Rapport de Recherche Bibliographique Mars 2004

DESS Réseaux d'information
et document électronique

L'archivage des sites web fédérateurs : Typologie, sélection, filtrage et métadonnées

Paul Belhouchat

Sous la direction de Omar Larouk
Maître de Conférences, Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques

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

Résumé :

Le web s'est imposé comme un canal majeur de diffusion de l'information, de la culture et du savoir. Les chercheurs ont pris conscience de la nécessité de préserver ces ressources qui sont menacées de disparaître. Cette question s'est transposée sur des sites web qui recensent des ressources à forte valeur informationnelles. Ce sont des sites fédérateurs, également appelés sites fédérateurs à qualité contrôlée. Ils offrent une alternative aux chercheurs et aux étudiants qui peuvent ainsi disposer de ressources fiables grâce à des bases de signets qui peuvent les orienter vers des pages ou des sites qu'ils ne trouveraient pas grâce à des moteurs de recherche classique, ce sont des ressources du Web Invisible. Cette recherche bibliographique veut faire un tour d'horizon sur ce qui se fait en matière d'archivage des sites fédérateurs, sur les méthodes utilisées, la périodicité d'enrichissement de l'archive, l'utilisation des métadonnées et les contraintes rencontrées par ces archivistes du web.

Descripteurs :

Site fédérateur, Site Fédérateur de Qualité Contrôlée, Archivage, Archivage du Web, Métadonnée, Filtrage d'informations, Périodicité, Web Invisible, Deep Web.

Abstract :

Since the 1990s, the Web has become one of the main channel of diffusion of information, culture and knowledge. Many researchers observe the necessity to think about the preservation of these contents which are threatened to disappeared if nothing is done. This question applies to another type of web sites which gather high quality resources, the Subject Gateway or Quality-Controlled Subject Gateway. Gateways propose an alternative to researchers and students thanks to reliable, relevant and high quality resources and documents which couldn't be found with classical search engine because these resources belong to the Deep Web or Invisible Web. This paper is a bibliographic research and a panorama of gateway archiving, of means of selection of relevant sites, archiving periodicity, use of metadata, and the difficulties encountered by web archivists in their projects.

Keywords :

Subject Gateway, Subject-Based Information Gateway, SBIGs, Quality-Controlled Subject Gateway, Archiving, Web Archiving, Metadata, Information Filtering, Periodicity, Invisible Web, Deep Web.

Sommaire

Introduction

Grâce à une évolution rapide et une croissance exponentielle, le web s'est imposé comme un vecteur incontournable de diffusion de l'information, de la culture et du savoir. Malheureusement, une grande partie de ces ressources disparaissent à la suite d'une mise à jour, de l'arrêt ou du déplacement d'un serveur web. L'enjeu est d'autant plus important lorsqu'il s'agit de préserver des sites dont le contenu est strictement évalué et filtré par des spécialistes d'une ou plusieurs disciplines. Cette question de la préservation du web et surtout des sites fédérateurs commence à occuper le quotidien de certains spécialistes du web, mais aussi du monde des bibliothèques – institution et infrastructure vouée à la conservation et à la valorisation de tout ce qui touche à la culture- qui réfléchit également au rôle qu'il pourra jouer dans ce domaine.

C'est dans cette optique que M. Omar Larouk, maître de conférences à l'Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques (ENSSIB) nous à chargé de mener une recherche bibliographique sur l'archivage des sites fédérateurs. Les contours de notre sujet prirent forme grâce aux premières recherches et aux lectures effectuées ainsi qu'au cours de nombreuses rencontres avec notre commanditaire.

Notre rapport va s'articuler autour de trois grandes parties : La première va présenter la méthode de recherche documentaire et les outils utilisés afin de répondre au mieux à notre commande en essayant de recenser les références les plus pertinentes mais aussi les plus récentes. La seconde partie est une synthèse d'articles jugés pertinents voire référentiels pour notre propos. Dans le cadre de ce rapport, la question de l'archivage des sites fédérateurs a été envisagée autour de trois grandes entrées : La typologie de ces sites, les types et méthodes d'archivage et le référencement (c'est-à-dire l'utilisation des métadonnées)[1]. Enfin, une bibliographie va présenter les références sélectionnées.

[1] Des thèmes connexes ont été abordés, le Web Invisible et les contraintes (notamment juridiques) de l'archivage.

Partie 1 : Méthodologie

1. Démarches préliminaires : L'état de l'art et la définition des mots-clés

1.1. L'existence de travaux récents et d'un congrès

Face à un sujet tel que l'archivage des sites fédérateurs, la tentation d'aller directement interroger les ressources électroniques est grande, mais il faut faire preuve de pondération et surtout d'organisation dans notre quête d'informations et de ressources. Tout d'abord, il a fallu réfléchir au sens même de la problématique, une réflexion nécessaire pour pouvoir délimiter d'une part le périmètre de notre étude mais surtout de dégager des mots-clés qui vont être notre fil d'Ariane tout au long de notre recherche.

Lors de la présentation de sa commande, O. Larouk nous a fourni de précieux renseignements sur l'état de l'art de l'archivage du web en général et sur les sites fédérateurs. En effet, il avait déjà soumis un rapport de recherche en 2003 sur l'utilisation des métadonnées dans les sites fédérateurs[2]. La même année, un mémoire de conservateur de bibliothèque ayant pour sujet l'archivage des sites web régionaux fut soutenu à l'ENSSIB[3].

Incontestablement, ces deux travaux vont former le socle de notre recherche puisque nous espérons y trouver des informations sur les sites fédérateurs et sur l'archivage du web en général, mais également une base bibliographique qui va nous permettre de cerner les principaux acteurs de ces recherches.

Notre commanditaire, intéressé par les avancées des diverses équipes internationales en matière d'archivage a également écrit un article qu'il a présenté

[2] Saidi S., *Utilisation des métadonnées dans les SBIG (Suject-Based Information Gateways) ou Sites Fédérateurs de Qualité Contrôlée*. Rapport de recherche bibliographique, Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques (ENSSIB), Villeurbanne, 2003, 55 p.

[3] Haettiger M., *L'archivage des sites Web d'intérêt régional*. Diplôme de Conservateur de Bibliothèque, Ecole Nationale Supérieure des Sciences de l'information et des Bibliothèques (ENSSIB), Villeurbanne, 2003, 121 p.

en 2002 lors d'un congrès où il donne une définition et une typologie des sites fédérateurs[4].

Enfin, une information capitale pour notre rapport est l'existence d'une conférence annuelle sur la recherche dans le domaine des bibliothèques numériques, l'ECDL (*European Conference on Resarch and advanced Technology for Digital Librairies*). Depuis 2001, l'ECDL organise un *Workshop* (atelier) sur l'archivage du web. La vocation de ce congrès est bien sûr de fournir une vision transversale sur la recherche et les pratiques dans un domaine récent et émergeant que représente l'archivage du web, mais aussi un forum pour permettre à divers spécialistes, qu'il s'agisse de bibliothécaires, d'archivistes, de chercheurs universitaires ou industriels, de présenter leurs méthodes et les améliorations ou les solutions nouvelles qu'ils préconisent pour développer ce domaine.

Avant même d'avoir débuté notre recherche, nous voici en possession d'informations très précieuses. Il s'agit de documents récents pour les mémoires, qui furent soutenus en 2003 et d'une information *a priori* a haute valeur ajoutée avec cette conférence sur l'archivage du web.

1.2. La définition des mots-clés

Avant d'aborder notre organisation et notre stratégie de recherche, une première étape consiste à cerner le vocabulaire qui tourne autour de ce qu'on peut appeler le champ sémantique de l'archivage du web. C'est peut-être une évidence, mais elle est utile à rappeler car elle prend tout son sens dans le cadre de notre étude. Une recherche efficace et réussie ne peut l'être que grâce à l'utilisation de mots-clés pertinents. La définition des mots-clés est l'étape qui précède la planification de notre stratégie.

Par contre, il faut garder à l'esprit que cette définition n'est pas figée *stricto sensu*. Il peut arriver qu'au cours de notre recherche, certains mots ne soient pas très efficaces, cette liste peut-être appelée à subir quelques mesures correctives.

---

[4] Larouk O., *Classifications et Méta-données pour le Web sémantique: Typologie et évaluation des sites fédérateurs*. In : Conférence internationale SETIT : Sciences Electroniques, technologies de l'Information et des Télécommunications, Ecole Nationale Supérieure des Télécommunications de Bretagne, ENST-Brest, 2003, p.235-251.

Voici comment nous avons établi cette liste de descripteurs et comment elle a pu être amenée à évoluer.

1.2.1. Les mots-clés en langue française

Bien que nous partions sur un *a priori* assez négatif sur les ressources que l'on pourrait retrouver en français, nous avons quand même décidé d'établir une liste de descripteurs dans la langue de Molière pour notre stratégie de recherche.

Nous verrons un peu plus tard que les résultats obtenus furent assez décevants puisque la majeure partie de la littérature qui traite de l'archivage du web est rédigée dans la langue des chercheurs, en anglais.

Un premier recensement donne les résultats suivants : "site fédérateur", "annuaire spécialisé", "répertoire spécialisé", "archivage", "métadonnée", "fusion d'informations", "filtrage d'informations", "périodicité", "web invisible", "deep web".

Explications : Pour les termes "site fédérateur" et "archivage", il n'y a aucun problème, ce sont les deux descripteurs de base. "Annuaire spécialisé" et "répertoire spécialisé" sont en fait des synonymes de "site fédérateur" et leur utilisation dans le cadre d'une requête peut être utile.

Le terme "métadonnée" s'impose aussi comme descripteur puisque comme nous le verrons dans la synthèse, les métadonnées sont utilisées dans les sites fédérateurs, elles représentent un gage de qualité de ces sites et sont essentielles dans le cadre d'un archivage.

Les termes "fusion d'informations" et "filtrage d'informations" ont été sélectionnés dans la mesure où les sites fédérateurs filtrent l'information pour former un outil de diffusion de ressources spécialisées dans un ou plusieurs domaines (d'où fusion également), un outil de ressources validées.

Le terme de "périodicité" est en corrélation avec l'archivage. Quand une institution a décidé de mener une politique d'archivage de sites web, elle doit définir la périodicité de cet archivage, par exemple une fois par semestre, tous les trois mois, etc.

Enfin, il faut également compter sur les notions de "web invisible" et son synonyme anglais qui est également utilisé en français, "*Deep Web*". Nous expliquerons plus tard ce qui se cache derrière ces termes mystérieux ; simplement, il faut savoir que les ressources d'informations scientifiquement validées sur le web, ainsi que de nombreux sites fédérateurs offrent des ressources qui sont incluses dans cette partie dite invisible du web. Donc, parler des sites fédérateurs, c'est également parler du Web Invisible.

1.2.2. Les mots-clés en anglais

Il s'agit *grosso modo* de la traduction des termes en français, mais avec quelques variantes. On vient de voir que l'on peut désigner un site fédérateur en français par répertoire ou annuaire spécialisé. En anglais, les spécialistes de la question ont développé une véritable terminologie pour dresser une typologie des sites fédérateurs. Ainsi, il n'existe pas moins d'une demi-douzaine de termes pour parler de ces types de sites : "*Subject Gateway*"[5], "*Information Gateway*"[6], "*Academic Subject Gateway*"[7], "*Subject-Based Information Gateway*" ou "*SBIGs*"[8, 9].

A la lecture du rapport de S. Saidi, mais aussi d'un article qu'elle a utilisé pour définir cette terminologie, nous avons appris qu'un terme a été développé par T. Koch pour rassembler l'ensemble de cette terminologie. Il parle de ce qu'on pourrait traduire par "site fédérateur de qualité contrôlée" ou "*Quality-Controlled Subject Gateway*"[10].

Pour les autres termes, nous pouvons utiliser comme descripteurs : "*Archiving*", "*Metadata*", "*Information Filtering*", "*Information Fusion*", "*Periodicity*", "*Invisible Web*", "*Deep Web*".

---

[5] Campbell D., *Australian subject gateways -metadata as an agent of change*. In : Books and bytes: technologies for the hybrid library : proceedings, 10th biennial conference and exhibition, 16-18 February, 2000, Melbourne Convention Centre / Victorian Association of Library Automation Inc.p.421-430.

[6] Hiom D., *SOSIG: an Internet hub for the social sciences, business and law*. Online Information Review, 2000, vol. 24, n°1, 2000, p.54-58.

[7] Peereboom M., *DutchESS: Dutch Electronic Subject Service - a Dutch national collaborative effort*. Online Information Review, 2000, vol.24, n°1, p.46-48.

[8] Ardoe A., Berggren M., Koch T., Kringstad R., *Nordic Interconnected Subject Based Information Gateways (NISBIG). Final report*. Nordinfo-Nytt, vol.23, n°3, 2000, p.7-33.

[9] Sur cette terminologie, je me réfère au travail de S.Saidi, *op. cit.*, p.23.

[10] Koch T., *Quality-controlled subject gateways: definitions, typologies, empirical overview*. Online Information Review, 2000, vol.24, n°1, p.24-34.

Cette partie n'avait pas la prétention immédiate d'entrer tout de suite dans le vif du sujet, nous aborderons ces problèmes de typologie et de terminologie au cours de la synthèse. Nous avons suivi les pistes de lecture de départ suggérée par notre commanditaire, des pistes qui nous ont permis d'avoir une première approche du sujet. La lecture des mémoires de S. Saidi et M. Haettiger ont grandement contribué à dresser une liste de mots-clés, une liste de base qui est susceptible d'être alimentée au cours de nos recherches.

1.3. Planification de la stratégie de recherche

Après la présentation de la commande et des recommandations de départ données par O. Larouk, puis un premier travail de documentation à travers des lectures préliminaires qui ont permis de dresser une première liste de descripteurs, nous voilà parés pour entamer la recherche bibliographique.

Deux questions vont se poser dans le cadre de la méthodologie : Où ? et comment ?

La première démarche consiste à choisir les outils de recherche qui sont à notre disposition et qui seraient susceptibles de répondre à notre sujet, c'est à dire des outils qui recenseraient des documents parlant de l'archivage des sites fédérateurs et de l'utilisation des métadonnées dans le cadre de ce travail d'archivage.

