet.

Si le taux de licences accordées a été indéniablement supérieur en ce qui concerne la collection de blogs, un autre aspect m'a aidé à faire grandir cette collection. Les blogs sont publiés par différents profils, au nombre desquels se trouve le profil des auteurs scientifiques ou intéressés par l'informatique et l'Internet. En cherchant à collecter des sites s'intéressant aux nouvelles technologies, j'ai réalisé que le taux de sites protégés par une licence Creative Commons était bien au dessus de la moyenne.

En général, la licence choisie pour les blogs interdisait une utilisation commerciale et autorisait une reproduction ainsi qu'une diffusion du site. Toutefois, elle interdisait parfois la modification du site. Le photoblog headphoneland utilise par exemple la Licence « Attribution-No Derivative Works 3.0 Unported » dont les caractéristiques sont reproduites ci-dessous:



Bien évidemment, le fait qu'un blog soit sous Creative Commons n'a jamais suffi à sa sélection cependant, grâce à ce genre de licence, le travail était facilité et cela a représenté un gain de temps parce que la correspondance devenait inutile.
Cependant, ces sites restent minoritaires dans la collection de blogs et sont pratiquement inexistants dans celle des sites Non Conformistes.

Au terme de ce travail, j'ai tenté de distinguer différents cas récurrents de non réponse ou de refus au cours de ces quatre mois mais aussi lors de mes recherches concernant les sites d'art.

Ces cas sont:

- **La multiplicité des collaborateurs**

- **L'utilisation fréquente de citations**

- **La reproduction de contenu appartenant à un tiers**

- **La peur de perte d'indépendance.**

Certains propriétaires de sites, le plus souvent politiques ou religieux, se sont manifestés parce qu'ils étaient inquiets à l'idée de perdre leur indépendance éditoriale. Si avoir son site archivé et préservé par la BL peut être extrêmement gratifiant, cela implique aussi une responsabilité accrue en terme de ligne éditoriale que certains auteurs préfèrent éviter.

- **L'anonymat**

Ce point a été significatif pour ce qui concerne la collection de blogs, car un certain nombre d'auteurs de blogs souhaitent rester anonymes.
Au Royaume-Uni, les données personnelles collectées doivent être protégées, cependant, certains auteurs craignent la publicité possiblement générée par la participation à ce genre de projet et préfèrent l'éviter.

L'un des sites sélectionné, Petite Anglaise, est représentatif de ce qui peut se produire. L'auteur, aujourd'hui célèbre, a été renvoyée par son employeur parce qu'il considérait son blog comme néfaste pour son entreprise. Aujourd'hui, l'auteur de ce blog publie un livre et a attaqué son employeur en justice.
Cet exemple n'en est qu'un parmi tant d'autres et montre que les bloggers ne réalisent pas toujours l'impact que peut avoir le Web. Si certains prennent leurs précautions, d'autres pensent être anonymes parce qu'ils ne donnent pas leur nom. Internet donne encore l'illusion de l'anonymat parce que l'on écrit seul devant son ordinateur.

- **La modestie**

Certains auteurs sont préoccupés par l'idée que leur site n'est pas suffisamment intéressant pour faire partie d'une collection archivée par la BL.
Des auteurs de blogs m'ont par exemple contactés pour connaître les raisons de cette sélection. D'autres étaient flattés mais intrigués par l'idée qu'un journal intime puisse avoir un quelconque intérêt pour les chercheurs.
A ceux qui m'ont contactés, j'ai répondu qu'une collection devait être représentative de la production de blogs sur Internet mais aussi de la culture britannique. Il me semble que

les journaux intimes en ligne, comme les correspondances ou journaux intimes des siècles précédents, sont une représentation intéressante de cette culture.

- **Les universitaires**

Un dernier exemple est celui des chercheurs et universitaires. De nombreux blogs et sites Internet sont créés dans le cadre d'un travail de recherche. Il s'agit par exemple d'acquérir une notoriété dans son domaine, de pouvoir communiquer sur ses recherches ou d'en discuter avec des spécialistes.
La plupart des sites et blogs que j'ai sélectionnés m'ont donné leur accord très rapidement. Cependant, l'un d'entre eux a refusé. Il s'agissait d'un étudiant utilisant un blog, en accord avec son université, pour rédiger son mémoire. Son refus a été motivé par le fait qu'il s'agissait pour lui d'un projet à part entière et participer à un autre projet, tel que celui de la BL n'était pas souhaitable. Il me semble que cette utilisation de sites personnels ou de blogs dans le cadre de travaux de recherche non publiés pourrait amener un certain nombre de problèmes relatifs au droit d'auteur ainsi qu'à l'édition en général.

Tous ces exemples montrent que publier sur Internet n'est pas un acte anodin. Le Web possède ses célébrités et les questions de droit d'auteur sont d'autant plus importantes que la plupart de ressources sont mises en ligne et accessibles à tous gratuitement. Internet Archive en est un bon exemple. Ce projet à but non lucratif, situé aux États-Unis, a soulevé un certain nombre de polémiques puisque l'archive est constituée de sites qui n'ont pas accordés leur autorisation et ce dans des pays où la loi l'interdit.
Cela explique, en plus des raisons techniques évidentes, pourquoi la plupart des pays choisissent d'archiver uniquement les sites enregistrés dans leur nom de domaine, comme c'est le cas pour UKWAC.

3 Aspect technique

3.1 Processus d'archivage

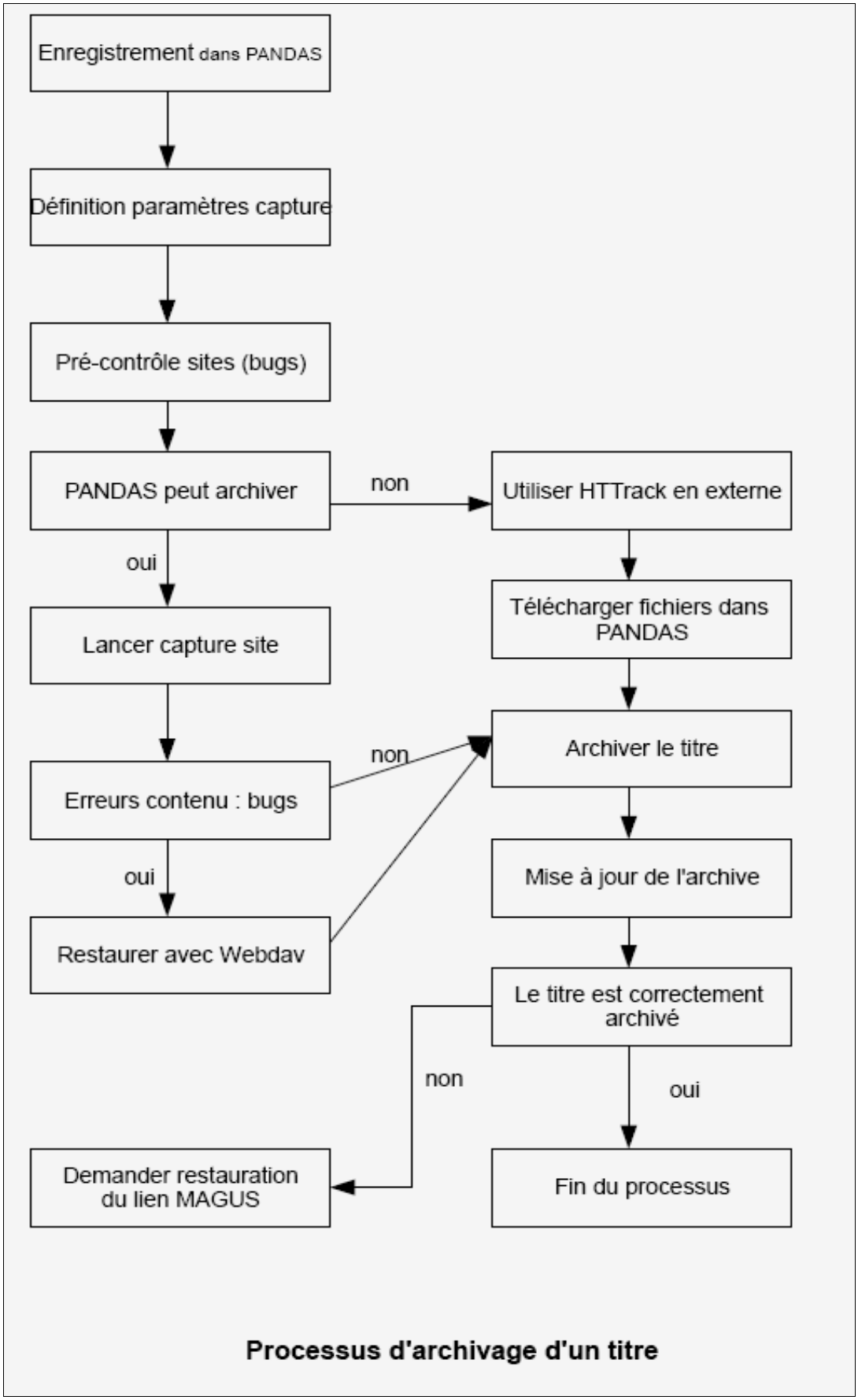
Lors de la sélection de sites Internet, les possibilités techniques des logiciels utilisés pour capturer les sites constituent un aspect à prendre en compte.