Une remarque s'impose d'emblée néanmoins, l'archivage du web et *a fortiori* l'archivage des sites fédérateurs est un sujet de travail relativement récent qui commence tout juste à mobiliser les énergies. On l'a dit, la conférence annuelle sur l'archivage du web arrive en 2004 à sa quatrième édition (prévue courant 2004 à Bath en Grande-Bretagne).

De cette nouveauté et de cette spécificité d'un sujet qui commence tout juste à sortir du confinement des centres d'études et de recherches, on peut légitimement émettre comme hypothèse que des monographies sur la question ne doivent pas être légion et que l'essentiel de la littérature doit être composé d'articles et d'actes de congrès.

Voilà pour ce qui est des hypothèses de départ concernant les ressources qui peuvent être à notre disposition. Et voici dans l'ordre chronologique, les outils qui ont permis de faire notre recherche.

Malgré la tentation d'aller directement interroger les bases de données en ligne vu le côté technique du sujet, nous avons préféré démarrer la recherche documentaire de façon classique, c'est-à-dire de consulter les catalogues de trois bibliothèques, celle de l'ENSSIB, de la Bibliothèque Municipale de Lyon et de la Bibliothèque Nationale de France. Dans un second temps, nous avons fait appel au catalogue des thèses du SUDOC via le site web de l'Abes.

Deuxième étape, la recherche via les bases de données en ligne. Etant donné que l'ENSSIB dispose d'un certain nombre d'abonnement à des bases de données, il fallait faire un choix vers celles qui donneraient les résultats les plus performants. Nous avons choisi deux bases de données de références bibliographiques : Celle de l'entreprise canadienne Dialog, et une de ses concurrentes américaines, la base du Cambridge Scientific Abstracts (CSA). Nous avons aussi opté pour une base de périodique en ligne, Emerald.

Enfin, une troisième étape est une recherche sur le web par l'intermédiaire du moteur de recherche Google. Un annuaire qui est aussi un site fédérateur va être utilisé, il s'agit en l'occurrence du site écossais Bubl.

Toujours dans le cadre du web, nous nous sommes penchés sur le site du World Wide Web Consortium (W3C), un site de référence pour tous les chercheurs du web qui recense tous les travaux et toutes les normes propre au fonctionnement du web mondial ; il s'agit donc d'une source potentiellement intéressante pour notre sujet, notamment pour ce qui est du référencement des sites.

Nous verrons que ces trois étapes vont nous permettre de faire un focus aussi précis que possible sur notre sujet. Les résultats ainsi obtenus pourront être également comparés avec ceux que nous avons observés au cours des lectures introductives des rapports de S. Saidi et M. Haettiger.

2. Première étape : les bibliothèques et les thèses

2.1. La bibliothèque de l'ENSSIB[11]

Ecole spécialisée dans les sciences de l'information, la consultation de son catalogue dans le cadre de notre sujet s'imposait d'elle-même.

La recherche se fait dans le formulaire de recherche simple, à noter que l'on peut utiliser un opérateur de troncature grâce à l'arobase "@", un opérateur d'adjacence avec le tiret "-" et les opérateurs booléens "et", "ou", "sauf". Voici les résultats obtenus en interrogeant dans tous les champs :

N°	REQUETE	RÉSULTATS
1	Gateway@	3
2	Site@-federateur@	0
3	Archivage	32
4	Metadonn@	4
5	Metadata	7
6	Gateway@ ET archiv@	0
7	Web ET archiv@	3

Le bilan de la consultation de l'OPAC de l'ENSSIB est assez décevant. La requête n°2 sur les sites fédérateurs ne donne rien, même si prochainement elle devrait mettre en lumière le rapport de S. Saidi. Son équivalent anglais "gateway" donne trois résultats, mais ils ne répondent pas à notre demande.

Par contre, la requête n°3 sur l'archivage donne 32 résultats qui concernent essentiellement l'archivage dans sa globalité et deux références pertinentes : Le rapport de M. Haettiger et celui de J. Masanès sur l'archivage des sites web. A noter également un travail sur le dépôt légal des sites web, c'est une référence intéressante même si elle n'entre pas directement dans notre sujet.

En ce qui concerne les métadonnées, la requête formulée en français donne quatre références qui traitent toutes sur le Dublin Core, quant à la version en

[11] <http://www.enssib.fr/>

anglais, elle débouche sur sept résultats dont deux très intéressants : Il s'agit de la thèse de Y. Amerouali[12] sur les différents standards et du travail de S. Lazinger sur la préservation des données numériques et des métadonnées utilisées dans le but de cette préservation[13].

Enfin, une dernière requête avait pour but de combiner l'archivage et le web, elle ne donne que peu de résultats probants, si ce n'est un croisement avec le mémoire de M. Haettiger.

Finalement, la consultation du catalogue de l'ENSSIB a été assez décevante ; cela peut s'expliquer du fait de la nouveauté du sujet, mais aussi à cause d'une description bibliographique des ressources assez lacunaire, hypothèse corroborée par une recherche manuelle dans les périodiques de l'ENSSIB qui nous ont permis de mettre la main sur des articles pertinents pour notre propos et que l'OPAC passait sous silence[14].

2.2. La bibliothèque municipale de Lyon et la BnF[15]

La consultation de l'OPAC de la bibliothèque de la Part-Dieu n'a pas fourni de surprise particulière, les requêtes "gateway@" et "archivage" donnent respectivement seize et treize résultats qui ne sont pas pertinents. Les requêtes concernant les métadonnées ne donnent aucun résultat.

Enfin, nous avons choisi la BnF en espérant qu'elle pourra nous offrir des ressources intéressantes pour notre recherche. En qualité de bibliothèque nationale, elle possède le catalogue le plus riche de France et il est important de souligner que le département de la bibliothèque numérique de la BnF mène une politique de préservation à long terme des sites web sous l'impulsion de C. Lupovici et de J. Masanès. On peut donc espérer collecter des informations sur l'archivage du web.

---

[12] Amerouali Y., *Métadonnées basées sur l'association d'éléments de description de ressources et d'éléments de profil d'utilisateur*. Thèse de doctorat, Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques (ENSSIB),Villeurbanne 2001, 243 p.

[13] Lazinger S. S., *Digital Preservation and Metadata: History, Theory, Practice*, 2002.

[14] C'est le cas, entre autres, de la revue *Archimag*, qui recèle quelques articles intéressants. Cf. Roumieux O., *La quadrature du Web : Le Web à la recherche de sa mémoire (The quadrature of the Web)*. Archimag, 2001, n°145, p.30-32.

La recherche se fait sur le catalogue BN-Opale-plus avec le choix d'utiliser un formulaire simple, une recherche avancée/experte et l'utilisation d'équation dans un troisième formulaire.

Malheureusement, les résultats sont relativement décevants, le fait d'interroger seulement dans les champs "sujet" et "titre" constitue une limite. Par exemple la requête "archivage" donne comme résultats seize références pour l'archivage numérique et douze pour l'archivage électronique, pas une ne sera retenue.

Par ailleurs, nous avons essayé une recherche par auteur avec les deux spécialistes du département de la bibliothèque numérique de la BnF, J. Masanès n'est pas recensé, et C. Lupovici y figure une fois, mais il s'agit d'un travail qui n'est pas en rapport avec notre sujet.

2.3. Les thèses[16]

Après avoir consulté les différents catalogues de bibliothèques, nous nous sommes intéressés directement à un catalogue de thèses. Un sujet récent a dû susciter de l'intérêt chez les jeunes chercheurs et il est fort probable que des thèses de doctorat soient publiées. L'OPAC de l'ENSSIB a révélé l'existence du travail de Y. Amerouali. Il existe deux outils pour consulter cette littérature : Le CD-rom Docthèse et le Sudoc en ligne. Nous avons sciemment laissé de coté le CD-rom pour des raisons évidentes de mise à jour, puisqu'il existe une base de donnée en ligne il est plus logique de s'y fier. Le Sudoc est donc disponible en ligne sur le site de l'Abes. On peut y utiliser les opérateurs booléens, l'étoile "*" pour la troncature, le dièse "#" pour faire une recherche d'expression et enfin le point d'interrogation "?" comme opérateur de proximité.

N°	EQUATION DE RECHERCHE	RESULTATS
1	Site*#federateur*	0
2	Metadonn*	5
3	Metadonn* ? archiv*	0
4	Metadata	20

[15] <http://www.bm-lyon.fr/> ; <http://www.bnf.fr>
[16] <http://www.sudoc.abes.fr>

Nous n'avons rien trouvé concernant l'archivage du web en général et celui des sites fédérateurs en particulier à l'exception du travail de M. Haettiger. Quelques ressources sont relevées sur les métadonnées dont la thèse de Y. Amerouali. Le bilan est également décevant.

2.4. Bilan de la première étape

Pour dresser un bilan sur cette première étape, les bibliothèques via leurs OPAC n'ont pas été à la hauteur de nos espérances, mais il y a une justification qui ne met nullement en cause la qualité de leur catalogue. D'abord, il s'agissait ici de retrouver des monographies, des thèses éditées : L'archivage du web et des sites fédérateurs est un nouveau sujet de recherche qui a moins d'une décennie, il est logique qu'il y ait peu de publications de ce type, beaucoup de travaux seront amenés à faire leur apparition dans les années à venir.

Par contre, ces bibliothèques possèdent des collections de périodiques qui traitent des sciences de l'information, de l'archivage et notamment des enjeux de l'archivage des matériaux numériques. Ces revues sont justement une des tribunes utilisées par les chercheurs pour publier leurs avancées, leurs propositions. Les OPAC ne fournissent pas une description suffisamment détaillée, ils ne descendent pas à un degré de granularité tel qu'ils pourraient décrire le contenu d'un numéro d'un périodique quelconque. Bien sûr, une recherche manuelle dans ces périodiques constitue une contrainte en matière de gestion du temps.

La première étape étant bouclée, nous pouvons dès lors interroger les serveurs de base de données qui vont pouvoir nous permettre de recenser les articles publiés sur notre sujet et de connaître les principaux périodiques sur la matière.

3. Deuxième étape : les bases de données en ligne

Nous entrons ici à un niveau de profondeur supplémentaire dans notre quête. C'est grâce à ces bases de données que nous allons constituer notre bibliographie puisque la première étape ne nous a permis d'aborder le sujet que sur

un angle introductif. Il reste à sélectionner les bases les plus pertinentes. L'ENSSIB a contracté plusieurs abonnements auprès de différents serveurs commerciaux, de même qu'il existe des ressources en ligne gratuites.

Nous avons donc opté pour trois bases qui fournissent une information très bien identifiée, de qualité, émanant de revues faisant autorité. Dialog, CSA et Emerald.

3.1. L'interrogation du serveur Dialog[17]

3.1.1. Présentation du serveur

Avec ses 450 bases de données qui couvrent les besoins en information de pratiquement tous les secteurs de la recherche et des principaux domaines industriels, le serveur de la société canadienne Thomson est quelque part un archétype de l'exhaustivité de l'information scientifique et technique. Mais elle n'est pas dénuée de qualité puisqu'il s'agit ici d'information analysée et validée.

Dialog permet une interrogation multibase et fournit une description détaillée de ses ressources grâce à des résumés assez bien rédigés. Ce type de description, couplé avec une interrogation multibase, apporte un gain de temps précieux dans l'analyse de la pertinence puis dans la sélection d'un document.

3.1.2. Les catégories de Dialog sélectionnées

Dialog regroupe ses bases de données dans des catégories thématiques appelées *One Search Category*. Après observation de l'ensemble des catégories, il est apparu que celles qui nous intéressaient étaient classées dans deux grandes catégories puisque notre sujet possède deux entrées ; en effet l'archivage des sites fédérateurs touche au domaine des sciences de l'information (catégorie *Social Science and Humanities*), mais aussi du web, donc d'Internet et par extension à l'informatique (catégorie *Science and technology*).

Pour la catégorie *Social Science and Humanities*, c'est la sous-catégorie *Library Information Science* (11 bases de données) qui va nous être la plus utile.

[17] <http://www.dialogweb.com>

Pour la catégorie *Science and Technology*, il faut compter sur la sous-catégorie *Computers, Electronics and Telecommunications* (16 bases)

3.1.3. Les bases de données sélectionnées

Par contre, il est hors de question d'interroger simultanément 27 bases, un travail de tri s'impose pour ne garder que les bases les plus pertinentes. Ce tri est possible grâce à la lecture des fiches techniques de chacune des bases, les *Bluesheets*. Sur les 27 bases de départ, 2 sont à dédoublonner puisqu'on les trouve dans les deux grandes catégories (*Dissertation Abstract Online* [35] et *Information Science and technology Abstracts* [202]).

Après lecture des fiches techniques, nous avons retenu les bases qui traitent des sciences de l'information en relation avec les nouvelles technologies d'information et de communication. De cette seconde sélection, 9 bases sont retenues.

Enfin, un troisième "écrémage" est réalisé. Pour cela, nous avons lancé notre première équation relative aux sites fédérateurs (S1). Une analyse rapide des résultats a permis de localiser les bases les plus pertinentes, deux bases sont éliminées puisqu'elles ne fournissent aucun résultat.

Voici les bases retenues :

- ***Inspec*** [2]: Base spécialisée en physique, électronique et informatique (plus de 4100 périodiques). Cette base constitue à elle seule une référence, d'ailleurs, elle est également disponible sur *Science Direct*[18].
- ***NTIS*** (*National Technical Information Service*) [6]: Spécialisée, entre autres, en informatique et en sciences de l'information, cette base recense notamment les références des travaux des plus grands centres de recherches américains, mais aussi d'autres institutions telles que le CNRS.
- ***Social SciSearch*** [7]: Base produite par l'*Institute for Scientific Information* (ISI), elle recense des références sur les sciences sociales et les sciences de l'information à travers plus de 1500 périodiques, mais aussi à

[18] C'est une des raisons qui nous a conduit à ne pas interroger Science Direct.

travers 2400 autres périodiques touchant d'autres sciences mais pouvant contenir des articles relatifs à son domaine de prédilection.