L'outil utilisé par UKWAC, PANDAS, fonctionne comme spécifié plus haut, avec HTTrack. Cet aspirateur de sites Internet est un logiciel libre qui possède de nombreux atouts, comme par exemple sa simplicité d'utilisation. Cependant, un certain nombre de caractéristiques techniques ne peuvent être capturées de manière optimale.

Avant de lister et d'expliquer ces limitations, il semble important de montrer les étapes nécessaires à l'archivage d'un site Internet.

Après les étapes de sélection et d'obtention de l'autorisation de copier, d'archiver et de diffuser le site par le biais du site Internet, les Web archivistes utilisent PANDAS.

Le circuit du « document » se décompose de la manière suivante:



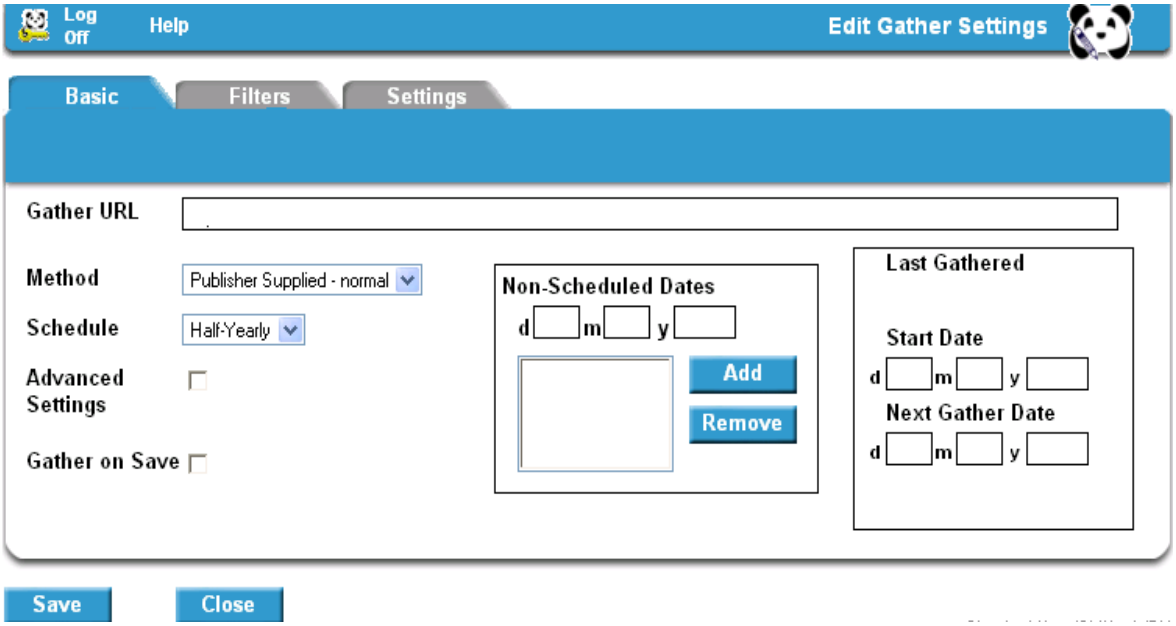
Processus d'archivage d'un titre

Ce schéma montre donc les étapes nécessaire à l'archivage d'un site à l'aide de PANDAS.

La première étape consiste donc à enregistrer les données relatives au titre dans PANDAS. Il faut entrer le nom du site, l'url, la date d'enregistrement, les coordonnées du propriétaire du site, les descripteurs, le format du site, le nom du Web archiviste en charge, etc, comme dans une base de données classique. On notera au passage que l'indexation est très basique.

- **PANDAS peut archiver**

Il faut contrôler le site pour détecter d'éventuels bugs lors de la capture. Si l'archiviste considère que PANDAS peut archiver le site, il le programme dans PANDAS.
Il lui faut programmer la capture du site dans le temps (dates et fréquences) ainsi qu'indiquer les urls à capturer et les filtres éventuels, permettant d'éviter les bugs, dans l'interface suivante:



Une fois tous les « ordres » entrés, il lance le processus de capture. Le site est donc dans la file d'attente et prêt à être archivé.
Le travail du Web archiviste est donc essentiel car tout « bug » prévisible et non repéré avant de lancer PANDAS est source non seulement de perte de temps mais aussi de ralentissement du processus technique d'archivage.
Je l'ai dit en première partie de ce document, le nombre de sites pris en charge par PANDAS est limité et la liste de sites programmés à l'archivage ou en attente de l'être est parfois très longue.
Lorsque PANDAS est surchargé, un signal se matérialise sur l'écran de contrôle sous la forme d'un feu rouge qui interdit tout archivage supplémentaire de site. Le but est donc de vérifier la taille des sites programmés ainsi que les bugs possibles afin d'éviter cette situation. Il faut aussi contrôler la taille et l'avancement des sites en cours d'archivage par les autres membres du Consortium et en tenir compte.
A la fin de ce processus de capture du site, l'archiviste récupère la copie du site capturé afin de la comparer au site original. Il effectue d'abord un contrôle visuel pour vérifier que le site est conforme puis utilise un programme Perl développé en interne afin de récupérer une liste des erreurs survenue durant la capture.

Le fichier se présente ainsi:

```
00:20:23        Warning:        Unable to test www.
00:20:31        Warning:        Unable to test www.
00:25:08        Error:  "Not Found" (404) at link www.
00:28:07        Error:  "Not Found" (404) at link www.
00:28:07        Warning:        Retry after error -2 (Receive Time Out)
00:28:07        Warning:        Retry after error -2 (Receive Time Out)
00:28:07        Warning:        Retry after error -2 (Receive Time Out)
```

Chaque erreur est répertoriée dans le Wiki de l'équipe et doit être traitée, le cas échéant avant la mise en ligne de la copie et son archivage. Je détaillerais quelque uns des « bugs » les plus courants dans la seconde partie. Cependant, j'ajouterai ici qu'il faut récupérer ou réparer chaque fichier manquant ou non conforme en dehors de PANDAS puis les réintégrer au fichier de la copie afin d'éliminer les erreurs. Heureusement,

certaines erreurs n'ont pas d'impact sur la qualité de la copie et ne nécessitent donc pas de correction.
Le travail est donc assez long puisqu'il faut constater les erreurs, trouver un moyen de les corriger, récupérer les fichiers en externe et enfin remplacer les fichiers défectueux.

- **PANDAS ne peut pas archiver**

Dans le cas ou les « bugs » détectés par l'archiviste ne peuvent pas être pris en charge par PANDAS, il faut utiliser HTTrack en externe. La capture des fichiers se fait alors en externe et les fichiers sont ensuite transférés dans PANDAS.

Dans les deux cas de figures, il faut de nouveau vérifier la copie du site puis lancer la phase finale d'archivage par PANDAS et mettre l'archive à jour.
Cette étape terminée, il faut contrôler sur le site public la conformité de la copie par rapport à l'original et le cas échéant, le bon fonctionnement des liens. Les liens externes ne sont bien entendu pas capturés lors du processus pour des raisons techniques et légales.
Si l'archiviste constate que l'un des liens ne fonctionne pas, il doit contacter le support technique afin que ce lien soit réparé dans les plus brefs délais. UKWAC a en effet un impératif de qualité à assurer pour ce projet et se doit de contrôler régulièrement l'état de l'archive.

La vérification du site doit donc se faire avant, pour déterminer quel logiciel utiliser, pendant, pour vérifier les erreurs éventuelles et après pour s'assurer que le lien du site Web est valide et que la copie du site est bien accessible et de bonne qualité.

3.2 Limitations techniques

Le contrôle qualité occupe donc une grande part du travail des Web archivistes.
Les erreurs les plus courantes sont répertoriées dans le Wiki de l'équipe et sont pour la plupart dues à:

- **Flash:**

Les fichiers Flash sont ceux qui posent le plus de problèmes parce que HTTrack ne parvient pas toujours à les capturer mais aussi parce que la plupart du temps, il est impossible de savoir si un site créé à l'aide de Flash sera capturé dans sa totalité ou non. Certains sites ne possédant qu'une bannière composée de Flash sont capturés alors que d'autres ne le sont pas.
Bien souvent, il est nécessaire pour les Web archivistes de relancer la capture de manière manuelle en espérant obtenir un meilleur résultat mais la plupart du temps, la consigne est d'éviter ces sites sauf s'ils sont exceptionnels.

- **Les pièges ou « crawler traps »**

HTTrack ne peut pas capturer les sites contenant des programmes qui bouclent comme les calendriers par exemple. Cela est dû au fait que ce logiciel ne peut pas capturer plus de 2GB et pas plus de 100,000 urls, au-delà, le système se bloque.
Or, les calendriers présents sur certains sites vont à l'infini et le logiciel se trouve pris au piège dans ce type de programme. Tout programme présent sur un site et fonctionnant à la manière de ces calendriers constitue une difficulté technique lors de l'archivage qu'il faut repérer avant la capture du site afin d'instaurer manuellement des filtres permettant d'éviter à l'aspirateur de tomber dans ces pièges. Par exemple, pour un calendrier, on notera dans la configuration préalable à l'archivage une limitation des dates qui se traduira par: « ne capture que les liens situés entre telle et telle date », ainsi l'aspirateur n'est pas pris dans la boucle et stoppe toute action une fois qu'il a atteint la date donnée.
Durant ma sélection de blogs, nous avons découvert que les commentaires postés sur les blogs par les lecteurs pouvaient fonctionner comme ces calendriers parce qu'ils sont tous reliés les uns aux autres. Il a donc fallu mettre en place un nouveau filtre pour ce type de site limitant l'accès au commentaires. Cette solution n'est évidemment pas idéale parce que cela revient à préserver un site Web pour de futurs chercheurs en l'amputant de ce qui constitue pour une grande part son intérêt, à savoir son caractère interactif.

- **Taille des sites**

Un autre paramètre à prendre en compte lors de la sélection d'un site est celui de sa taille. Cela est lié au point précédent, à savoir que HTTrack n'est pas capable de capturer

plus de 2 GB ou plus de 100 000 urls. Au-delà, les archivistes doivent relancer le logiciel utilisé en interne afin de capturer le site dans son entier.
Ceci est à prendre en compte lors de la sélection parce que les archivistes sont actuellement au nombre de deux, or toute intervention manuelle ralentit le travail de l'équipe. Dans l'idéal, l'archiviste ne devrait être là que pour vérifier la qualité de la copie capturée du site, ce qui lorsqu'il n'y a aucun problème, ne prend que quelques minutes.

- **Nombre d'url**

Les sites comportant différents liens dont la racine de l'url diffère sont également une source de problèmes parce que HTTrack ne peut prendre en compte que deux urls lors de sa configuration pour la capture. Ces sites comme ceux des exemples précédents demandent donc un temps de travail supplémentaire à l'archiviste.
Il lui faut tout d'abord choisir les deux urls les plus « rentables » c'est-à-dire couvrant la majorité du site. Si cela ne suffit pas, il faut relancer une capture des parties du site manquantes de manière manuelle (HTTrack interne à la BL) puis ajouter ces fichiers à la copie faite par le logiciel.

- **Les CSS**

Les CSS posent également problème au logiciel parce que ce dernier reformule, lors de la capture, le code de chaque page sur lesquelles elles apparaissent. Ce sont souvent les images (bannières, bandeaux,...) ainsi que l'organisation des pages qui sont ou manquantes ou déplacées. L'un des « bugs » les plus courant est celui-ci:

« /path/to/stylesheet.css » devient « path/to/stylesheet.css »

On constate que l'élément en rouge est manquant, le fichier ne peut donc être trouvé.
La procédure est là encore manuelle. Après capture du site, l'archiviste effectue un contrôle visuel du site puis le compare à l'original.
La BL a développé un programme Perl capable de lister les erreurs survenues lors de la capture. Les CSS sont listées dans ces erreurs, il faut donc corriger le code, c'est-à-dire indiquer le bon emplacement du fichier au robot. On relance ensuite la capture de chacun de ces fichiers que l'on intègre au bon emplacement afin que la copie soit conforme à l'original. Ceci est l'erreur la plus fréquente d'HTTrack et prend un temps parfois considérable pour ce qui est de la correction car il faut traiter les CSS une à une.

- **Les contenus dynamiques**

Les bases de données dynamiques comme les formulaires ou sondages en ligne ne peuvent évidemment pas être copiés par HTTrack à moins qu'ils ne soient téléchargeables. Il est donc impossible de préserver des contenus pourtant riches tels que les bases de données d'images.
Cette limitation a posé problème notamment pour les sites d'art utilisant des systèmes de base de données d'artistes ou d'oeuvres ou encore pour certains sites politiques dont l'une des nouvelles tendances est de créer un espace de discussion avec les électeurs potentiels contenant bien souvent toutes sortes de sondages d'opinion.

La liste de ces contraintes techniques incite donc à penser que la migration vers le nouvel outil WCT Tool et surtout l'utilisation d'Heritrix sont inévitables pour l'évolution du projet.

Partie 3 Analyse du projet

1 Les difficultés rencontrées

1.1 Les contraintes de temps

Il m'a été extrêmement difficile de planifier mes projets durant ce stage.
En effet, le temps de création estimé pour les collections du même type que les miennes, c'est-à-dire, hors actualité, était d'environ 6 mois. Je l'ai déjà expliqué, mes conditions de travail étaient cependant différentes car il avait été décidé au départ que je me concentrerai sur certains projets, les autres devant être occasionnels. Cependant, si j'ai consacré la plupart de mon temps de travail à la création de deux collections, j'ai également travaillé de manière active sur les sites d'art, testé les liens contenus dans la BDD afin de vérifier que l'accès aux sites Internet fonctionnait correctement, participé à des réunions, assisté à des conférences et démarrer un projet relatif à la publicité du projet par la BL.
Or, la plupart de ces tâches étaient impossibles à planifier au début de mon stage.

Un autre événement a marqué ce stage. A mon arrivée, Ravish Mistry, le « Permission Officer », venait d'être promu au poste de Web Archiviste. Le poste vacant de « Permission Officer » n'a trouvé de remplaçant qu'à la fin du mois d'août, c'est-à-dire après mon départ.
Dès le début de mon stage, nous avons convenu que je gérerai les collections dans leur ensemble, c'est-à-dire de la sélection à l'archivage. Cependant, l'absence de la personne chargée de la correspondance a fait que mon temps de travail a été pour une grande part, consacré à l'envoi de demandes de licence ainsi qu'à l'entrée des titres sélectionnés dans la BDD. Il devenait alors extrêmement difficile de quantifier le temps de travail nécessaire à la création de ces collections car les collections créées précédemment

l'avaient été par plusieurs personnes, dont une était chargée uniquement de gérer la BDD et la correspondance.
Au fil des mois, j'ai cependant remarqué que j'effectuais chacune de ces tâches de plus en plus rapidement. En effet, ces tâches m'ont permis non seulement de me familiariser avec les outils employés par l'équipe mais aussi de connaître mes collections et les sites sur lesquels je travaillais. Ceci a constitué un gain de temps non négligeable pour ce qui a concerné la préparation de réunions ou même pour ma présentation orale sur les blogs à la fin de mon stage.
Le fait de gérer les courriers et de créer moi-même les statistiques concernant le nombre de titres archivés par exemple, a fait que j'étais constamment en relation avec l'évolution de ces collections. Il me semble que mon implication dans le projet s'en est également trouvée décuplée parce que je me devais d'atteindre les livrables fixés au départ en terme de nombre de sites sélectionnés et prêts à être archivés. Suivre les statistiques et recevoir des emails et lettres montrant un intérêt des auteurs pour ce projet à bel et bien été une source de motivation supplémentaire.
Participer à chaque tâche a également facilité ma prise de contact avec les membres de l'équipe. Il m'a semblé que mon travail était utile et ma présence justifiée, ce qui a créé un sentiment presque immédiat d'appartenance à cette équipe.

Enfin la rédaction du rapport de stage, commencée à la mi-juillet, a été ralentie au début du mois d'août car la fin du stage comportait la rédaction d'un descriptif pour le site Internet de la collection de blogs ainsi que la préparation d'une présentation orale sur le sujet le 15 août. Pour des questions de gestion des connaissances, il m'a aussi fallu harmoniser mes documents afin de les ajouter à l'historique du projet contenu dans le Wiki.

1.2 Les contraintes politiques

Dans une institution aussi importante que la British Library, il faut prendre en compte des questions d'image et de relations.
C'est ainsi que j'ai vu certains des projets fixés au départ évoluer au cours de mon stage.

- **Les sites d'art**

La sélection, la classification ainsi que les questions de droit et les questions techniques relatives aux sites d'art ont constitué l'un des projets principaux de ce stage. Or, plusieurs problèmes se sont posés.

D'abord, il m'a semblé, ainsi qu'à mon superviseur, que l'ampleur de la tâche aurait pu suffire à remplir mes quatre mois de stage. Nous avons donc revu cette demande à la baisse. Il nous a semblé que le problème majeur de cette catégorie de l'archive était le peu d'autorisations reçues. Il a donc été décidé que ce projet serait une sorte de projet pilote destiné à trouver différentes stratégies afin d'améliorer le taux de réponses positives pour ces sites.

Ensuite, il avait été décidé que je collaborerai avec Chris Michaelides, chargé de créer la collection Art, afin d'améliorer ce taux. Cependant, il a été extrêmement difficile pour lui de se libérer car une importante exposition concernant l'art était en préparation au même moment à la BL.

Enfin, après avoir récupéré les données, contacté tous les sites et élaboré différentes stratégies d'approches, notamment en recherchant les anciennes correspondances dans les archives du projet, les résultats étaient peu significatifs, incitant à penser que ce projet nécessitait des appuis à un niveau hiérarchique élevé. Le travail concernant ces sites a donc été suspendu.

- **Les sites de religion Non Conformistes**

Cette tâche était définie dans la commande comme secondaire, après la création d'une collection de blogs et le travail sur les sites d'art. Cependant, elle est devenue un travail à part entière après une demande forte de la hiérarchie. Ce qui devait être une sélection de quelques sites, s'est transformé en la création d'une collection.