- ***Information Science and Technology Abstracts*** [202]: Base qui s'appuie sur plus de 300 revues.
- ***Library Literature and Information Science*** [438]: Recense plus de 229 périodiques "phares" sur les sciences de l'information publiées aux Etats-Unis et dans le reste du monde.
- ***Inside Conference*** [65]: Base produite par la British Library. Comme son nom l'indique, elle recense tous les congrès, colloques, symposium, ateliers (Workshop) qui sont enregistrés à la British Library. (près de 5 millions de références en novembre 2003).
- ***Pascal*** [144]: Base pluridisciplinaire produite par l'Institut National de l'Information Scientifique et Technique (INIST), elle peut nous fournir des résumés voire des articles en français.

Les bases étant choisies, nous pouvons commencer la recherche.

3.1.4. Equations et stratégie de recherche

La stratégie adoptée tourne autour de quatre entrées qui vont être croisées entre elles : Les sites fédérateurs, l'archivage, les métadonnées et le Web Invisible.

Pour démarrer la recherche sur le serveur, il faut d'abord sélectionner les sept bases grâce à la commande : b 2,6,7,202,438,144,65.

Vu son coté spécifique, la recherche s'est faite en interrogeant tous les champs. Dialog permet la combinaison des équations et l'utilisation de nombreux opérateurs. Citons un opérateur de proximité (W), de troncature (?), les booléens. La commande "PY "permet de sélectionner des références en fonction de leur date, la commande RD permet de supprimer les doublons.

Voici les équations et les résultats obtenus :

A : <u>Cerner la littérature sur les sites fédérateurs</u>

Set	Term Searched	Items
S1	SUBJECT(W)GATEWAY? OR INFORMATION(W)GATEWAY? OR SUBJECT(W)BASED(W)INFORMATION(W)GATEWAY? OR QUALITY(W)CONTROLLED(W)SUBJECT(W)GATEWAY? OR SBIG?	445
S2	S1 NOT ASTRONOM? NOT ECLIPS? NOT STELL?	376
S3	S2 AND PY=>2000	225
S4	RD S3 (unique items)	175

L'équation S1 permet de faire ressortir l'ensemble des articles traitant des sites fédérateurs grâce à ses différentes dénominations. Par, contre le terme SBIG pose un problème d'homonymie avec une célèbre entreprise de matériel astronomique (télescope, appareils de photographie spatiale, etc.) : La société Santa Barbara Instrument Group. Les 175 résultats obtenus correspondent bien à ce que nous recherchons.

B : <u>L'archivage ou la préservation des ces sites</u>

S5	S4 AND ARCHIV?	10
S6	S4 AND PRESERV?	3

Ces deux équations ont pour but de demander des références sur l'archivage de ces sites. On a également utilisé le terme de préservation, souvent utilisé par les spécialistes.

C : <u>La typologie des sites fédérateurs</u>

S7	S4 AND TYPOLOGY	1
S8	GATEWAY AND TYPOLOGY	11

Ces deux équations fournissent de bons résultats concernant la typologie de ces sites. Notons que l'équation S7 isole l'article de T. Koch qui est une référence pour notre sujet.

D : Les métadonnées : liens avec les sites fédérateurs typologie et rôle dans l'archivage

S9	METADATA	6285
S10	S4 AND S9	22
S11	METADATA AND TYPOLOGY	14
S12	S11 and PY=>2000	6

Première étape, la recherche de renseignements généraux sur les métadonnées, puis d'articles assez récents (à partir de 2000) pouvant établir une typologie.

S13	METADATA AND ARCHIV?	718
S14	METADATA AND WEB AND ARCHIV?	157
S15	S14 AND GATEWAY?	10

Deuxième étape: le rôle des métadonnées dans l'archivage en général puis au niveau du web et enfin à l'échelle des sites fédérateurs. C'est à partir de la requête S15 qu'on s'aperçoit que les résultats se croisent avec ceux des autres équations.

E : Le Web Invisible : Généralités, collecte des sites et archivage

S16	INVISIBLE(W)WEB	144
S17	DEEP(W)WEB	45
S18	DEEP(W)WEB OR INVISIBLE(W)WEB	179
S19	RD S18 (unique items)	136
S20	S19 AND PY=>2000	119

Les 119 résultats ci-dessus représentent des travaux relatifs au web invisible. Pour bien cerner ce thème, il fallait également employer le terme de "deep web".

S21	S19 AND ARCHIV?	8
S22	S19 AND PRESERV?	2
S23	S19 AND HARVEST?	4

A partir de ces 119 résultats, on combine des termes relatifs à l'archivage (S21), à la préservation (S22), mais aussi à la collecte ou aspiration de ces sites dans le cadre d'une politique d'archivage. Le terme anglais utilisé est le verbe "*to harvest*", les logiciels qui font cette collecte de sites sont des *harvesters*.

D'une équation à l'autre, on trouve les mêmes articles, mais l'ensemble est relativement satisfaisant au niveau de la pertinence.

F : L'archivage du Web en général

S24	WEB(W)ARCHIV? OR INTERNET(W)ARCHIV ?	199
S25	RD S24 (unique items)	170
S26	S25 AND PY=>2000	111

Nous avons fait une équation sur l'archivage du web en général puisque les équations précédentes sur l'archivage du Web Invisible ont donné un nombre très faible de résultats. Ici, nous obtenons 111 résultats dont une bonne partie concerne en fait l'organisation américaine Internet Archive. L'équation "INTERNET(W)ARCHIV?" est formulée sous la forme d'une recherche d'expression ici, il est logique que le moteur de Dialog puise dans son index tout ce qui touche à cette société, mais comme l'équation est couplée avec "WEB(W)ARCHIV?", la requête conserve néanmoins sa pertinence. En S29, nous reformulerons cette requête.

G : Corrélations entre l'archivage du Web et les sites fédérateurs, entre l'archivage du Web et les métadonnées

S27	S26 AND S4	0
S28	S26 AND METADATA	8

La requête S28 donne approximativement les mêmes références que la requête S15.

S29	(WEB OR INTERNET) AND ARCHIV?	3741
S30	S29 AND S4	5

Ici, reformulation de l'équation S24 qui génère trop de bruit. On combine ces résultats avec la requête S4 sur les sites fédérateurs. Les références obtenues se retrouvent en partie dans les résultats trouvés en S5, S6 et S8.

H : Filtrage et fusion d'informations

S31	INFORMATION(W)FILTER?	1471
S32	S4 AND S31	1
S33	S31 AND METADATA	18

Les résultats ne sont pas très probants, le problème du filtrage des informations dans les sites fédérateurs commence tout juste à être traité.

S34	INFORMATION(W)FUSION	2042
S35	S34 AND S4	0

Ici aussi, les résultats sont décevants puisqu'on obtient du bruit puis du silence. La notion de fusion d'information n'est pas encore traitée en tant que telle comme un descripteur.

I : La question de la périodicité de l'archivage du Web

S36	PERIODICITY AND S4	0
S37	ARCHIV? AND PERIODICITY	46
S38	S37 AND GATEWAY?	0
S39	S37 AND WEB	0

La requête S37 donne des résultats hors-sujet pour notre propos, nous avons beaucoup d'articles sur la fréquence d'observation des étoiles, cette équation s'avère très pertinente pour ce qui touche aux domaines de l'astronomie, de l'astrophysique, de la physique nucléaire, mais aussi de la médecine.

J : <u>La recherche de référence sur la conférence annuelle de l'ECDL</u>

S40	EUROPEAN(W) CONFERENCE(W) ON(W) RESEARCH(W) AND(W) ADVANCED(W) TECHNOLOGY(W) FOR(W) DIGITAL(W) LIBRARIES OR ECDL	552
S41	S40 AND ARCHIV?	23

Vu que nous étions déjà au courant de l'existence de cette conférence, une requête la concernant s'imposait d'elle-même. Les résultats sont intéressants, même si beaucoup ne font que citer cette conférence sans aller au -delà.

K : <u>L'archivage des sites fédérateurs et les métadonnées: recherche en langue française</u>

S42	SITE?(W)FEDERATEUR?	0
S43	ARCHIVAGE AND (WEB OR INTERNET)	191
S44	S43 AND PY=>2000	108
S45	RD S44 (unique items)	107
S46	REPERTOIRE?(W)SPECIALISE ?	4
S47	METADONN?	1080
S48	S47 AND ARCHIVAGE	45

Après analyse des résultats, le bilan est assez mitigé, on n'a pratiquement rien sur les sites fédérateurs et sur l'archivage du web. Mais, il faut souligner que les chercheurs français publient leur article en anglais. Même si Dialog fournit quelques résumé en français, on y trouvera que très peu d'articles rédigés dans la même langue. Par exemple, les références renvoyant aux travaux de J. Masanès sont toutes en anglais.

3.1.5. Bilan de Dialog

En ayant concentré nos efforts autour des sites fédérateurs, de l'archivage, des métadonnées et du Web Invisible, Dialog a permis de dresser une première esquisse de bibliographie. Le fait d'avoir formulé différemment nos requêtes et d'avoir systématiquement croisé les équations nous a permis de cerner des références pertinentes qui étaient listées dans plusieurs écrans de résultats.

Maintenant, nous allons utiliser une autre base de donnée, précisément spécialisée en science de l'information afin de corroborer ou infirmer la pertinence des résultats obtenus avec Dialog. Cette base peut également nous informer sur d'autres articles passés sous silence par le serveur canadien.

3.2. L'interrogation de LISA[19]

3.2.1. Présentation de la base

La base LISA (*Library Information Science Abstracts*) est une base de donnée bibliographique de plus de 440 titres spécialisés en bibliothéconomie et en science de l'information. Elle est la propriété du diffuseur américain CSA (*Cambridge Scientific Abstracts*). Il va sans dire que sa consultation dans le cadre de notre recherche est incontournable et peut constituer un complément de qualité aux résultats collectés précédemment.

3.2.2. Stratégie adoptée.

La stratégie d'interrogation adoptée ici est sensiblement la même que celle réalisée sur Dialog. Nous avons mené la recherche grâce au formulaire avancé. La troncature est symbolisée par l'étoile (*), la recherche d'expression par les parenthèses. On peut interroger sur tous les champs, mais aussi plus précisément sur les mots-clés (KW) et le titre (TI). On peut donner des intervalles temporels.

[19] <http://www.csa.com/>

3.2.3. Résultats de la recherche[20]

La première colonne de résultats (Items) représente les références contenues dans la base LISA, quant à la seconde colonne que nous avons appelé "Web", elle recense des pages ou des sites web *a priori* pertinents à la requête formulée.

N°	EQUATION DE RECHERCHE	ITEMS	WEB
1	KW=((subject gateway*) or (information gateway*) or (subject based information gateway*) or (quality controlled subject gateway*) or sbig*)	151	1
2	TI=((subject gateway*) or (information gateway*) or (subject based information gateway*) or (quality controlled subject gateway*) or sbig*)	43	1
3	(KW=archiv* and KW=((subject gateway*) or (information gateway*) or (subject based information gateway*))) or KW=((quality controlled subject gateway*) or sbig*) Limited to:2000-2004	10	0
4	(KW=((deep web) or (invisible web)) and KW=((subject gateway*) or (information gateway*) or (subject based information gateway*))) or KW=(quality controlled subject gateway*) Limited to:2000-2004	5	0
5	(KW=metadata and KW=((subject gateway*) or (information gateway*) or (subject based information gateway*))) or KW=(quality controlled subject gateway*)	16	0
6	KW=((deep web) or (invisible web)) Limited to:2000-2004	52	1
7	KW=((deep web) or (invisble web)) and KW=archiv* Limited to:2000-2004	4	0
8	(KW=(information filter*) and KW=((subject gateway*) or (information gateway*) or (subject based information gateway*))) or KW=Sbig* Limited to:2000-2004	2	0
9	KW=((web archiv*) or (internet archiv*))	30	1
10	KW=preserv* and KW=((deep web) or (invisible web))	0	0
11	KW=(site* federateur*)	0	0

[20] Les équations présentées ici sont une retranscription des requêtes réalisées dans le formulaire de recherche avancée. Autrement dit, il s'agit de la forme qu'il aurait fallu mettre en place dans le formulaire simple.

12	KW=((repertoire* specialise*) or (annuaire* specialise*))	0	0
13	KW=metadonn*	6	0

Comme pour la recherche sur Dialog, ces équations ont permis de bien situer notre sujet, puisque nous recensons peu de références hors-sujet. L'équation n°1 fait appel à un recensement global des ressources relatives aux sites fédérateurs, puis nous avons associé à ces premiers résultats le thème de l'archivage, des métadonnées et du Web Invisible. Les résultats obtenus sont globalement satisfaisants et corroborent les données collectées sur Dialog. Toutefois, il faut dire que la recherche sur Dialog a débouché sur de meilleurs résultats, LISA ne nous apporte pas d'éléments nouveaux à notre recherche.
Un dernier point pour signaler des ressources en langue française très lacunaires, mais ce n'est pas une surprise.

3.3. L'interrogation d'une base d'article : Emerald[21]

Nous avons choisi d'interroger une dernière base, en l'occurrence Emerald. Cette base donne accès à une centaine de périodiques spécialisés en management, en sciences de l'ingénieur et surtout en sciences de l'information et des bibliothèques. Cette base est relativement attrayante pour notre étude d'autant plus qu'elle propose un accès direct au texte intégral, contrairement aux autres bases.

Si nous avons décidé de consulter Emerald en dernier, c'est bien sûr parce qu'il fallait déjà avoir une idée des ressources qui seraient les plus pertinentes en travaillant d'abord sur Dialog qui est la base la plus puissante que nous avons consulté, puis sur LISA qui a permis d'identifier grâce aux croisements de ses résultats avec ceux de Dialog des auteurs référentiels ou des revues incontournables qui s'intéresseraient de près à l'archivage des sites fédérateurs. Ainsi, on espère qu'Emerald corrobore nos premières investigations et mette à notre disposition des documents en texte intégral[22].

[21] <http://www.emeraldinsight.com>
[22] La stratégie d'interrogation de la base Emerald obéit à la même logique de ce que nous avons fait sur Dialog et LISA. L'interrogation est effectuée dans le formulaire de recherche avancé, avec utilisation des opérateurs booléns, de proximité et une limitation de date entre 2000 et 2004.