Cependant, le projet sur les sites d'art ayant du mal à démarrer au début de mon stage, je me suis impliquée avec enthousiasme dans ce nouveau projet. De plus, j'ai eu la chance de collaborer pour ce domaine que je connaissais peu avec des spécialistes et tout particulièrement avec le Dr. Clive D. Field.

Il m'a ainsi semblé que collecter des sites dans un domaine qui m'était inconnu et dans une langue qui n'est pas ma langue maternelle était une chance. Ce travail m'a appris bien plus que ce que j'en attendais.

- **Le 1000 ème site**

Ce projet publicitaire a pour but de faire connaître le programme Archivage du Web de la BL à son public. Lorsque j'ai commencé mon stage, il s'agissait de proposer différents sites à la hiérarchie de la BL, susceptibles d'être utilisés lors d'une campagne publicitaire.
En collaboration avec Alison Hill, j'ai proposé une liste de ces sites. En travaillant ensuite sur la conception de la collection de blogs, nous avons pensé qu'un blog serait une image très représentative de l'archivage du Web. J'ai donc proposé une courte liste de blogs pouvant être utilisés.
Ce projet soutenu au départ, n'a finalement pas abouti parce qu'il a été considéré comme inopportun à un moment où l'avenir du Consortium était en discussion.

Cet aperçu du fonctionnement d'une grande institution où les décisions se prennent à plusieurs niveaux a été très enrichissant pour moi.

1.3 La langue

La langue m'a posé beaucoup moins de problèmes que ce que j'avais imaginé.
Le travail de sélection était en effet effectué à mon rythme et j'ai utilisé toutes les ressources à ma disposition pour m'aider, comme par exemple les ressources en ligne de la BL, les dictionnaires, les encyclopédies ou les BDD spécialisées.
Lorsque j'avais un doute à propos d'un terme familier ou d'une allusion culturelle, je faisais appel à l'un des membres de l'équipe. Ma seule préoccupation était en effet liée à la sélection des sites et j'ai parfois hésité devant leur contenu.
Du point de vue culturel, politique ou religieux, il est parfois difficile d'effectuer une sélection lorsque l'on appartient à une autre culture. Ce problème n'en a pas vraiment été un car il m'a permis d'échanger à ce propos, en particulier avec Alison Hill, la conservatrice et ces conversations m'ont apportées énormément et m'ont permis de me familiariser avec la culture britannique.
Après un certain laps de temps, j'ai correspondu avec les propriétaires de sites qui avaient des réserves concernant le projet et j'avoue ne pas avoir éprouvé de grandes difficultés à le faire.

Au terme de ce stage, il me semble que ma compréhension de l'anglais et de la culture britannique, s'est beaucoup améliorée.

2 Bilan

2.1 Les solutions trouvées

La commande figurant dans le document de départ a dû être modifiée plusieurs fois au cours du stage. Ce qui aurait pu être considéré comme négatif m'a en réalité été utile et a fait de ce stage une réelle expérience professionnelle.
Hormis la création de la collection de blogs, tous les projets prévus dans la commande de départ ont dû être modifiés. J'ai dû m'adapter et trouver des solutions afin de rester aussi près que possible de la commande de départ. Mon implication dans l'équipe a rendu mon travail extrêmement riche car il était important pour moi, au même titre que chaque membre de l'équipe d'être utile au service et de faire en sorte de pallier à l'absence d'un des employés.
En plus de mes projets, j'ai donc participé à la vie du service et pris en charge un certain nombre de tâches ponctuelles.
J'ai vérifié les liens d'une liste de sites (environ 600) et contacté le support technique afin que les liens brisés du site d'UKWAC soient réparés.
J'ai contrôlé et le cas échéant corrigé la base de données ainsi que les informations entrées dans PANDAS pour chacun de ces sites.
J'ai participé aux réunions de l'équipe ainsi qu'au présentations d'Alison Hill devant les spécialistes chargés d'effectuer la sélection, afin d'exposer mon travail sur les collections de blogs et de sites religieux.
J'ai participé aux réunions concernant la refonte du site et aux tests d'utilisateurs.

Tout ceci m'a aidé à me sentir intégrée à l'équipe et à ressentir la même motivation pour la réussite du projet que tous ses membres.

Afin de répondre au mieux à la commande de départ, et malgré les contraintes et le manque de temps, j'ai tenté de regrouper certaines demandes.
Par exemple, le projet de départ qui consistait à sélectionner des sites d'Irlande du Nord a été intégré à mes projets de collection. J'ai procédé à une recherche spécifique de blogs d'Irlande du Nord grâce à des moteurs de recherche et annuaires spécialisés comme par exemple, Britblog.com[31]. J'ai fait de même pour les sites de religion Non Conformiste. J'ai indiqué dans la base de données la localisation géographique de ces sites afin qu'ils puissent être regroupés par la suite si cela était le souhait de l'équipe, en une collection sur l'Irlande du Nord. Cette indication permettra aussi, de pouvoir simplement spécifier la région sur le site d'UKWAC, comme c'est le cas aujourd'hui pour les sites d'Écosse ou du Pays de Galles.

Pour augmenter le nombre de sites composant la collection de blogs et remplir mes objectifs, j'ai utilisé différentes stratégies.
Je me suis d'abord intéressée aux sites protégés par une Licence Creative Commons en utilisant le site des Creative Commons ainsi que des moteurs de recherche spécialisés, comme par exemple Yahoo.
Ce mode de recherche n'a pas fait augmenter le nombre de sites de manière spectaculaire mais m'a tout de même aidé à augmenter le nombre de sites autorisés à être archivés.
J'ai également utilisé plusieurs fois la sélection de blogs liés à de grands groupes comme le Guardian Unlimited ou le Times Online par exemple. Cela a non seulement remarquablement fait augmenter le nombre de sites sélectionnés mais encore celui de sites autorisés. Le Times Online nous a par exemple accordé une licence pour la totalité de ses blogs, soit plus d'une dizaine de licences en une seule fois.
De même, j'ai tenté de grouper mes sélections de sites religieux. Un mouvement religieux à l'origine d'associations diverses constitue me semble-t-il une grande chance de récupérer plusieurs sites en ne demandant qu'une licence.
J'ai misé, chaque fois que les événements le permettaient, sur mes contacts, lors de conférences par exemple ou par courriers afin de trouver de nouveaux propriétaires de sites. J'ai notamment expliqué à chaque propriétaire de site religieux à quel point il m'était difficile de récupérer des licences pour ce type de sites en leur demandant de parler du projet autour d'eux et de me recommander tout site auquel ils penseraient.

31 Section Northern Ireland

Cette dernière stratégie alliée à l'aide du Dr. Clive D. Field m'a paru commencer à porter ses fruits à la fin de mon stage.

Il m'a semblé que la communication à l'intérieur ainsi qu'à l'extérieur de la BL était un élément primordial de la réussite de ces projets de collections.
Alison Hill m'a énormément soutenue pour ce qui est de cet aspect de mon travail, en me présentant systématiquement toute personne susceptible d'être intéressée et en me permettant d'intervenir durant ses présentations et de participer à des conférences.
Enfin, la présentation orale que j'ai effectuée à la fin de mon stage, et qui portait sur la collection de blogs, m'a valu quelques retours au sein de la BL et notamment, deux personnes possédant un blog se sont manifestées afin qu'il soit archivé.

2.2 Les collections créées

J'ai donc créé deux collections lors de ce stage, l'une regroupant des blogs et l'autre des sites de religion Non Conformiste. Les statistiques de la base de données montrent qu'entre le 23/04/07 et le 17/08/07, environ 600 sites ont été contactés pour demander la licence nécessaire à l'archivage et à la diffusion des sites. Sur ces 600 sites 230 ont accordés cette autorisation dans la même période. Une grande part de ces chiffres est en rapport avec ces deux collections.

- **Les sites Non Conformistes**

Au terme de ce stage, la collection de site Non Conformistes comprend 159 sites se répartissant de la manière suivante:

- ➔ Méthodistes, environ 18 %
- ➔ Baptistes, environ 14%
- ➔ Presbytériens, environ 11%
- ➔ Quakers, environ 11%
- ➔ Autres Mouvements (Salvation Army, Vineyard, Pentecostal, etc.), environ 46 %

Ici une première constatation peut être effectuée. Lors de mes recherches concernant la pratique des religions au Royaume-Uni, je me suis intéressée à la répartition des pratiquants chrétiens Non Conformistes. Les chiffres trouvés dans les différents ouvrages de référence que j'ai pu consulter donnaient des pourcentages très proches les uns des autres, répartissant les pratiquants ainsi:[32]

- ➔ Presbytériens, environ 17 %
- ➔ Méthodistes, environ 7 %
- ➔ Pentecostalistes, environ 4 %
- ➔ Baptistes, environ 3,5 %
- ➔ Autres Mouvements, environ 5 %

Ainsi, je n'ai pu respecter de manière précise la répartition de ces pratiques comme je l'avais d'abord décidé. Ma motivation de départ était de pouvoir justifier les choix de cette collection et d'éviter ainsi toute réclamation de tel ou tel groupe religieux.
Cependant, après avoir commencé cette sélection, il m'a semblé que si je devais essayer de respecter cette répartition, je devais également tenir compte de la présence de ces même mouvements sur Internet. Or, force est de constater que certains mouvements sont plus présents sur la toile que d'autres et aussi plus innovants.
Ainsi, j'ai décidé de favoriser la diversité et la qualité de la collection tout en collectant une majorité de sites représentatifs des mouvements les plus importants. En plus des critères de qualité et de diversité, il m'a paru important de revoir mes objectifs de départ tout simplement parce qu'en élaborant ces collections, il m'a semblé que mon travail était davantage de représenter la population présente sur Internet.
Par exemple, les mouvements presbytériens sont plus représentés dans le pays qu'ils ne le sont sur Internet. J'ai toutefois tenté d'en collecter un maximum, en diversifiant ma sélection, de l'église au collège ou au site historique. Cependant, j'ai trouvé bien plus d'exemples de sites tels que les organismes de charité, les associations culturelles ou les églises virtuelles pour les mouvements méthodistes, baptistes et quakers.

La sélection regroupe dans l'ordre:

- ➔ Mouvements religieux: Églises / Organisations

32 Christian Research Association's Religious Trend

- ➔ Associations: Caritatives, historiques, éducatives
- ➔ Magazines en ligne, journaux
- ➔ Bibliothèques, églises virtuelles, blogs, etc.

Malheureusement, le taux de réponses positives pour cette collection est extrêmement bas. Cependant, il m'a semblé que les changements de tactique consistant en l'ajout du nom du Dr. Clive Field ainsi que mon dialogue systématique avec les responsables me contactant, commençaient à porter leurs fruits au moment où j'ai quitté la BL.

Les résultats concernant cette collection se décomposent ainsi au 17 août:

24 sites au total ont donné leur accord, dont:

- ➔ Quakers, 11 sites
- ➔ Baptistes, 6 sites
- ➔ Méthodistes, 4 sites
- ➔ Autres, 3 sites

Il me semble que ces résultats reflètent davantage la présence et les efforts de communication, comme la création de nouvelles communautés religieuses en ligne, par certains mouvements. En effet, les Quakers sont très représentés sur la toile et possèdent énormément de sites de type communautaires tels que des « groupblogs », c'est-à-dire des blogs publiés en groupe ou des forums. Les sites baptistes et méthodistes sont à peu près autant représentés les uns que les autres sur Internet et tentent d'innover sur le plan technique. Les deux églises virtuelles sélectionnées pour la collection sont d'ailleurs méthodistes.[33]

Cette collection parvient ainsi à montrer le côté traditionnel de ces mouvements religieux tout en mettant l'accent sur l'évolution des religions au Royaume-Uni. Les modes de communication et de diffusion de l'information sont indéniablement en pleine mutation. Il existe aujourd'hui des associations proposant aux églises locales de leur créer un site Internet. Les sermons sont transmis à l'aide de podcasts (ou fichiers audio accessibles sur le site) et il est possible de visionner la messe du dimanche en ligne.

33 Ex.: Church of Fools

- **Les blogs**

En comptant les blogs déjà dans la BDD à mon arrivée, 364 blogs ont été sélectionnés pour cette collection. Au 17 août, nous avions une autorisation d'archiver pour 154 de ces blogs, en comptant les sites sous Creative Commons.

J'ai organisé les titres au fur et à mesure que la collection grandissait dans une base de données, en séparant les descripteurs de la base de données de la BL en 3 sujets: Sujet 1, Sujet 2, Sujet 3. J'ai ensuite comptabilisé les différents sujets par catégorie en reprenant la catégorisation de la BL. J'ai ajouté à cette catégorisation une catégorie « collections » qui regroupe les blogs appartenant à d'autres collections créées pour UKWAC, par exemple, « London Terrorist Attack 7th July 2005 ».

- ➔ Society & Culture: environ 18 %
- ➔ Reference Works (contient les journaux intimes): environ 15 %
- ➔ Art & Humanities: environ 14 %
- ➔ Government & Politics: environ 13,5 %
- ➔ Science & Technology: environ 10 %
- ➔ News & Media: environ 8 %
- ➔ Business & Economy: environ 8,5 %
- ➔ Collections: environ 7 %
- ➔ Education & Research: environ 4 %
- ➔ Health: environ 2 %

Les pourcentages obtenus montrent que la catégorie Society & Culture est davantage représentée que les autres. Durant mon stage, j'ai essayé au maximum d'éviter cette catégorie qui semble être utilisée en règle générale parce qu'elle est peu dénominative. Si elle arrive tout de même première au classement, c'est parce que deux des sous catégories qui la composent sont très représentées dans les blogs de cette collection, à savoir le sport et les festivals d'une part et les communautés, d'autre part. Le sport et les festivals parce que je trouvais important de représenter ces aspects de la vie quotidienne des britanniques. Les communautés parce que l'un des aspects l'un plus intéressants en ce qui concerne les blogs est me semble-t-il le phénomène d'interaction.

Le développement des communautés virtuelles est un phénomène qui va de pair avec le succès des blogs et j'ai bien souvent constaté durant mes recherches que de nouvelles communautés naissaient de cette manière.

Le fonctionnement des blogs est très intéressant du point de vue sociologique puisqu'il démontre le besoin qu'ont les auteurs de se regrouper par centre d'intérêt, lieu d'habitation, etc., dans leur vie virtuelle, comme ils le feraient dans la vie réelle. Les « blogrolls » ainsi que tous les systèmes permettant la citation ou le renvoi à un autre blog à partir du sien propre, (par exemple en utilisant un outil tel que Trackback [34]), sont de bons exemples des liens qui se tissent entre « bloggers ». De nombreux « Groupblogs » sont d'ailleurs issus de ce type de relations.

Dans le domaine de la recherche universitaire, auquel je me suis tout particulièrement intéressée, les blogs sont utilisés comme un outil de diffusion de l'information, voire même comme un outil aidant à la réflexion et permettant de tisser des liens avec la communauté de son domaine d'étude. Il existe de nombreux sites personnels de professeurs d'université, qui disent utiliser ce média parce qu'il permet une communication moins formelle et plus directe avec les étudiants. J'ai également trouvé un certain nombre de blogs utilisés par des étudiants, le plus souvent en thèse, souhaitant échanger sur leur sujet de recherche, partager leurs conclusions et s'aider de la discipline liée à l'écriture journalière d'un blog pour s'astreindre à la rédaction de leur thèse.

Il existe également un certain nombre de « Social Networks » plus organisés tels que Nature Network Boston et London[35]. Cet exemple, qui est une initiative de l'éditeur Nature Publishing Group (NPG)[36], fonctionne à la manière de n'importe quel autre réseau Internet (MySpace, Facebook, etc.). Il permet de bénéficier d'un espace propre au sein d'une communauté de scientifiques. Il s'agit de partager mais aussi de trouver un emploi ou de créer des relations, c'est-à-dire de gagner en notoriété.

Il est intéressant de remarquer que le blog est souvent décrié parce qu'assimilé au journal intime d'adolescent. Le porte drapeau de cette opinion au Royaume-Uni est Andrew Keen, auteur de l'ouvrage, *The Cult of the Amateur*. Il y critique les blogs qu'il juge trop nombrilistes et la qualité moindre de la culture et des idées véhiculées sur Internet qui auraient une influence néfaste sur la société.

34 Ou retroliens, système de liens inter-blogs semi-automatisé. (Wikipédia)
35 Nature Network Londres
36 http://www.nature.com/npg/

Andrew Keen écrit de nombreux articles sur le sujet et participe à de nombreuses conférences. Il possède lui-même un blog[37] qui lui permet de diffuser son opinion sur le sujet et de vendre ses livres. La proportion de journaux intimes sélectionnés pour cette collection est en effet importante car il me semble que ce mode d'expression est représentatif d'une partie de la population et de l'utilisation des blogs, elle se devait donc d'être représentée. Cependant, j'ai également collecté ces blogs parce que tous témoignent de la vie quotidienne dans diverses régions et milieux du Royaume-Uni. Ce type de blogs tant décriés avaient pour moi l'intérêt d'être l'expression de la culture britannique, or préserver cette culture est l'un des impératifs de la politique de développement de la collection Web.

J'ai par ailleurs collecté de nombreux sites ayant trait aux Sciences Humaines et Sociales ainsi qu'aux nouvelles technologies. Si les sites créés par des amateurs existent bel et bien sur Internet, il m'a semblé qu'une majorité de blogs politiques, artistiques ou technologiques étaient au contraire l'oeuvre de professionnels.

De nombreux journalistes spécialisés en un domaine possèdent un blog afin de pouvoir trouver une liberté d'expression et de choix de sujet. Des artistes ou des entreprises utilisent le blog pour exposer leur travail. Les uns se font ainsi connaître ou vendent leurs oeuvres. Les autres augmentent leur chiffre d'affaire en élargissant leur clientèle.

L'édition est également touchée par ce phénomène alors que des universitaires ou des auteurs utilisent leur blog pour publier sans contraintes ou qu'à l'inverse une maison d'édition leur propose un contrat comme cela a été le cas pour certains blogs.