Voici les résultats obtenus :

N°	EQUATION	RÉSULTATS
1	“subject gateway” and archiv*	23
2	“subject based information gateway” and archiv*	5
3	Sbig* and archiv*	4
4	“subject based information gateway”	1
5	“deep web” or “invisible web”	23
6	Metadata and gataway	9
7	Metadata	119

Emerald a plus ou moins corroboré nos premières recherches, les résultats les plus pertinents sont ceux qui ont été recueillis précédemment sur Dialog et LISA. L’avantage de cette base, hormis son accès direct au texte intégral, c’est qu’il met en lumière peu de résultats sur certaines équations (excepté l’équation n°7), mais ils sont pratiquement tous pertinents.

3.4. Conclusion sur les bases de données

Indiscutablement, l’utilisation des bases de données en ligne a donné une autre dimension à notre recherche. Remarquons le rôle de Dialog qui, grâce à une interrogation sur plusieurs bases a permis d’identifier rapidement des ressources pertinentes. Soulignons aussi le poids de la base INSPEC, véritable centre névralgique de notre recherche et “vivier” de références pertinentes. L’interrogation de LISA a eu pour objectif de confirmer la pertinence où non des premiers résultats collectés, enfin Emerald à également servi à corroborer ou infirmer nos références et à fournir des documents en texte intégral.

Nous pouvons dès lors entamer la dernière grande étape de notre recherche. Puisque notre sujet touche au web, il est tout à fait légitime de s’interroger sur ce que pouvons nous retrouver à travers ses outils de recherche.

4. Troisième étape : La recherche sur le Web : moteurs, sites fédérateurs et listes de discussions

Un élément paraît évident à nos yeux, la question de l'archivage des sites fédérateurs et de l'utilisation des métadonnées, tout comme l'archivage du web en général, ne peut pas être passé sous silence à travers les millions de références stockées dans les index des moteurs et des annuaires de recherche du web.

Il nous paraissait plus sage d'établir une recherche sur le web après le dépouillement des OPAC et des bases de données en ligne, le web étant un outil d'auto-publication, nous pourrions très facilement nous faire piéger par des références *a priori* séduisantes mais qui ne seraient en aucun cas pertinentes.

Pour mener à bien cette recherche, nous avons décidé d'utiliser le moteur de référence du web, Google. Au départ, nous voulions travailler aussi avec son concurrent Alltheweb, mais il est vite apparu que les résultats étaient globalement les mêmes, donc au bout de dix équations environ, nous nous sommes uniquement concentrés sur Google. La même démarche fut effectuée avec l'interrogation d'un métamoteur, en l'occurrence Copernic qui a corroboré les redondances observées avec Alltheweb.

Par ailleurs, nous avons choisi un site fédérateur, le site crée par l'université d'Edimbourg, l'annuaire spécialisé Bubl qui contient une rubrique sur les sciences de l'information et sur Internet en général.

Un autre site va être passé au crible, il s'agit de celui du Consortium du World Wide Web, susceptible de nous fournir de précieuses informations sur les métadonnées et éventuellement sur l'archivage des sites web.

Enfin, le rapport de S. Saidi et le Département de la bibliothèque numérique de la BnF nous ont appris qu'il existait des listes de discussions pour les spécialistes de l'archivage des ressources numériques. C'est une piste intéressante à étudier.

4.1. Le moteur Google[23]

Les moteurs présentent plusieurs contraintes par rapport aux bases de données. On ne peut pas combiner les équations et Google limite sa requête à dix mots.

En second lieu, nous nous sommes fixé une "charte" d'évaluation des références fournies : Ainsi une page ou un site peut attirer notre attention et être jugé pertinent non pas en fonction de son classement mais en fonction de ces éléments :

- L'URL contient le suffixe ".org", ".edu", ".ac.uk", ".gov". Le site est institutionnel, c'est un gage de qualité, une garantie que l'information contenue est validée.
- Le document fait figurer les mentions de responsabilité, c'est-à-dire l'auteur, l'organisme et si le document est daté, c'est également un signe de qualité. En général un site identifié par les URL citées plus haut est accompagné de ce type de mentions.
- Enfin, si le document se réfère à des sources ou des auteurs faisant autorité et s'il fournit une bibliographie, nous avons aussi affaire à un document digne de foi.

Voilà les règles de recherche fixées, nous pouvons commencer la recherche.

4.1.1. Les équations de recherches sur Google

N°	EQUATION	RESULTATS
1	"information gateway"	314 000
2	"subject gateway"	16 900
3	"subject based information gateway"	377
4	SBIG*	187 000
5	"quality controlled subject gateway"	209
6	"quality controlled subject gateway" or "information gateway" or "subject based information gateway"	19
7	"information gateway" or "subject gateway" or SBIG* or	80

[23] <http://www.google.com>

	"information gateway"	
8	"information gateway" or "subject gateway" or SBIG* or "information gateway" and archiv*	2
9	"quality controlled subject gateway" or "subject based information gateway" and archiv*	3
10	"information gateway" or "subject gateway" or SBIG* or "information gateway" and metadata	70
11	"quality controlled subject gateway" or "subject based information gateway" and metadata	18
12	"web archiving" or "internet archiving"	28
13	web or internet archiv* or preserv*	99
14	"site* federateur*"	289
15	"site federateur"	4270
16	"site federateur" and archiv*	8
17	Métadonnées	37500
18	Metadonnees	36500
19	Métadonnées and archive	9
20	Métadonnées and "site federateur"	7
21	"web harvesting" and metadata and gateway*	69
22	ECDL and digital	36100
23	“European Conference on Resarch on advanced Technology for Digital Librairies”	4240
24	Metadata SBIG	365

4.1.2. Commentaires sur les résultats obtenus : un bilan mitigé

Google offre un grand nombre de résultats, mais il génère surtout beaucoup de bruit. L’outil n’est pas forcément adéquat pour réaliser une recherche d’une telle ampleur, Google est un moteur index qui recherche une suite de caractère, il n’y a pas d’analyse sémantique ou linguistique du contenu de la requête.

Toutefois nous avons des raisons d’être satisfait des résultats fournis par certaines requêtes. En effet, les premières équations n°1 et n°2 et surtout les n°3 et 5 nous renvoient tout de suite vers des références institutionnelles malgré la masse de résultats (UKOLN[24], ou d’autres sites .ac.uk), des sites fédérateurs (Bubl,

[24] *United Kingdom Office for Library and Information.*

NISS[25], RDN[26], etc.) ou des projets internationaux (Renardus, DESIRE[27], NEDLIB[28], Nordic Web Archiv).

A contrario, des requêtes de type SBIG (n°4) ou les requêtes en français sur les sites fédérateurs ne donnent pas satisfaction. La notion de site fédérateur est encore perçue d'un point de vue commercial, voire corporatif, donc hors de notre propos. Pour les SBIGs, il suffit de se rappeler notre mésaventure sur Dialog, le web est aussi une tribune pour les passionnés d'astronomie… Ici, il y a un problème lié à la polysémie et à l'homonymie, Google peut fournir de bons résultats qui sont accompagnés par beaucoup de bruit, un bruit qui peut nous faire passer à coté de références intéressantes. Néanmoins, lorsque nous avons associé le terme SBIG avec le terme de métadonnée (équation n°24), nous avons obtenu un résultat tout à fait honorable avec de nombreuses références sur les projets internationaux tels que ceux de l'UKOLN, Renardus, des pages de l'IFLA[29], etc.

Les équations du n°8 au n°12 donnent d'assez bons résultats dont certains relatifs aux conférences de l'ECDL (n°12), sauf qu'il y a redondance dans les références. Ce sont souvent les mêmes projets qui sont en tête de liste. Ici, on entrevoit un des problèmes posé par l'algorithme de popularité Page Rank qui est celui de la récence des ressources publiées. En effet, ce sont logiquement les références les plus anciennes qui sont les plus populaires et il faut un certain temps à un article, une page ou un site nouveau pour qu'il soit bien classé. Google permet de localiser des travaux de références d'un certain âge, mais on ne peut pas trouver en tête de classement un travail récent et référentiel.

Enfin, signalons les équations n°21 et n°22 qui ont permis d'obtenir des résultats sur la collecte des sites dans le cadre d'une politique d'archivage. La requête n°22 avait pour but de localiser des pages sur la conférence de l'ECDL, elle fait paraître en tête les travaux de la Bnf et notamment ceux de C. Lupovici.

Malgré tout le bilan est assez mitigé car on recense toutefois énormément de ressources intéressantes pour notre propos où ne figurent aucunes mentions de

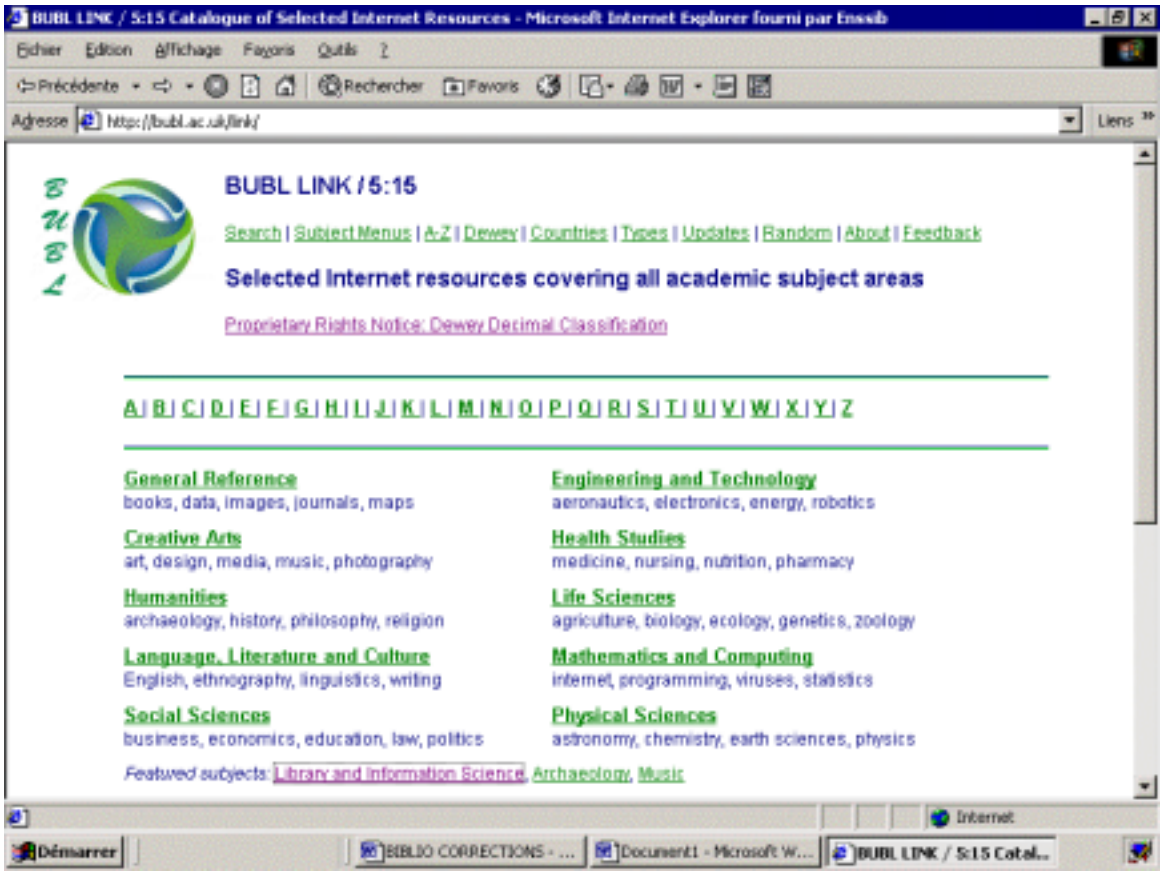
---

[25] *National Information Services and System*.
[26] *Resource Discovery Network*.
[27] *Development of a European Service for Information on Research and Education*.
[28] *Networked European Deposit Library*.

responsabilité, aucune description, ce qui les rend *de facto* pratiquement inexploitables.

4.2. Le site fédérateur Bubl

Ce site dont l'originalité est d'utiliser la classification Dewey, regroupe plus de 11000 ressources validées. Nous pouvons interroger directement l'index, mais il était plus facile de se laisser guider par les catégories. Deux nous intéressaient plus particulièrement : La catégorie *Library and Information Science* et celle relative à Internet.



Voici les rubriques jugées les plus pertinentes pour notre sujet :

[29] *International Federation of Library Association.*

Internet-Regulation-Standard ; *Internet Ressource Classification* ; *Internet Ressource Evaluation* ; *Library and Information Science Research* ; *Library And Information Science News* ; *Information Retrieval*.[30]

4.3. Le Consortium World Wide Web (W3C)[31]

Il s'agit du site de référence des spécialistes du web, des chercheurs du monde entier collabore dans le cadre des travaux du W3C qui se réunit régulièrement lors de congrès. C'est ici qu'on décide des nouvelles normes, des règles qui vont régir le web par l'application de nouveaux langages, de scripts, etc.

Nous avons décidé d'interroger son moteur de recherche pour éventuellement trouver des informations sur les sites fédérateurs, sur l'archivage et sur les métadonnées. Concernant les deux premiers thèmes, nous n'avons rien obtenu. Par contre la recherche sur les métadonnées a été plus fructueuses puisque le W3C recense toutes les normes relatives aux métadonnées.

4.4. Les listes de discussions

Bien que nous n'ayons pas couvert ce domaine, il était essentiel de signaler qu'il existe des listes de discussion relatives aux archivistes du web.

La première est celle du site Imesh (inscription sur imesh@mailbase.ac.uk). C'est un véritable réseau de coopération internationale qui regroupe la communauté des acteurs-clés engagés dans le développement des sites fédérateurs[32].

Une seconde est l'initiative du département de la BnF qui est chargé de l'archivage des sites web (web-archive@cru.fr). Ces listes peuvent donc permettre de connaître les principaux spécialistes de l'archivage des sites web, d'être tenu au courant des dernières avancées et du calendrier des conférences qui peuvent se tenir.