S'il est indéniable qu'il existe une communauté de bloggers, possédant leur vocabulaire, leurs fêtes, leurs récompenses et leurs codes, parler d'une culture Internet qui abêtirait la population me semble très extrême. Il m'a semblé après avoir visité un grand nombre de blogs, que la majorité d'entre eux étaient créatifs, pédagogiques et orientés vers un but. L'un des mouvement actuel dans cette communauté est d'ailleurs de pousser à l'action les citoyens au niveau politique, environnemental ou associatif. De nombreux blogs utilisent aujourd'hui des bannières d'organisations humanitaires ou de collectifs appelant à signer telle ou telle pétition ou à manifester, incitant leurs lecteurs à s'engager de manière active.

37 The great seduction

Finalement, l'écart entre la « vie réelle » et la « vie virtuelle » m'a semblé bien plus ténu que ce qui peut se dire ou s'écrire à ce sujet, c'est pourquoi archiver le Web a pour moi une visée réellement patrimoniale.

Le dernier travail concernant cette collection a consisté en la rédaction d'un descriptif pour le site UKWAC (ci-dessous). Le travail effectué pour créer cette collection a également fait l'objet d'une présentation PowerPoint[38] auprès d'un certain nombre de bibliothécaires et conservateurs de la BL et en présence de l'équipe WAP.
Cette présentation orale a été une expérience extrêmement enrichissante car j'ai pu partager le travail de recherche effectué durant ces 4 mois ainsi que les solutions trouvées pour sélectionner et organiser ces sites Internet.
Les questions qui m'ont été posées lors de cette présentation m'ont été très utiles et les retours positifs ainsi que l'intérêt porté par ce public de professionnels m'ont beaucoup touchés.

Blogs

The UK Blogosphere (multiple connected communities of web logs) has burgeoned since the late 1990s and early 2000s, due in part to a more convenient publishing process. Writing on the Web is open to a more diverse, less technical population than at any time before. Individuals are creating and posting a wide range of material from academic research, political commentaries, news items etc to personal thoughts and insights. This collection represents a cross section of UK web logs containing a wealth of material which will be of value to researchers now and in the future.

Top ^

38 Présentation Blogs

Conclusion

La British Library, en tant qu'institution nationale et membre leader du UK Web Archiving Consortium, occupe un rôle clé dans ce projet d'archivage et de préservation du domaine national.

L'indépendance dont j'ai bénéficié, ainsi que la volonté d'échange d'idées et d'opinions voulue par la conservatrice, Alison Hill, soutenue dans cette approche par les autres membres dirigeants de la British Library, m'ont énormément apporté en terme d'enrichissement personnel et professionnel.

J'ai pu planifier mon travail afin de répondre aux objectifs chiffrés par la hiérarchie et créer des collections dignes d'intérêts et représentatives d'une culture qui n'était pas la mienne. J'ai quantifier et organiser les sites sélectionnés afin de pouvoir justifier chacun de mes choix. Enfin, j'ai combiné des critères de sélection capable d'identifier la qualité d'un site et sa cohérence par rapport à l'ensemble de la collection future.

Au delà des tâches qui m'ont été confiées, ce stage effectué dans le service Web Archiving de la British Library m'a laissé entrevoir les enjeux inhérents à la préservation de notre patrimoine et m'a permis d'y contribuer.

J'y ai aussi appréhendé la nécessaire évolution du rôle de professionnel de l'Information et les défis à relever afin d'y parvenir.

Bibliographie

ARTICLES

Andrieu, Olivier. «L'archivage du Web: utopie ou projet stratégique?». *Technologies Internationales,* [En ligne]. no 115 (juin 200(). http://www.adit.fr/SP/pdf/ti/115.pdf (Page consultée le 30 juin 2007)

Bachimont Bruno; Drugeon Thomas; Piéjut Geneviève. *«Documenter et partitioner une archive du Web: vers le dépôt légal d'un domaine média*». In Archives & Museum Informatics Publication. Site d'Archives & Museum Informatics, [En ligne]. ,http://www.archimuse.com/publishing/ichim05/Bachimont.pdf (Page consultée le 02 aout 2007)

Charlesworth, Andrew . «Legal issues to the archiving of Internet resources in the UK, EU, US, and Australia». In Web Archiving. *Site* de la Wellcome Library, [En ligne] http://library.wellcome.ac.uk/assets/wtl039230.pdf (Page consultée le 26 avril 2007)

Day, Michael. «Collecting and preserving the world wide web: a feasability study undertaken for JISC and The Wellcome Trust». In Web Archiving. *Site* de la Wellcome Library, [En ligne] http://library.wellcome.ac.uk/assets/wtl039229.pdf (Page consultée le 26 avril 2007)

Day, Michael. «The long-term preservation of Web content». In Index of preservation publications 2006. *Site de UKOLN, University of Bath*, [En ligne]. http://www.ukoln.ac.uk/preservation/publications/2006/web-archiving/md-final-draft.pdf (Page consultée le 26 juillet 2007)

Electronic Publishing Services Ltd. «*Refining the map of the universe of electronic publications potentially eligible for legal deposit*». In EPS Report to LDAP. Site du Department for Cuture, media and sport, [En ligne]. http://www.culture.gov.uk/NR/rdonlyres/02E46A4E-F53F-4071-B80E-C98DCE92761C/0/EPS_Report_to_LDAP_Nov_2006.pdf

Game Valérie; Illien Gildas. «Le dépôt légal d'Internet à la Bibliothèque Nationale de France». *Bulletin des Bibliothèques de France (BBF),* [En ligne]. no 3 (2006). http://bbf.enssib.fr/sdx/BBF/frontoffice/2006/03/document.xsp?id=bbf-2006-03-0082-013/2006/03/fam-dossier/dossier&statutMaitre=non&statutFils=non (Page consultée le 19 avril 2007)

Hill, Alison; Price Richard. «*Collection development policy for UK Websites».* In Modern British Collections. Site de The British Lirary, [En ligne]. http://www.bl.uk/collections/britirish/pdf/modbritcdpwebsites.pdf (Page consultée le 23 avril 2007)

Masanes, Julien. «Préserver les contenus du Web». In bibnum.bnf.fr. Site de la Bibliothèque Nationale de France (BNF), [En ligne].http://bibnum.bnf.fr/conservation/migration_web.pdf (Page consultée le 16 juillet 2007)

Patel, Majula; Pennock Maureen. «Digital Preservation Coalition Forum on Web Archiving». *ADRIADNE* [En ligne]. no 48 (juillet 2006). http://www.ariadne.ac.uk/issue48/dpc-web-archiving-rpt/ (Page consultée le 21 juin 2007)

Sugden, Joanna. «30 Most influential religion blogs». In Faith Central. *Site du Times Online* [En ligne] http://timesonline.typepad.com/faith/2007/07/articles-of-fai.html (Page consultée le 16 juillet 2007)

Thompson, Dave; Bailey, Steve. «UKWAC: Building the UK's First Public Web Archive». *D-Lib Magazine* [En ligne]. Vol.12 no 1 (janvier 2006). http://www.dlib.org/dlib/january06/thompson/01thompson.html (Page consultée le 11 juin 2007)

MONOGRAPHIES

Keen, Andrew. *The Cult of the Amateur: How Today's Internet is Killing Our Culture And Assaulting Our Economy,* Paperbook, 2007, London

Fry, Eileen; Weller, Paul; Michele, Wolfe. *Religions in the UK, Directory 2001-2003,* University of Derby in association with the Inter Faith Network for the UK, 2001

Masanés, Julien. *Web Archiving,* Springer Berlin Heidelberg , 2006

MCKeating, Justin. *The Blog Digest 2007, Twelve months of the best writing from the Web,* Friday Book, 2006, London

Religious Trend, Christian Research UK, No. 2, 2001/2002, Millenium Edition

UK Christian Handbook 2007/2008, Christian Research UK, 2006

Worstall, Tim. *2005 Blogged, Dispatches from the blogosphere,* Friday Book, 2005, London

SITES INTERNET

7th International Web Archiving Workshop: *Site de The International Web Archiving Workshop (IWAW),* [En ligne]: http://www.iwaw.net/07/index.html (Page consultée le 24 juin 2007)

Apache Lucene: *Site de The Apache Software Foundation,* [En ligne]: http://lucene.apache.org/java/docs/ (Page consultée le 10 juin 2007)

Blog Directory: *Site de Bloscholar,* [En ligne]: http://www.blogscholar.com (Page consultée le 12 mai 2007)

Blog Edublogs: *Site de Edublogs*: [En ligne]: http://edublogs.org/category/blog/ (Page consultée le 11 juillet 2007)

Blog Terms Glossary: *Site de Whatis:* [En ligne]: http://whatis.techtarget.com/definition/0,,sid9_gci1186975,00.html (Page consultée le 20 juillet 2007)

Site de The Blogging Libraries, [En ligne]: http://www.blogwithoutalibrary.net/links/index.php?title=Welcome_to_the_Blogging_Libraries_Wiki (Page consultée le 12 mai 2007)

Cataloguing Cultural Objects: *Site de Visual Ressources Association (VRA)*, [En ligne]: http://www.vraweb.org/ccoweb/cco/selections.html (Page consultée le 05 mai 2007)

Cedars project (Curl Exemplars in Digital Archives): *Site de University of Leeds*, [En ligne]: http://www.leeds.ac.uk/cedars/ (Page consultée le 12 mai 2007)

Copyright: *Site de the UK Intellectual Property Office,* [En ligne]: http://www.ipo.gov.uk/copy.htm (Page consultée le 28 juillet 2007)