---

[30] Notons que Bubl a pu être localisé grâce au site fédérateur de l'ENSSIB, SIBEL (Sciences de l'Information et Bibliothèque en Ligne). <http://sibel.enssib.fr>

[31] <http://www.w3c.org>

[32] Imesh fut créée à l'issue de la Seconde Conférence Européenne sur la Recherche et la Technologie des Bibliothèques Numériques (Crète, ECDl 1998)

5. Conclusion et coût

Notre recherche s'est faite de façon progressive, nous sommes partis des outils qui ont offerts le moins de ressources, c'est-à-dire, les bibliothèques, les bases de thèses pour ensuite aborder des outils qui ont donnés beaucoup de résultats, mais qui ont nécessité un important travail de tri. Dialog, LISA et Emerald ont apporté une solide base bibliographique, mais peu d'articles traitent en fait de l'archivage des sites fédérateurs. Les articles retenus à l'issue de la recherche font la présentation d'un ou plusieurs sites fédérateurs, de projets nationaux ou de collaborations internationales relatives à l'homogénéisation du référencement et du catalogage de ces sites. Concernant la question de l'archivage, on a surtout retrouvé des articles relatifs à la conservation des sites web en général, avec la question délicate de l'archivage du Web Invisible. L'archivage des sites fédérateur est quand même abordé dans le cas d'archivage de sites thématiques qui va elle-même créer une base de signet vers des sites représentatifs de ces thèmes.

La partie consacrée à la synthèse montrera que peu de choses diffèrent entre l'archivage du web classique et celui des sites fédérateurs.

Temps et coût de la recherche :

TACHE	TEMPS	COUT
Travail sur Internet (Google, navigation)	70h.	/
Interrogation de Dialog	3h.	70 Eur.
Interrogation des autres bases	3h.	/
Lecture des articles	48h.	/
Rédaction du rapport	32h.	/
Heures travaillées[33]	**156**	**1833 Eur.**
TOTAL général		**1903 Eur.**

[33] Rémunération calculée sur la base du salaire d'un ingénieur d'études et de recherches en début de carrière soit 11.75 Euros de l'heure.

Partie 2 : Synthèse

Tout d'abord, nous allons faire un bref récapitulatif sur le développement des ressources en ligne et du Web Invisible, puis essayer de donner une définition et une typologie des sites fédérateurs. Ensuite, cette synthèse va essentiellement s'articuler autour de quatre entrées principales. Une partie va aborder les différentes méthodes d'archivage, une seconde va se pencher sur les critères de sélection des sites qui vont être aspirés dans l'archive et la périodicité de visite de ces sites. Un troisième point s'intéressera au référencement des sites fédérateurs. Enfin, nous nous pencherons sur les diverses contraintes techniques, budgétaires et juridiques liées à l'archivage de ces sites.

1. Introduction : Mise en contexte

1.1. Le développement des ressources en ligne

Avant d'entrer au cœur du sujet, quelques chiffres sur l'évolution du Web sont essentiels pour comprendre que les ressources numériques du Web représentent un volume énorme en constante croissance.

L'étude de S. Lawrence et de C. Lee Giles estimait la taille du Web en 1999 à 800 millions de pages HTML[34]. Depuis, il est passé à environ 4 milliards de pages, mais ce chiffre représente une faible proportion du Web global puisque comme le rappelle de nombreux spécialistes tels J.P. Lardy ou P. Lyman les moteurs de recherche classiques n'indexeraient que 4 milliards de pages, faisant fi des 550 milliards de documents composant le Deep Web (v. 1.2)[35].

---

[34] Lawrence S., Lee Giles C., *Accesibility of information on the Web*. Nature, 1999, n°400, p.107-109.

[35] J.P. Lardy, *Le Web Invisible*, *compte rendu à la commission de l'ADDNB*, mars 2002 rappelle l'évolution : 800 millions de pages pour Lawrence et Lee Giles juillet 1999, 1 milliard janvier 2000 pour Inktomi et le NEC research Institute, 2 milliards en juillet 2000 selon Cyveillance et 500 milliards de documents d'après l'étude de Bright Planet.

Selon P. Lyman, le Web connaît une croissance journalière de 7 millions de pages, mais *a contrario*, il perdrait aussi une masse très importante de documents[36].

1.2. Le Web Invisible ou Deep Web

Pour bien comprendre notre sujet d'étude, il est nécessaire de se pencher sur ce qu'est le Web Invisible. Il s'agit d'une expression récente développée dès 1994 par J. Ellsworth pour désigner les informations qui sont inaccessibles aux moteurs de recherche traditionnels[37]. Les robots ou *crawlers* n'ont pas accès à ces sites ou à ces pages parce qu'elles peuvent être protégées par un mot de passe, contenir des balises d'exclusion des robots. Il s'agit surtout d'informations contenues dans des bases de données : Ce sont donc des pages qui n'existent pas sur un serveur, elles sont créées à la demande d'un client par le biais d'un formulaire. Il s'agit ici de pages ou de sites dynamiques.

Il s 'avère que l'information contenue dans le Deep Web est organisée et dont l'indexation est d'une grande qualité[38]. Dans le cadre de notre problématique, le Web Invisible ne va pas manquer de nous intéresser puisqu'une écrasante majorité des sites fédérateurs est comprise dans cette partie du Web.

1.3. La masse de l'information et le problème de l'auto-édition

Le Web propose de l'information accessible en masse, toute la difficulté est de retrouver de l'information pertinente. De plus, Internet est un ensemble de réseaux interconnectés très faiblement centralisé où les ressources documentaires disponibles ne sont pas organisées. O. Larouk résume bien ce problème en disant qu' "*il n'existe pas de base documentaire centralisée affectée à l'organisation ou à la recherche de ces ressources*"[39].

---

[36] Lyman P., *Archiving the World Wide Web. In Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving*, 2002, p.39-51.

[37] Ouf R., *Le dynamisme du Worl Wide Web : taille, croissance, visibilité, distribution et accessibilité de l'information*. Rapport de recherche bibliographique, Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques (ENSSIB), Villeurbanne, 2001, 58p.

[38] *Ibid..,* p.34.

[39] Larouk O., *op.cit.*, p.235-251.

Enfin, une des principales conséquences du Web est l'évolution des logiques éditoriales. Le web simplifie la diffusion de documents pour un moindre coût avec les aléas que cela implique. On y trouve le meilleur comme le pire puisqu'en dehors des sites fédérateurs, il y a rarement évaluation ou de filtrage de l'information.

1.4. Des sites fédérateurs pour de l'information de qualité

Les sites dits "fédérateurs" représentent l'alternative a ces problèmes d'explosion de l'information. Ils sont crées par des spécialistes d'un domaine particulier. De part leur nature qui leur donne une logique bibliothéconomique, il n'est pas surprenant de constater que les bibliothécaires et les documentalistes furent les précurseurs de la mise en place de ces passerelles d'informations où le contenu est structuré avec des liens valides vers les sites les plus intéressants.

Selon la description de O. Larouk (entre autres), ces sites centraliseraient les accès par discipline, zone géographique, langue, etc.

2. Définition et typologie des sites fédérateurs

Le Web est un vaste espace documentaire, ses sites ne sont pas tous de même nature. Avant d'aborder le cas des sites fédérateurs, un bref rappel technique s'impose sur les différents types de sites que l'on peut trouver sur le Web.

2.1. Les sites web statiques et dynamiques [40]

Pour être synthétique, le site statique est un site dont le contenu est stocké effectivement dans le serveur qui est chargé de diffuser l'information demandée par un utilisateur quand il tape l'URL de la page ou du site ou lorsqu'il clique sur un lien hypertexte qui pointe vers la page ou le site désiré.

[40] Pour plus de détails sur cette typologie, v. MasanEs J., *Préserver les contenus du Web* [en ligne]. Disponible sur <http://bibnum.bnf.fr/conservation/migration_web.pdf >

Les sites dynamiques utilisent des bases de données pour stocker tout ou partie de leur contenu. En suivant l'étude de J. Masanès , on peut observer deux grandes catégories de sites dynamiques :

La première consiste à stocker dans une base de données le contenu de pages. L'usager envoie sa requête, un script va générer les éléments demandés qui sont stockés dans la base de données et les afficher coté client. J. Masanès nomme ce type d'architecture "HTML éclaté" puisque les pages sont éclatées dans la base de données.

Deuxième catégorie : C'est ce que J. Masanès appelle les passerelles documentaires, ce sont les sites fédérateurs. Les bases de données servent alors soit à stocker directement l'information visée, soit à obtenir des informations qui vont permettre d'accéder aux documents. Cela ne veut pas dire que les sites fédérateurs sont des sites dynamiques, ils peuvent posséder des moteurs internes qui leur permettent de chercher dans une base de données.

2.2. Qu'est ce qu'un site fédérateur ?

Il s'agit d'une base de signets mise au point et validée par des spécialistes d'un ou plusieurs domaine en fonction de l'éventail de disciplines ou de thèmes que le site veut présenter. Le problème posé par cette notion de site fédérateur, c'est de comprendre ce dont il s'agit vraiment. Ces sites diffèrent bien sûr de par leur nature, mais surtout de par la valeur ajoutée quant à leur qualité informationnelle.

2.3. Typologie en fonction de la vocation du site

O. Larouk se fait précis sur la typologie ou plutôt la typologie des vocations de ces sites à travers cinq catégories :

- Les sites qui présentent l'institution (vitrine).
- Les sites utilitaires et portails exhaustifs de connaissances pour les universités et les organismes de formation.
- Les sites offrants des ‘‘services documentaires’’ qui fournissent un outil de recherche adapté aux besoins du lecteur.

- Les sites orientés sur la fusion de compétences qui regrouperaient plusieurs institutions couvrant plusieurs domaines afin de permettre des recherches globales[41].
- Les sites thématiques limités sans aucune classification qui représentent les pôles d'intérêt d'un ou plusieurs organismes ou de passionnés d'un même sujet. Cette dernière catégorie nous intéresse plus particulièrement car beaucoup de projets d'archivage tournent autour de ces sites thématiques.

Toujours selon l'article de O. Larouk, ces catégories contiendraient deux types de sites : Les sites fédérateurs généralistes du type Bubl qui donnent accès à de nombreux sujets évalués et sélectionnés et les sites spécialisés par domaine comme SOSIG (*Social Science Information Gateway*) pour les sciences sociales ou EEVL (*Edinburgh Engineering Virtual Library*) pour l'ingénierie.

2.4. Typologie en fonction de la qualité du contenu

Beaucoup de chercheurs ont proposé une typologie de ces sites. Les termes de "*Subject Gateway*", "*Information Gateway*", "*Academic Subject Gateway*", "*Subject-Based Information Gateway*" ou "*SBIGs*" désignent bien sûr une même entité, mais le critère de qualité n'est pas explicitement cité dans ces dénominations.

Selon l'analyse de S. Saidi le terme le plus adéquat pour parler d'un site fédérateur dont le contenu est strictement contrôlé et validé par des experts serait "*quality controlled subject gateway*" ("site fédérateur de qualité contrôlée"), terminologie développée par T. Koch[42]. Le site fédérateur de qualité contrôlée met en œuvre un ensemble de règles de qualité pour rendre la découverte systématique de ressources possibles. Un important travail humain est demandé afin de garantir le choix judicieux des ressources qui vont être publiées mais aussi pour décrire de façon très précise les documents grâce à l'utilisation de métadonnées.

---

[41] O. Larouk, *op.cit*.
[42] Koch T., *Quality-controlled subject gateways: definitions, typologies, empirical overview*. Online Information Review, 2000, vol.24, n°1, p.24-34.

Régulièrement mis à jour, ces sites doivent permettre aux usagers de trouver facilement une information pertinente grâce à une indexation de haut niveau avec l'utilisation d'un vocabulaire contrôlé, une structure de classification profonde pour permettre la recherche avancée et la navigation.

Ce qui importe dans la définition de T. Koch, c'est l'implication d'un travail humain, comme le montre la définition d'O. Larouk :

"*Il existe des sites fédérateurs assimilés à des passerelles d'information spécialisés par sujet ou SBIG. Leur lancement est l'œuvre de bibliothécaires et/ou d'informaticiens-documentalistes qui souhaitent ''corriger'' certaines insuffisances des services Web. La principale caractéristique des concepteurs est d'y appliquer les méthodes issues de la bibliothéconomie et des sciences de l'information. Ces démarches nécessitent des indexeurs ou modérateurs de la documentation pour valider le contenu informationnel* "[43].

2.5. La responsabilité de mémoire : vers la conservation et l'archivage des sites web

L'enjeu fondamental entraîné par cette diffusion de savoir et de culture va être de préserver ces informations puisqu'une très grande partie de ces ressources est définitivement perdue. Cette prise de conscience est nette puisque au long des lectures effectuées dans le cadre de ce rapport on constate que les spécialistes sont à l'unisson et ont commencé à œuvrer dès 1996 pour préserver le contenu du Web et bien sûr plus particulièrement celui des sites fédérateurs dont le contenu ne peut pas se permettre d'être perdu[44].

Cette conservation pose des problèmes, notamment pour la pérennisation des archives de sites web, pour les sites qui forment le Deep Web, des sites qui se généralisent de plus en plus.

Préserver un site, c'est préserver un ensemble de fichier, c'est aussi préserver la structure de cet ensemble et de son mode de fonctionnement, notamment la navigation qui est le mode naturel d'accès à ce type de contenu. Nous verrons que

[43] Larouk O., *op.cit.*
[44] Outre les travaux de M. Haettuger et J.Masanès, signalons aussi ceux de P. Hallgrimsson et S. Bang sur les archives nordiques, de B. Reilly sur les archives à caractère politique, etc. (pour les références complète, cf. bibliographie)

la tâche est difficile pour les architectures techniques complexes comme les sites dynamiques.

3. L'archivage des sites web et des sites fédérateurs

3.1. Démarche initiale avant de commencer un projet d'archivage

3.1.1. Le site web : un objet complexe face à l'obsolescence technique

En qualité d'objet multimédia, un site web peut contenir du texte, de l'image, du son, de l'animation, des fichiers aux formats divers ainsi que des fonctionnalités. C'est un ensemble d'éléments structurés par des liens hypertextes. M. Haettiger rappelle l'étude de P. Lyman selon laquelle une page web contiendrait en moyenne une quinzaine de liens et cinq objets différents[45].

On entrevoit ici un problème, celui des formats de fichiers face à l'évolution technologique, autrement dit pourra t-on par exemple lire un fichier PDF ou JPEG dans dix ans ?

Pour les pages dynamiques, il faut fixer des règles pour savoir quelle est la partie qu'il faut considérer comme originelle et donc à conserver.