Digital Preservation: *Site de UKOLN*, [En ligne]: http://www.ukoln.ac.uk/preservation/publications/ (Page consultée le 30 juin 2007)

Digital Preservation: *Site de University of London Computer Center (ULCC),* [En ligne]: http://www.ulcc.ac.uk/digital-preservation/current-activities/ukwac.html (Page consultée le 15 juillet 2007)

Site de The Digital Preservation Coalition, [En ligne]: http://www.dpconline.org/graphics/index.html (Page consultée le 15 juillet 2007)

Site de the European Archive, [En ligne]: http://www.europarchive.org/ (Page consultée le 28 avril 2007)

Site de The Joint Joint Information Systems Committee (JISC), [En ligne]: http://www.jisc.ac.uk/ (Page consultée le 12 avril 2007)

Legal deposit: *Site de the Departement for Cuture Media and Sport (DCMS),* [En ligne]: http://www.culture.gov.uk/what_we_do/Libraries/legal_deposit/ (Page consultée le 28 juillet 2007)

Library and Information studies: *Site de Intute,* [En ligne]: http://www.intute.ac.uk/artsandhumanities/cgi-bin/browse.pl?id=artifact859 : (Page consultée le 10 mai 2007)

MINERVA: *Site de The Library of Congress*, Web Archiving and Preservation project, [En ligne]: http://lcweb2.loc.gov/cocoon/minerva/html/minerva-home.html (Page consultée le 20 avril 2007)

netpreserve.org: *Site de the International Internet Preservation Consortium (IIPC),* [En ligne]: http://www.netpreserve.org/ (Page consultée le 21juillet 2007)

PANDORA Archive: *Site de The National Library of Australia,* [En ligne]: http://pandora.nla.gov.au/apps/PandasDelivery/WebObjects/PandasDelivery.woa/ (Page consultée le 12 avril 2007)

Projects: *Site de The Wellcome Library,* [En ligne]: http://library.wellcome.ac.uk/projects.html (Page consultée le 12 avril 2007)

Resources for Research: *Site de la British Library,* [En ligne]: http://www.bl.uk/collections/resres.html (Page consultée le 04 mai 2007)

Staff Papers 2007: *Site de The National Library of Australia,* [En ligne]: http://www.nla.gov.au/nla/staffpaper/2007/index.html (Page consultée le 15 juillet 2007)

UKWAC: *Site de MAGUS,* [En ligne]: http://www.magus.co.uk/aboutus/casestudy_ukwac.html (Page consultée le 26 mai 2007)

Web Archiving, PADI (Preserving Access to Digital Information): *Site de The National Library of Australia*, [En ligne]: http://www.nla.gov.au/padi/topics/92.html (Page consultée le 28 avril 2007)

Web Curator Tool: *Site de SourceForge.net,* [En ligne]: http://webcurator.sourceforge.net/ (Page consultée le 08 juillet 2007)

Women Business Blogging conference: *Site de De Monfort University* Leicester, [En ligne]: http://www.hum.dmu.ac.uk/blogs/nlabwomen/ (Page consultée le 6juillet 2007)

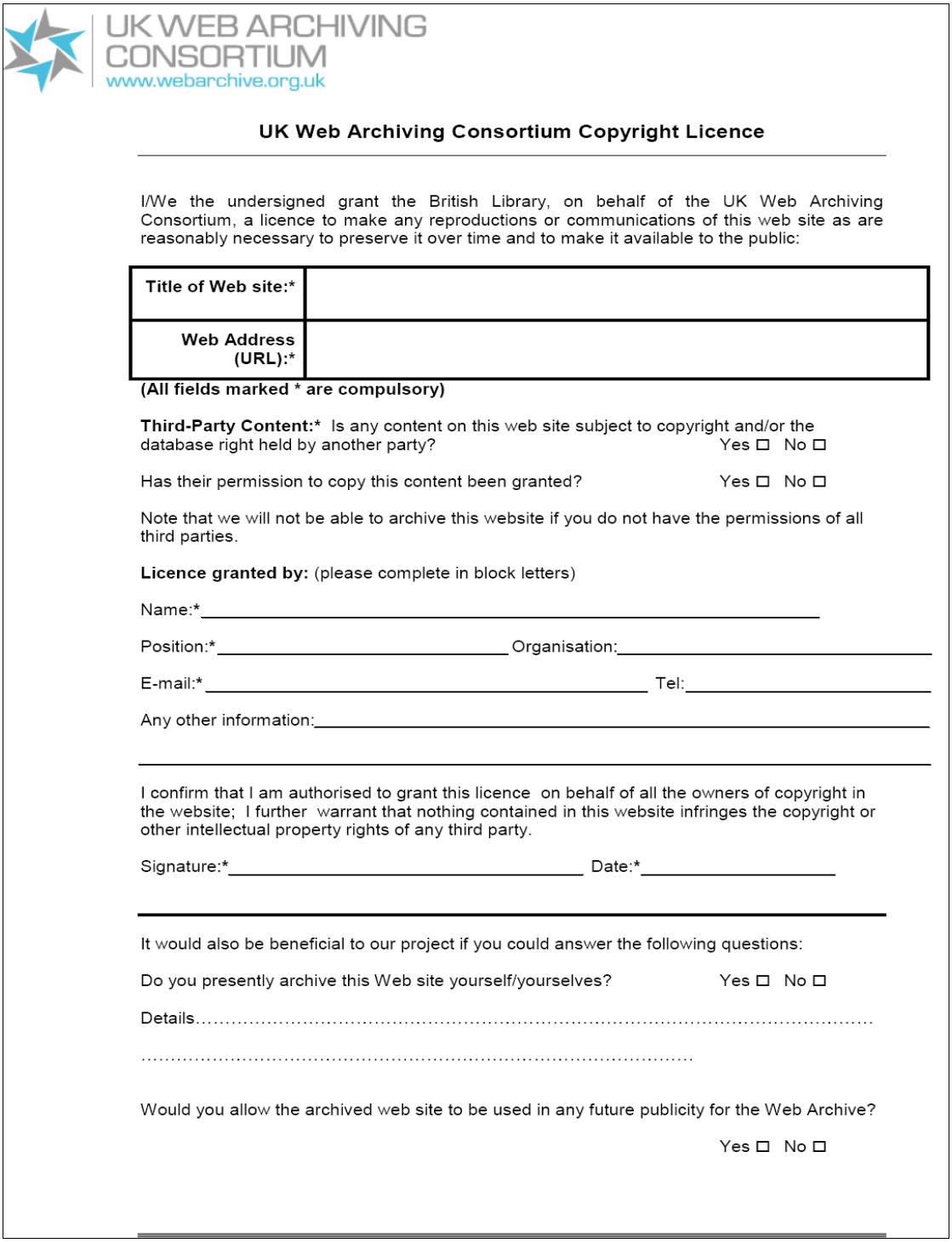
Table des annexes

Annexe 1 Fiche sélection

Selection Form for websites to be archived. Please forward by email to ..... or internal post to ....., Web Archivist, Modern British Collections, Floor 2, Zone 8, St. Pancras.	
Name (Subject Specialist)	
Date	
Name of Website	
URL (address of website)	
Contact (name/position/postal address/telephone)	
Frequency of archiving *	
Subject area (choose between one and three from the subject list)	
Notes. Please include a brief description of the website, including justification for archiving	

* Please select one of the following: **One-off, daily, weekly, fortnightly, monthly, quarterly, half-yearly, 9-monthly, annual, 18-monthly, biennial**

Annexe 2 Licence



UK WEB ARCHIVING CONSORTIUM
www.webarchive.org.uk

UK Web Archiving Consortium Copyright Licence

I/We the undersigned grant the British Library, on behalf of the UK Web Archiving Consortium, a licence to make any reproductions or communications of this web site as are reasonably necessary to preserve it over time and to make it available to the public:

Title of Web site:*	
Web Address (URL):*	

(All fields marked * are compulsory)

Third-Party Content:* Is any content on this web site subject to copyright and/or the database right held by another party? Yes ☐ No ☐

Has their permission to copy this content been granted? Yes ☐ No ☐

Note that we will not be able to archive this website if you do not have the permissions of all third parties.

Licence granted by: (please complete in block letters)

Name:*_____

Position:*_____ Organisation:_____

E-mail:*_____ Tel:_____

Any other information:_____

I confirm that I am authorised to grant this licence on behalf of all the owners of copyright in the website; I further warrant that nothing contained in this website infringes the copyright or other intellectual property rights of any third party.

Signature:*_____ Date:*_____

It would also be beneficial to our project if you could answer the following questions:

Do you presently archive this Web site yourself/yourselves? Yes ☐ No ☐

Details……

……

Would you allow the archived web site to be used in any future publicity for the Web Archive?

Yes ☐ No ☐

Annexe 3 Lettres Type

Invitation to participate in Web preservation programme

Dear Sir/Madam

Click and enter site address

"The Religious Archives Group, an affiliated organisation of the Society of Archivists, attaches great importance to the long-term archiving of a significant number of the country's religion-related websites, which increasingly offer unique sources of information for historians of the future. The Group therefore strongly endorses the efforts of the United Kingdom Web Archiving Consortium in including religion-related sites in its programme. On behalf of the Group, I very much hope that you will feel able to consider sympathetically the Consortium's request to archive your own site. This has been carefully selected for inclusion in the collection of religious websites, and your permission will enable your website to be preserved for posterity.