Un autre problème est relatif aux navigateurs : Netscape, Internet Explorer, Mozilla, Opera ne présentent pas un site de la même manière. L'accès à la forme originelle n'est pas garanti.

3.1.2. Que doit-on archiver ?

La question est de savoir à partir de quel niveau faut-il commencer à archiver un site web. Conserve t-on la présentation du document, l'ensemble des fonctionnalités du document ? Bien sûr, l'archiviste voudrait archiver l'ensemble de ces éléments, mais il faut être prudent : Si l'on veut intégrer toutes les caractéristiques d'un document, les conditions de conservation deviennent alors de

plus en plus complexes parce que cela demande l'intégration de formats, de logiciels et de périphériques différents[46]. On peut aussi se poser la question de l'archivage des sites qui ont un lien sur le site à archiver.

3.1.3. Authentification des sites

Le problème ne se pose pas pour les sites fédérateurs, mais il est bon de rappeler que beaucoup de sites ont des lacunes en terme d'authentification et d'identification. Les mentions de responsabilités (auteur, créateur, date de mise à jour, etc.) ne sont pas toujours explicitement publiées.

M. Haettiger remarque que les sites n'ont pas d'identificateur unique comme l'ISBN, ce qui pose un problème dans le cadre d'une politique de conservation. L'identificateur unique permettrait de mieux gérer les différentes versions d'un même site. On pourrait s'assurer qu'une version enregistrée à un instant T est bien la suite d'une autre version enregistrée dans un autre établissement[47].

3.2. Créer une archive web : une politique d'acquisition

La constitution d'un fonds de sites web doit entrer dans une politique d'acquisition qu'on peut calquer sur celle effectuée pour n'importe quel type de document. Il faut définir des critères de pertinence comme la langue, le type de public visé, le sujet traité, le type de site, etc.

Ces critères peuvent poser des problèmes : Si on fait une politique d'acquisition de sites web français, on pose la question de savoir ce qu'est un site français : un serveur hébergé en France ? Un domaine **.fr** ? Un site francophone ? Un site d'une entreprise étrangère installée en France ? Quel est donc le degré de granularité qu'il faut choisir en sachant que la taille du web est une contrainte pour repérer les sites les plus pertinents ?

Il existe donc deux méthodes pour sélectionner des sites dans le but de créer une archive, une manuelle et une autre automatique.

---

[45] Haettiger M., *op.cit.*, p.14.
[46] Haettiger M., *op.cit.*, p.15.
[47] Haettiger M., *op.cit.*, p.18.

3.2.1. La sélection manuelle.

Méthode choisie notamment par les Bibliothèques Nationales australienne et canadienne, elle consiste à repérer, sélectionner et acquérir manuellement les sites à archiver. Pour le cas du Projet australien PANDORA[48], cette méthode permet de filtrer les sites créés par un Australien, portant sur l'Australie (sujet social, politique, économique, culturel, scientifique, religieux).

Cette solution présente des avantages : Le suivi du nombre de sites archivé et la prévision sur les coûts de l'archivage. C'est un suivi plus fin, on peut envisager de conserver un nombre plus important de mises à jour pour un même site. Le travail humain se révèle plus précis que celui du robot.

3.2.2. La sélection automatique ou *snapshot*

Cette solution est permise grâce à des logiciels appelés "collecteurs" ou "*harversters*". Ce sont des moteurs de recherche qui parcourent le Web et qui aspirent les sites repérés en fonction des critères de pertinence paramétrés dans le robot. La Bibliothèque Royale de Suède, la BnF mais aussi Internet Archive ont privilégié cette démarche.

Le produit obtenu par le robot est un *snapshot*, littéralement un "instantané" du Web ou d'une portion de celui-ci. C'est le meilleur moyen d'avoir une capture exhaustive du Web ou d'une de ses parties. Pourtant, cette méthode présente un certain nombre d'inconvénients :

3.2.3. Contraintes du snapshot

On a dit que le collecteur est paramétré selon un algorithme de pertinence. Lorsqu'elle s'appuie à l'image de Google sur la popularité, elle peut poser des problèmes dans le cadre d'un archivage : Par exemple, la Bnf a paramétré son robot de façon à ce qu'il récupère les sites les plus populaires.

---

[48] *Preserving and Accessing Networked Documentary Resources of Australia*

Avec un tel paramétrage, on risque de retrouver dans son archive toujours les mêmes sites. Les plus importants laissant les plus rares, les moins connus et peut être les plus pertinents *de facto* de coté.

Le second problème concerne l'accès aux ressources du *Deep Web*. Les bases de données en ligne, mais aussi les OPAC ne sont pas archivés puisque les robots sont stoppés. Un *snapshot* est donc inopérant sur le Web Invisible.

Une autre contrainte est qu'on ne peut en faire qu'un nombre limité de *snapshot* dans l'année, jamais plus de deux en général parce que le nombre de données qui sont aspirées est énorme et prend donc beaucoup de temps. Entre deux *snapshots*, de nombreux sites connaîtront d'importantes modifications, certains disparaîtront d'autres peuvent apparaître et disparaître entre les deux instantanés, sans qu'on puisse le savoir et donc les archiver.

Finalement, on s'aperçoit qu'un archivage performant et pertinent ne peut se réaliser que grâce à la combinaison des deux méthodes, la méthode manuelle permettant de mieux juger la ressource mais surtout étant capable de repérer les sites du Web Invisible où les robots sont stoppés, donc des informations à forte valeur ajoutée.

3.3. L'archivage et la conservation des sites fédérateurs[49]

Si l'on veut assurer la conservation d'un site fédérateur, il est nécessaire de conserver les moyens d'accéder facilement à la masse de documents à laquelle il renvoie. On ne peut se contenter d'archiver uniquement le formulaire et les documents, il faut conserver la passerelle documentaire dans son intégralité.

J. Masanès cite une expérimentation réalisée à la BnF qui consiste à faire migrer ces passerelles vers un format, le XML, qui tout en étant pérenne, permet des fonctionnalités de recherche d'information. L'utilisation d'une archive en

[49] L'archivage des sites statiques ne pose pas vraiment de problème, par contre le cas des sites dynamiques se révèle beaucoup plus complexes et fait l'objet aujourd'hui de nombreux travaux. Masanes J., *Towards continuous Web archiving*. D-Lib Magazine [en ligne]. 2002, vol.8, n°12. Disponible sur : <http://www.dlib.org/dlib/december02/masanes/12masanes.html>

format semi-structuré (XML) permet de garantir un accès futur à cette information[50].

4. Etude de cas sur les critères de sélection et la périodicité

Nous avons soulevé précédemment la question des critères de sélection pour constituer une archive web, nous allons ici entrer davantage dans le détail en se penchant sur plusieurs projets internationaux.

4.1. Le cas des archives thématiques

Une archive thématique peut être le fruit du travail d'un groupe de spécialistes, de passionnés ou d'une institution qui s'intéresse à un domaine précis. Il peut aussi résulter d'un événement majeur susceptible d'avoir une portée historique ou représenter un suivi de l'activité politique d'un pays ou d'une région du globe.

Dans le cadre de ce rapport, nos lectures nous ont fait découvrir les projets d'archives thématiques de la Bibliothèque du Congrès sur les attentats du 11 septembre 2001 et sur les élections fédérales américaines de 2002[51], mais aussi sur la création d'une archive sur les communications politiques de groupes non-officiels autour de quatre grandes régions du globe.

4.2. Les archives relatives aux 11 septembre et aux élections de 2002 aux Etats-Unis[52]

4.2.1. Les critères de sélection

La Bibliothèque du Congrès a chargé Internet Archive de collecter chaque jour les URL relatives aux attentats et à leurs conséquences : Le personnel de la Bibliothèque du Congrès, les chercheurs de Webarchivist.org et les internautes du

[50] Masanès J., Préserver les contenus du web, *op.cit.*.
[51] La création de cette archive s'inscrit dans le projet MINERVA (*Mapping The Internet Electronic Ressources Virtual Archive*).
[52] Schneider S.M., Foot K., Kimpton M., Jones G., *Building Thematic Web Collections: Challenges and Experiences from the September 11 web Archive and the Election 2002 Web Archive*. In : Koch T., Sølvberg I.T. Eds. Research and advanced technology for digital libraries : 7th European conference, ECDL 2003, Trondheim, Norway, August 17-22, 2003, p.77-93.

monde entier ont soumis des URL pour constituer la collection. La collecte des sites des élections de 2002 a obéi à la même logique.

Pour les décrire les archives des élections fédérales, Webarchivist et Internet Archive ont développé leur type de métadonnées[53] (candidat, parti, gouvernement, etc.). On insiste sur le critère qualité des sites recueillis. Les URL capturées sont analysées par Webarchivist puis envoyée à Internet Archive. Elles sont *crawlées* sur une périodicité spécifiée, de même que la profondeur des niveaux visités est contrôlée également.

4.2.2. Identification des sites pertinents

L'identification se fait par l'étude des pages d'accueil des sites cibles. Une fois identifiée, l'URL doit être évaluée du point de vue technique, pour s'assurer que chacune représente bien un point de départ idéal pour pouvoir exploiter les liens qu'elle contient. Une fois l'URL identifiée et sélectionnée, il faut spécifier la profondeur interne et externe d'exploration.

4.2.3. Périodicité, vérification de la qualité réelle du site capturé et autorisation à la diffusion

Il n'est pas encore possible de détecter précisément la fréquence de changement d'un site ou d'une de ses pages en utilisant les techniques de *crawling*. Il faut donc visiter les pages d'un site régulièrement, en général deux fois par an. Des vérifications automatiques et manuelles sont effectuées pour s'assurer de la précision de l'indexation et que les pages archivées peuvent être correctement reconstituées pour être affichée dans l'archive. Enfin, La Bibliothèque du Congrès sollicite l'autorisation de publication au responsable du site.

4.3. L'archivage des sites de communications politiques[54]

[53] Voir § 5 sur le grand nombre de standards de référencement.
[54] Reilly B. et al., *Political Communications Web Archiving : Adressing Typology and Timing for Selection, Preservation and Access*, In : Koch T., Sølvberg I.T. Eds. Research and advanced technology for digital libraries : 7th European conference, ECDL 2003, Trondheim, Norway, August 17-22, 2003, p.94-102.

Projet initié par le *Center for Research Library*, son objectif est d'archiver les sites à caractère politique non-officiels de l'Asie du Sud-Est, l'Europe Occidentale, l'Afrique Subsaharienne et l'Amérique latine. Officiellement le but est de préserver des communications qui sont l'équivalent numérique des affiches, des pamphlets et toute sorte de littérature de rue, une littérature qui va survivre grâce à cette archive et pouvoir être analysée d'un point de vue historique, sociologique dans les années à venir.

4.3.1. Identification et sélection

Sont concernés dans cette archive des sites web statiques et des documents de formats divers, des groupes de news également. Beaucoup de ces sites sont instables et selon les régimes politiques, ils peuvent être hébergés hors de leur pays d'origine. Les spécialistes des quatre zones se sont entendus pour faire un travail de veille pour analyser les sites émergeants puis éventuellement les capturer.

4.3.2. Périodicité

La fréquence de capture est importante pour deux raisons : La durée de vie moyenne d'une page web est estimée à 44 jours. Ensuite, les mises à jour du contenu de quelques sites cibles peuvent évoluer un grand nombre de fois en un laps de temps très court alors que d'autres sites peuvent rester inchangés pendant des mois, voire plus. Les recherches sur les types de sites qui doivent être inclus dans l'archive ont révélé que la fréquence de mise à jour du contenu des sites varie largement et dépend de nombreux facteurs, dont la nature du groupe politique lui-même ou de l'événement sur lequel le site est concentré[55].

55 Sur l'instabilité de ces sites, notons deux exercices menés par l'équipe du projet : Le premier est conduit fin 2002 et concerne les sites du LANIC (*Latin American Network Information Center*), un répertoire de liens des sites d'élection en Amérique Latine. 148 URL de l'observatoire sont recherchées à travers la Wayback Machine d'Internet Archive. Ces URL correspondent à un total de 21 processus électoraux de 15 pays entre décembre 1998 et juin 2002. 61% de ces 148 sites ont déjà disparu du web au moment du début de l'exercice. 19 sites n'étaient pas dans l'Internet Archive. Pour les 129 restants, 36% n'incluent plus de documents sur la période de l'élection elle-même. Au printemps 2003, on se penche sur 38 archives de sites visités par un robot sur le processus électoral au Nigeria. On a un résultat similaire que celui observé pour l'Amérique latine.

Les chercheurs du projet ont proposé deux niveaux de périodicité : Une approche automatique et une autre sélective.

Un groupe de sites conservé et identifié est *crawlé* selon une durée spécifique (ex. deux fois par an). Il s'agit ici de sites dont le contenu connaît une certaine stabilité.

Un autre groupe de sites doit être identifié par l'archiviste sur la base "d'événements chocs" (coup d'état, guerre, dissolution d'assemblée, etc.) En conséquence, la fréquence de visite est régulière en fonction de l'importance de l'événement (ex. : quatre fois par jour pour les deux dernières semaines, chaque jour pendant deux mois, etc.)

5. Le référencement des sites : une pléthore de standards

La description d'un site consiste à produire des métadonnées sur le document. Les métadonnées permettent d'identifier, de décrire, de localiser un document, c'est un élément essentiel dans le cadre d'une politique d'archivage. Bien utilisée, la métadonnée, est aussi un gage de qualité du site.

Il existe trois types de métadonnées qu'on doit retrouver dans n'importe quel document :

- Les métadonnées descriptives : Auteur, titre, date, création du document…
- Les métadonnées structurelles ou de conservation : Informations sur la structure informatique des sites web, arborescence des fichiers, langage, etc.
- Les métadonnées de gestion : Informations sur l'histoire du document, sur son enregistrement, le nombre de modifications subies (migrations), les formats successifs sur lesquels les fichiers ont été enregistrés. Informations aussi sur la gestion des droits et des modalités de consultation de la ressource.

Ces métadonnées sont adaptables, elles peuvent permettre de décrire n'importe quel site, elles sont minimales. Par contre, elles ne montrent pas des informations comme le nombre de pages, les types de documents inclus, l'arborescence du site, sa structure.

5.1. Principaux standards : Dublin Core, les DTD et RDF

Certaines métadonnées ne sont pas adaptées au référencement de sites web comme le format MARC, d'autres comme IAFA Templates manquent de souplesse[56].

5.1.1. Le Dublin Core, le modèle le plus stabilisé

Standard créé par la NSCA (*National Center for Supercomputing Application*) et l'OCLC (*Online Computer Library Center*) en 1995 il comprend 15 éléments[57], sous forme de balise, qui peuvent être encodés en HTML ou XML.

Sa simplicité est son principal avantage. Par ailleurs, il existe des outils pour générer un référencement en Dublin Core des sites web à partir de recherche de mots effectuée dans le texte du site. Cela permet de rechercher des documents complexes (la Suède utilise le Dublin Core pour décrire les sites qu'elle archive).

Mais il faut être prudent, car dans le cadre de l'archivage de sites web, le Dublin Core ne renferme pas toutes les informations, il peut y avoir confusion par exemple entre l'auteur et le créateur du site.

5.1.2. XML et les DTD

XML est un langage à balises créé par le Consortium W3C supposé supplanter HTML. A la différence d'HTML, XML distingue un contenu à une présentation, ce qui permet à un même document d'appliquer de nombreuses présentations. XML intéresse les chercheurs car il pourrait permettre d'optimiser les recherches sur Internet en développant un véritable Web sémantique. Enfin,

[56] Saidi S., *op.cit.*, p.29-30.
[57] Les éléments sont : auteur, titre, sujet des mots-clés, description, éditeur, date, autre contributeur, identificateur, support, relation, langue, source, couverture, localisation spatiale et temporelle, droit auteur.

XML permet la création de DTD (Document Type Definition) qui définissent la structure et la liste d'éléments utilisés dans un document électronique[58].

Il existe beaucoup de DTD mises au point par les diverses équipes qui s'occupent d'archivage, mais signalons parmi les plus utilisées dans l'archivage des sites fédérateurs les standards EAD (Encoded Archival description) et TEI (Text Encoded initiative)[59]. La Bibliothèque du Congrès, pour cataloguer les sites du 11 septembre et l'archive sur les élections de 2002, a crée son propre standard, MODS[60] dérivé d'un autre standard, METS (*Metadata Encoding and Transmission Standard*).

En théorie, XML est un outil souple et adaptable : Avec une DTD particulière il est *a priori* adapté au référencement de sites web archivés. En fait, la qualité de XML est aussi son principal défaut, il y a énormément de DTD qui se sont développées. Chaque institution, chaque créateur de site peut créer sa DTD et cela pose un véritable problème, car l'archivage des sites web ne peut être efficace que si XML et les DTD se standardisent.

5.1.3. RDF (Ressource Description Framework)

Les modèles vus précédemment posent le problème de l'interopérabilité et de l'échange entre les différents types de description de ressources qui utilisent des applications différentes. L'alternative proposée par le Consortium W3C est RDF. Standard évolutif, il est également un métalangage à l'image de XML. D'ailleurs la syntaxe de RDF est inspirée du XML qui est lui-même sémantisé par RDF. Le schéma de ce standard s'articule autour d'un sujet, d'un prédicat et d'un objet, autrement dit une ressource, une propriété (ou action) et une valeur.

Il peut intégrer les autres standards comme EAD, MARC, DC. S. Saidi rappelle un article de référence sur RDF de A. De Robbio qui ne tarit pas d'éloges

---

[58] Une DTD peut être interne ou externe, pour les détails v. le site du W3C : *Introduction to DTD*. Disponible sur : <http://www.w3schools.com/dtd/dtd_intro.asp>

[59] Sur EAD, v. Saidi S., *op.cit.*, p.31 et Library of Congress, *Encoded archival description (EAD) Official EAD Version 2002 Web Site* [en ligne]. Disponible sur: <http://www.loc.gov/ead/>

[60] *Metadata Object Description Schema*. V. Library of Congress, *MODS Standard* [en ligne]. Disponible sur: <http://www.loc.gov/standards/mods/>

sur ses avantages[61]. Pour elle, RDF permet la description des ressources web pour faciliter l'élaboration automatique des informations et peut être utilisé dans divers domaines d'application tels que la recherche de ressource, le catalogage ou l'évaluation du contenu.

RDF est pour l'instant le standard le plus adapté au travail d'archivage de sites web… en attendant qu'une nouvelle norme apparaisse.

5.2. Les métadonnées dans les sites fédérateurs

Un bon usage des métadonnées est une condition *sine qua non* pour garantir la qualité et *a fortiori* l'intérêt d'un site fédérateur. Ce qui distingue les sites fédérateurs des outils de recherche classiques, c'est le travail manuel de description des ressources, plus précis qu'un travail automatique. On peut ainsi décrire de façon minutieuse des ressources sélectionnées. Les usagers pourront juger de la pertinence d'une source avant d'aller la consulter et les indexeurs peuvent savoir à quelle date une ressource doit être revue voire être retirée de la base de données.

On peut permettre à des moteurs internes au site fédérateur d'effectuer des recherches pertinentes grâce à une indexation précise. Enfin, l'échange de ressources ou de descriptions de ressources entre les sites fédérateurs est possible, favorisant ainsi les flux d'informations, la collaboration et la recherche sur plusieurs sites à la fois. Dans le cadre des sites fédérateurs, c'est l'interopérabilité entre les métadonnées qui est recherchée.

5.3. Nécessité d'une norme stable et d'une identification unique des sites

Etablir un catalogage ou un référencement des sites web et des sites fédérateurs se révèle finalement assez périlleux. Il existe beaucoup trop de normes, même si le modèle Dublin Core est le plus stabilisé et que RDF se présente comme un standard d'avenir.

---

[61] S. Saidi, *op.cit.*, p.32 qui cite ici A. De Robbio, *Metadati : Parola chiave per l'accesso alla biblioteca ibrida*. In La biblioteca ibrida : verso un servizio informativo integrato. Milan, 2002.

A défaut de normes stables sur les métadonnées, il existe une norme sur l'organisation générale des archives électroniques : L'OAIS (*Open Archival Initiative System*). Cette norme décrit toutes les étapes constitutives au traitement des données à archiver[62].

Pour conclure sur les standards, soulignons également un problème parallèle, celui d'une identification fiable des sites web. Dans le cadre de l'archivage des sites, il faut travailler sur des identificateurs uniques, persistants, donc fiables. L'URL (*Uniform Ressource Locator*) n'est pas un identifiant persistant puisqu'un site quelconque peut changer plusieurs fois d'URL[63].

6. Maintenance et contraintes de l'archivage

6.1. Stockage physique et prévention

Les supports de stockage utilisés sont en général des supports classiques, c'est-à-dire des disques durs IDE et quelques fois des bandes magnétiques de type DLT. Pour le cas des disques durs, il faut choisir des disques à forte capacité de stockage pour ne pas être pris au dépourvu rapidement : Faire une archive du web, cela veut dire que cette archive peut croître très rapidement. Les archives sont copiées et stockées dans des endroits différents pour éviter des pertes si une machine venait à être endommagée ou si le bâtiment abritant l'archive serait détruit.

Il faut aussi se prémunir contre les risques de pertes des informations archivées. Nombreux sont les aléas qui peuvent se produire au cours de l'archivage : Problèmes de réseaux, pannes, virus, malversations ou détournement des données collectées, saturation du disque dur, etc.

6.2. Solutions de conservation face à l'évolution technique

[62] Pour les détails, v. Haettiger, *op.cit.*, p.47-49.
[63] M. Haettiger rappelle les principales solutions qui vont dans ce sens : l'URN (*Uniform Ressource Name*), le PURL (Persistant URL) et le DOI (*Digital Object Identifier*)Pour les détails, v. Haettiger, *op.cit.*, p.41-42. Certains projets comme PANDORA ont développé leur propre système d'identification.

M. Haettiger résume bien les différentes possibilités de conservation de ces sites une fois aspirés. Une solution s'est imposée, la migration et une autre plus intéressante n'en est qu'au stade de la recherche, il s'agit de l'émulation[64].

6.2.1. La migration

Le processus consiste à transformer le document à conserver en suivant l'évolution des techniques, c'est un système déjà connu du monde des bibliothèques. Deux types de migrations existent : La première consiste à changer le support physique de stockage (Disquette vers Cdrom, Cdrom vers DVDrom, etc). La seconde porte sur la modification du format ou du codage des données du document (document Word98 vers WordXP, de ASCII à UNICODE).

La migration est en ce moment la seule solution qui permet une conservation à long terme des sites, elle est adoptée dans la plupart des bibliothèques lancées dans ces projets. Nous verrons par la suite que cette solution pose un problème juridique puisque la migration modifie le document archivé et donc porte atteinte aux droits moraux d'auteur.

6.2.2. L'émulation

Un émulateur est un dispositif qui simule le comportement d'un autre système. Il serait donc possible d'accéder aux sites web dans leur forme originale par le biais d'un émulateur capable d'exécuter des instructions écrites dans un langage obsolète pour une machine obsolète. Les avantages sont évidents avec un tel système puisque le site n'est pas modifié et on a accès à toutes ses fonctionnalités. Cette solution est encore au stade de la recherche et son développement est long. Et comme les bibliothèques vont être les seuls clients, le prix de cette solution risque d'être prohibitif.

6.3. Contraintes budgétaires, contraintes pour le personnel

[64] Il est possible de les conserver de façon analogique, sur papier ou microfiche, mais également dans leur environnement technique d'origine, ce qui posera bien sûr des problèmes lorsque les pièces et les supports de rechange ne seront plus fabriqué.

L'archivage des sites web intéresse les bibliothécaires, premiers concernés quand il s'agit de conservation et de valorisation d'un document. Cependant, un site web n'est pas créé à l'origine pour être archivé. Si une politique d'acquisition de sites web obéit aux logiques bibliothéconomiques classique, son archivage est complexe. Les contraintes sont nombreuses pour le personnel des bibliothèques : Création de postes, achat de machines pour créer et valoriser le fonds archiver, formations en informatique, nécessité d'avoir un budget conséquent[65]…

Seules les grandes institutions peuvent se permettre de mener une politique d'archivage du web. Les coûts en matériels, logiciels[66] et en personnels (formations, recrutement de spécialistes) ne sont pas vraiment compatibles avec des budgets qui, comme en France, sont revus à la baisse.

6.4. Les problèmes d'ordre juridiques

En France, il n'y pas de loi sur le dépôt légal des sites web, il est donc difficile d'être informé sur ce qui est crée sur le web. Ensuite, les modifications effectuées sur le site, changement des formats lors d'une migration, insertion de métadonnées pour cataloguer les sites, sont autant d'opérations qui portent atteintes droits moraux d'auteur avec les risques pénaux que cela implique pour les archivistes.

Les sites consultables par mot de passe ou par accès payant posent également problème. Toutes ces contraintes peuvent nuire à la constitution et à la valorisation de ces archives. Enfin, l'archivage à l'échelle des projets de coopération internationale pose aussi un problème puisque la législation sur ce dépôt légal n'est pas harmonisée.

---

[65] Zabicka P., *Archivig the Czech Web: Issues and Challenges*. In : Koch T., Sølvberg I.T. Eds. Research and advanced technology for digital libraries : 7th European conference, ECDL 2003, Trondheim, Norway, August 17-22, 2003, p.111-117, montre bien dans son article les problèmes et les retards connu par le projet de son équipe à cause d'un budget restreint.

[66] Par exemple, le coût d'un collecteur est élevé puisque le logiciel est crée sur mesure.

Conclusion

La partie méthodologique a montré quels sont les outils les plus adéquats pour mener une recherche documentaire sur l'archivage du web et des sites fédérateurs. Compte tenu de la nouveauté du sujet, l'essentiel de la littérature est composé d'article de périodiques et surtout d'actes de congrès spécialisés dans l'archivage ; les travaux publiés dans le cadre des conférences de l'ECDL sont bien sûrs référentiels pour notre sujet, la prochaine conférence prévue courant 2004 à Bath (Angleterre) présentera sans doute des travaux intéressants.

On a également observé qu'il existait peu de travaux relatifs à l'archivage des sites fédérateurs contrairement à un grand nombre de ressources sur l'archivage du web en général. Bien sûr les techniques d'archivage ne sont pas fondamentalement différentes, mais ça n'explique pas tout. Les projets internationaux sur les sites fédérateurs sont récents, les plus anciens remontent à la fin des années 1990 et traitent surtout du problème du référencement.

La partie consacrée à la synthèse à mis en lumière ces problèmes, mais il faut souligner que l'archivage du web et des sites fédérateurs est tributaire de nombreux "aléas" qui vont des contraintes techniques en passant par les différents modes de conservation et le besoin d'uniformiser le référencement même si le standard RDF affiche de belles promesses. Mais les contraintes sont surtout d'ordres économiques et juridiques. Mener une politique d'archivage coûte très cher ; dans le monde, seule Internet Archive peut se permettre une politique d'archivage à grande échelle[67]. Enfin, il est nécessaire de se pencher sur le coté juridique. Beaucoup de barrières concernant les droits moraux d'auteurs freinent cet archivage, de même que pour les projets internationaux, il est nécessaire d'harmoniser les législations de chaque pays.

---

[67] Internet archive appartient à la société Alexa qui a elle-même racheté Amazon.com. La Bibliothèque du Congrès dispose aussi d'un important budget. Les Etats-Unis comptent aussi d'importantes fondations privées qui mise sur l'archivage à long terme des ressources numérique comme la Fondation Mellon.

Bibliographie

1. L'archivage du web et des sites fédérateurs : Principales conférences et généralités

1.1. Les conférences sur les bibliothèques numériques

AGOSTI M., **THANOS** C. Eds., *Research and advanced technology for digital libraries : 6th European conference, ECDL 2002, Rome, Italy, September 16-18*, 2002, 664 p. ISBN: 3-540-44178-6.

CONSTANTOPOULOS P., **SØLVBERG** I.T.E. Eds., *Research and advanced technology for digital librairies : 5th European conference, ECDL 2001, Darmstadt, Germany, September 4-9, 2001*, 462 p. ISBN : 3-540-42537-3.