Dr Clive D Field, OBE, President, Religious Archives Group"

The British Library is a founding member of the UK Web Archiving Consortium (www.webarchive.org.uk) consisting of The British Library, JISC (Joint Information Systems Committee), the National Archives, the National Library of Scotland, the National Library of Wales and the Wellcome Library. The Consortium is the national effort to archive selective representative websites from UK web space in advance of the introduction of legal deposit for digital materials.

The British Library would like to invite you to participate in this work by allowing us to archive your web site under the terms of the appended licence. We select sites to represent aspects of UK documentary heritage and as a result, they will remain available to researchers in the future. We aim to subsequently include the archived copy of your web site in our permanent collections.

Invitation to have your website archived in a collection of blogs

Dear Sir / Madam,

Click and enter site address

The British Library is building a collection of blogs. This collection will form part of the UK Web Archiving Consortium (UKWAC) initiative to archive websites of research interest. Please visit www.webarchive.org.uk if you wish to see the current online archive which is publicly accessible.

We would like to invite you to have your site included in this important collection for Internet research. We will be selecting some 150 key sites to form the basis of the blog's collection until August 2007 but archiving will continue into the future. To carry out this archiving we need you to sign the Licence document.

If you are happy for your site to be included in this Web archive please complete the attached copyright licence form and return it to the address given below. If there are any other of your sites which you would like to be considered for archiving, and you are able to sign a licence document for them please make additional copies of the licence document. For more information about Copyright, the UK Web Archiving Consortium and how your archived web site will be made available please see the attached Further Information & FAQ document.

Alternatively, if you require any additional information, please do not hesitate to contact me.

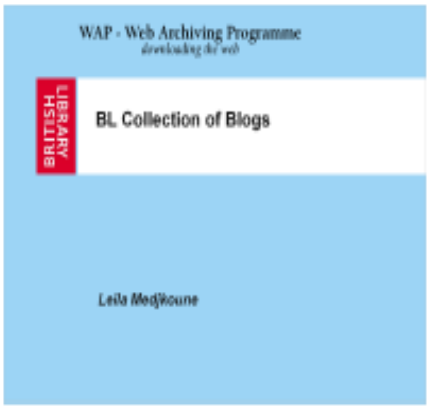
Best regards,

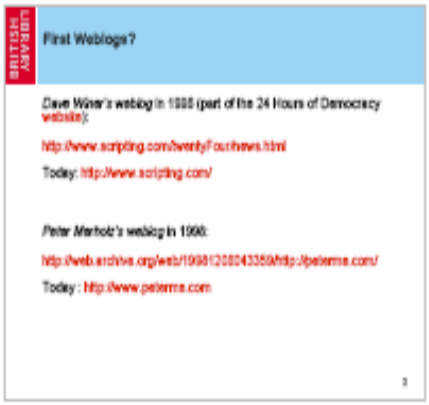
.....
Web Archivist

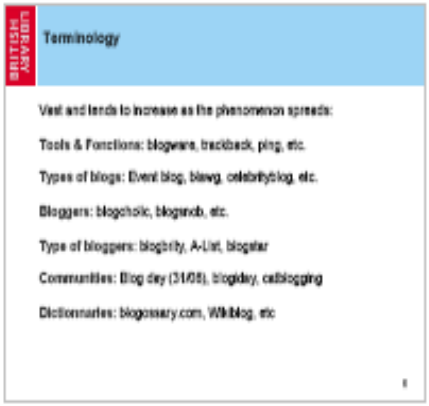
.....

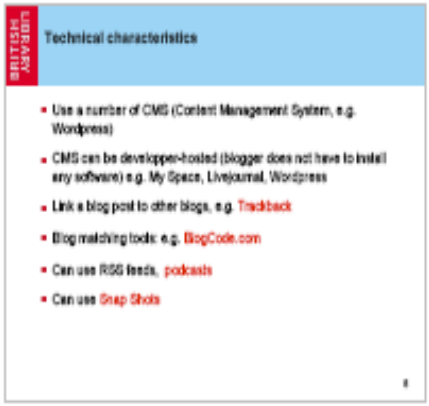
.....
Modern British Collections
The British Library
96 Euston Road
London

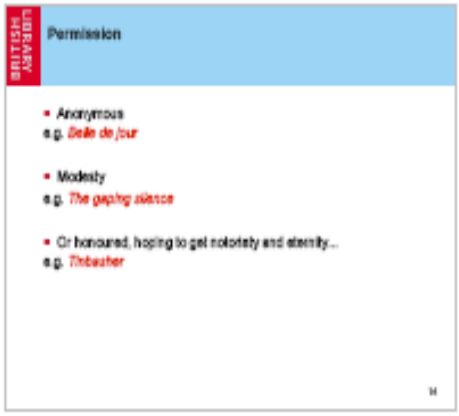
Annexe 4 Présentation Blogs











Annexe 5 Exemple de classement (BDD)

ID	Id BDD	Name	url	subject1	subject2	subject3	Type	archivé	date
1	4838	A Don's Life	http://timesonline.typej	Higher education	Arts & Humanities	Society & Culture	Journalist	1	19/07/07
2	4743	a hazy day today	http://therockmother.bl	Journals	Music			0	
3	4834	Access 2.0	http://www.bbc.co.uk/t	Computing & IT; the Web	Social problems & welfare		Journalist	0	
4	4487	Acerbia	http://www.acerbia.cor	Society & Culture	literature			0	
5	4780	ACHOCKABLOG	http://achuka.co.uk/ach	Literature	Journal	Children		1	02/07/07
6	4706	Adrenalin rush	http://howling-adrenali	Health	Society & Culture			1	22/06/07
7	4618	Alan in Belfast	http://alaninbelfast.blo	Journals				1	01/06/07
8	4488	alfred the ok	http://www.alfredtheok	Society & Culture	Journals			0	
9	4707	All about my movie	http://zummer.blogspot	Film	Journals			0	
10	3366	All Things Footie	http://www.allthingsfo	Sport				0	
11	4785	Alpha Mummy - Ti	http://timesonline.typej	Women	Society & Culture		Journalist	1	19/07/07
12	4532	Am I still me?	http://hollyfinch.blogsp	Society & Culture	Journals			0	
13	4750	An Unreliable Witn	http://www.unreliablew	Arts & Humanities	Journals			1	29/06/07
14	4619	And just then	http://andjustthen.blog	Journals				0	
15	4783	Andrew Keen	http://thecultoftheamat	Web & IT	Media	Society & Culture		0	
16	4804	Andrew McKie	http://blogs.telegraph.c	News & Media	Science & Technology	Society & Culture	Journalist	0	
17	3301	AndyPryke.com	http://www.andypryke.	Art	sports & recreation	Music		1	10/10/06
18	4489	angry chimp	http://www.angrychimp	Society & Culture	Literature			0	
19	4705	angry medic, The	http://angrymedic.blog	Health	Education & Research			0	
20	4612	Art blog and inspir	http://hedoesntloveyou	Art				0	
21	4787	Articles of faith	http://timescolumns.typ	Religion	Society & Culture		Journalist	1	19/07/07
22	4490	As a dodo	http://asadodo.blogspo	Society & Culture	Politics	economics	Group blog	0	
23	4617	Askew	http://medbh.blogspot.	Journals				0	
24	2336	Astronomy Blog	http://www.strudel.org	Physical sciences				1	17/01/05
25	4224	Bad Science	http://www.badscience	Popular science				1	28/06/07
26	2672	Baghdad Burning	http://riverbendblog.bl	Politics	News & Media	International relations		0	
27	4501	Banditry	http://www.johnband.o	Politics	Society & Culture	Computing, IT & the	web	1	10/05/07
28	4799	Baz Blog	http://bazblog.dailymai	Society & Culture				0	
29	4626	belfastbird	http://belfastbird.blogs	Women	Journals			0	
30	4798	Benedict Brogan's p	http://broganblog.daily	Government & Politics				0	
31	4491	Blatant optimism	http://www.sparklefluf	Society & Culture	Computing & IT, the Web			0	
32	4639	Blawging it in 'n.i.'	http://blawg-net-ni.blo	Society & culture				0	
33	4748	Blogdom of God	http://server.com/Web	Religion			Agrégateur	0	

Annexe 6 Principaux critères de sélection

CRITERES	OUTILS
Domaine .uk	→ Profil / Coordonnées → Moteurs / annuaires spécialisés (régions, pays, etc.) → DNS
Heritage culturel britannique	→ Moteurs / annuaires de recherche spécialisés → Ouvrages spécialisés → Sites d'Awards → Classements par les journaux → Aggrégateurs → Collaboration avec le personnel
Web innovation	→ Universités en ligne → Sites d'Awards → Sites répertoriant les sites innovants → Sites de professionnels
Hors du champs de sélection des autres membres	→ Politique de sélection des membres → Vérification de la base de données PANDAS
Technologies à éviter	→ Flash → Streaming (non téléchargeable) → Javascript → BDD dynamique → Taille des sites/Nombre d'urls
Politique de la BL	→ Neutralité → Représentativité → Contenu