KOCH T., **SØLVBERG** I.T. Eds., *Research and advanced technology for digital libraries : 7th European conference, ECDL 2003, Trondheim, Norway, August 17-22, 2003* [**en ligne**], 536 p. Disponible sur : <http://bibnum.bnf.fr/ecdl/2003/index.html> (Consulté le 02/12/2003) ISBN: 354040726X.

1.2. Travaux généralistes

BEAGRIE N., *Research and Community Collections from the Web*. **In** : AGOSTI M., THANOS C. Eds. Research and advanced technology for digital libraries : 6th European conference, ECDL 2002, Rome, Italy, September 16-18, 2002, 664 p.

BEAGRIE N., **GREENSTEIN** A., *Strategic policy framework for creating and preserving digital collections*. [**en ligne**]. Disponible sur : <http://www.ahds.ac.uk/old/manage/framework.htm> (Consulté le 06/12/2003)

BEAGRIE N., **POTHEN** P., *Web-archiving: managing and archiving online documents and records*. Ariadne, 2002, n°32.

BEARMAN D., *Reality and chimeras in the preservation of electronic records*. D-Lib Magazine [**en ligne**]. 1999, vol.5, n°4. Disponible sur: <http://www.dlib.org/dlib/april99/bearman/04bearman.html> (Consulté le 15/01/2004)

BERGMARK D., *Automatic Collection Building*. **In** : AGOSTI M., THANOS C. Eds. Research and advanced technology for digital libraries : 6th European conference, ECDL 2002, Rome, Italy, September 16-18, 2002, 664 p.

BERTHON H., **THOMAS** S., **WEBB** C., *Safekeeping: a cooperative approach to building a digital preservation resource*. D-Lib Magazine [**en ligne**]. 2002, vol. 8, n°1. Disponible sur : <http://www.dlib.org/dlib/january02/berthon/01berthon.html> (Consulté le 20/12/2003)

BRADLEY P., *Archiving the Internet*. Records Management Bulletin, 2002, n°108, p.7-10.

BRODIE N., *Collaboration entre les bibliothèques nationales en vue de conserver l'information numérique*. Nouvelles de la Bibliothèque Nationale [**en ligne**]. 1999, vol.32, n°3-4. Disponible sur : <http://www.nlc-bnc.ca/9/1/p1-259-f.html> (Consulté le 10/11/2003)

BROWNE G., *Selection criteria*. Online Currents, 2002, vol. 17, n°6, p.13-16.

CLAVEL-MERRIN G., *Initiatives in the field of long-term digital preservation and the need for a continued research effort*. Zeitschrift Fur Bibliothekswesen Und Bibliographie, 2001 vol.48, n° 3-4, p.184-187.

CLIFF P., *Building Resource Finder*. Ariadne [**en ligne**]. 2002, n°30. Disponible sur: <http://www.ariadne.ac.uk/issue30/rdn-oai/> (Consulté le 13/12/2003)

COUNCIL ON LIBRARY AND INFORMATION RESOURCES, *The state of digital preservation: an international perspective*, [**en ligne**]. 2002. Disponible sur : <http://www.clir.org/pubs/reports/pub107/pub107.pdf> (Consulté le 04/12/2003)

COBENA G., *Crawling Important Sites on the Web*. **In** : AGOSTI M., THANOS C. Eds. Research and advanced technology for digital libraries : 6th European conference, ECDL 2002, Rome, Italy, September 16-18, 2002, 664 p.

DAY M., *Collecting and preserving the world wide web*, [**en ligne**]. 2003. Disponible sur : <http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf> (Consulté le 20/12/2003)

FERRAN P., *Crawling, XML Storage and Query*. **In** : AGOSTI M., THANOS C. Eds. Research and advanced technology for digital libraries : 6th European conference, ECDL 2002, Rome, Italy, September 16-18, 2002, 664 p.

GRANGER S., *Emulation as a digital preservation strategy*. D-Lib Magazine [**en ligne**]. 2000, vol.6, n°10. Disponible sur: <http://www.dlib.org/dlib/october00/granger/10granger.html> (Consulté le 10/02/2004)

HAETTIGER M., *L'archivage des sites Web d'intérêt régional*. Diplôme de Conservateur de Bibliothèque, Ecole Nationale Supérieure des Sciences de l'information et des Bibliothèques (ENSSIB), Villeurbanne, 2003, 121 p.

HODGE M.G., *Best practices for digital archiving*. The Journal of Electronic Publishing [**en ligne**]. 2000, vol.5, n°4. Disponible sur: <http://www.press.umich.edu/jep/05-04/hodge.html> (Consulté le 20/12/2003)

HOLDSWORTH D., **WEATHLEY** P., *Emulation, preservation and abstraction*. RLG DigiNews [**en ligne**]. 2001, vol.5, n°4. Disponible sur : <http://www.rlg.ac.uk/preserv/diginews/diginews5-4.html#feature2> (Consulté le 10/01/2004)

HUC C., *Le modèle de référence pour les systèmes ouverts d'archivage*. Document numérique, 2000, vol.4, n°3-4, p.233-51.

KAVCIC-COLIC A., *Archiving the Web - some legal aspects*. Library Review, vol.52, n°5, 2003, p.203-8.

KAWANO H., *Web Archiving Strategies by using Web Mining Techniques*. **In** : Communications, computers and signal processing - Pacific rim conference; 9th IEEE Pacific Rim Conference On Communications Computers And Signal Processing (Victoria, Can), Vol.2, 2003, p.915-918. ISBN: 0780379780.

KENNEY A. & **al.**, *Preservation risk management for web resources: Virtual remote control in Cornell's project Prism*. D-Lib Magazine [**en ligne**]. 2002, vol.8, n°1.Disponible sur : <http://www.dlib.org/dlib/january02/kenney/01kenney.html> (Consulté le 20/12/2003)

LEE K.H. et **al.**, *The state of art and pratice in digital preservation*. Journal of Research of the National Institute of Standards and Technology [**en ligne**]. vol.107, n°1, 93-106. Disponible sur: <http://www.nist.gov/jres> (Consulté le 15/12/2003)

LEGER D., *Legal Deposit and the Internet : Reconciling Two Worlds*. **In** : CONSTANTOPOULOS P., SØLVBERG I.T.E. Eds. Research and advanced technology for digital librairies : 5th European conference, ECDL 2001, Darmstadt, Germany, September 4-9, 2001, 462 p.

LIU X., **MALY** K., **ZUBAIR** M., **NELSON** M.L., *DP9: an OAI gateway service for Web crawlers*. **In** :JCDL 2002. Proceedings of the Second ACM/IEEE-CS Joint Conference on Digital Libraries (Portland), 2002, p. 283-4. ISBN: 1 58113 513 0.

LUDASCHER B., **MARCIANO** R., **MOORE** R., *Preservation of digital data with self-validating, self-instantiating knowledge-based archives*. SIGMOD Record, 2001, vol.30, n°3, p.54-63.

LUPOVICI C., *Les besoins et les données techniques de préservation*. **In** : 67th IFLA Council and General Conference [**en ligne**]. 2001. Disponible sur: <http://www.ifla.org/IV/ifla67/papers/163-168f.pdf> (Consulté le 01/12/2003)

LYMAN P., *Archiving the World Wide Web. In Building a National Strategy for Digital Preservation: Issues in Digital Media Archiving* [**en ligne**]. 2002, p.39-51. ISBN: 1-887334-91-2. Disponible sur : <http://www.clir.org/pubs/reports/pub106/web.html> (Consulté le 30/11/2003)

MASANÈS J., *L'archivage des sites Internet*. Diplôme de Conservateur de Bibliothèque, Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques, Villeurbanne, 2000, 24 p.

MASANES J., *Towards continuous Web archiving*. D-Lib Magazine [**en ligne**]. 2002, vol.8, n°12. Disponible sur : <http://www.dlib.org/dlib/december02/masanes/12masanes.html> (Consulté le 10/12/2003)

MASANES J., *Préserver les contenus du Web* [**en ligne**]. Disponible sur <http://bibnum.bnf.fr/conservation/migration_web.pdf > (Consulté le 10/12/2003)

MASANES J., *Archiving the Deep Web*. **In** : AGOSTI M., THANOS C. Eds. Research and advanced technology for digital libraries : 6th European conference, ECDL 2002, Rome, Italy, September 16-18, 2002, 664 p.

MEDEIROS N., *Reap what you sow: harvesting the deep Web*. OCLC Systems & Services, 2002, vol.18, n°1, p. 18-20.

MITCHELL R. L., *Preserving Web history: the technical challenges of Web site archiving make it difficult for companies to maintain the records of past activities and....* Computerworld, 2002, vol.36, n°32, p.26-27.

OCLC, *Web characterization* [**en ligne**]. Disponible sur : <http://wcp.oclc.org> (Consulté le 20/12/2003)

PIERRE J.M., *Practical issues for automated categorization of web sites* [**en ligne**]. 2000. Disponible sur : <http://ww.ics.forth.gr/isl/SemWeb/proceedings/session3-3/html_version/semanticweb.html> (Consulté le 22/12/2003)

PITSCHMANN L.A., *Building Sustainable Collections of Free Third-Party Web Resources*, Digital Library Federation, Council on Library and Information Resources, Washington, 2001, 44p. ISBN: 1-887334-83-1.

PYMM B., *Preserving digital information: A how-to-do-it manual Hunter, GS - English, 2000*. Online Information Review, 2001, vol.25, n°1, p.67-68.

ROUMIEUX O., *La quadrature du Web : Le Web à la recherche de sa mémoire (The quadrature of the Web)*. Archimag, 2001, n°145, p.30-32.

THIBODEAU K., *Building the archives of the future: Advances in preserving electronic records at the National Archives and Records Administration*. D-Lib Magazine [**en ligne**]. 2001, vol.7, n°2. Disponible sur:

<http://www.dlib.org/dlib/february01/thibodeau/02thibodeau.html>
(Consulté le 29/12/2003)

WIGGINS R., *Digital preservation: paradox & promise*. Library Journal netconnect, 2001, p.12-15.

2. Cas pratiques : les travaux en Europe septentrionale, Aux Etats-Unis, en Australie et en Europe Occidentale

2.1. Les pays Nordiques

ALBERTSEN K., *The Paradigma Web Harvesting Environnment*. **In** : KOCH T., SØLVBERG I.T. Eds. Research and advanced technology for digital libraries : 7th European conference, ECDL 2003 [**en ligne**], Trondheim, Norway, August 17-22, 2003, p.49-62. Disponible sur : <http://bibnum.bnf.fr/ecdl/2003/index.html>
(Consulté le 03/12/2003)

ARVIDSON A., *Harvesting the Swedish web space*.**In** : CONSTANTOPOULOS P., SØLVBERG I.T.E. Eds. Research and advanced technology for digital librairies : 5th European conference, ECDL 2001, Darmstadt, Germany, September 4-9, 2001, 462 p.

ARVIDSON A., **PERSSON** K., **MANNERHEIM** J., *The kulturarw3 Project, the Royal Swedish Web archive: An example of "complete" collection of web pages*. **In** 66th IFLA Council and General Conference, Jérusalem, 2000 [**en ligne**]. Disponible sur : <http://www.ifla.org/IV/ifla66/papers/154-157e.htm>
(Consulté le 26/12/2003)

BRYGFJELD S.A., *Access to Web archives: The Nordic Web Archive Access Project approach*. Zeitschrift Fur Bibliothekswesen Und Bibliographie, 2002, vol.49 , n°4, p.227-231.

HAKALA J., *Harvesting the Finnish Web space - practical experiences*. **In** : CONSTANTOPOULOS P., SØLVBERG I.T.E. Eds. Research and advanced technology for digital librairies : 5th European conference, ECDL 2001, Darmstadt, Germany, September 4-9, 2001, 462 p.

HALLGRIMSSON P.**, BANG** S., *Nordic Web Archive*. **In** : KOCH T., SØLVBERG I.T. Eds. Research and advanced technology for digital libraries : 7th European conference, ECDL 2003 [**en ligne**], Trondheim, Norway, August 17-22, 2003, p.37-48. Disponible sur : <http://bibnum.bnf.fr/ecdl/2003/index.html>
(Consulté le 02/12/2003)

HENRIKSEN B., *The Danish Project netarchive.dk*. **In** : AGOSTI M., THANOS C. Eds. Research and advanced technology for digital libraries : 6th European conference, ECDL 2002, Rome, Italy, September 16-18, 2002, 664 p.

HENRIKSEN B., *Danish Legal Deposit on the Internet: Current Solutions and Approaches for the Future*. **In** : CONSTANTOPOULOS P., SØLVBERG I.T.E. Eds. Research and advanced technology for digital librairies : 5th European conference, ECDL 2001, Darmstadt, Germany, September 4-9, 2001, 462 p.

ROYAL LIBRARY OF SWEDEN, *Kulturarw3 heritage project* [**en ligne**]. Disponible sur: <http://www.kb.se/kw3/ENG/Defaults.htm>
(Consulté le 20/12/2003)

2.2. Etats-Unis et Australie

AMMEN C., *MINERVA: Mapping the INternet Electronic Resources Virtual Archive -Web Preservation at the Library of Congress*. **In** : CONSTANTOPOULOS P., SØLVBERG I.T.E. Eds. Research and advanced technology for digital librairies : 5th European conference, ECDL 2001, Darmstadt, Germany, September 4-9, 2001, 462 p.

BATES M.E., *The Wayback Machine*. Online, 2002, vol. 26, n° 6, p.80.

BREWSTER K., *The Internet Archive*. RLG DigiNews [**en ligne**]. 2002, vol.6, n°3. Disponible sur: <http://www.rlg.org/preserv/diginews/diginews6-3.html>
(Consulté le 10/02/2004)

CATHRO W., **WEBB** C., **WHITING** J., *Archiving the Web: the PANDORA archive at the National Library of Australia. Preserving the present for the future*, **In** : Web archiving conference [**en ligne**], 2001, Copenhague. Disponible sur : <http://www.nla.gov.au/nla/staffpaper/2001/cathro3.html>
(Consulté le 25/01/2004)

FEISE J., *Accessing the history of the Web: a Web way-back machine*.**In** REICH S., ANDERSON K.M. Eds. Open Hypermedia Systems and Structural Computing. 6th International Workshop, OHS-6. 2nd International Workshop (San Antonio), SC-2. Proceedings May 30-June 3, 2000, p.38-45. ISBN: 3 540 41084 8.

GUPTA A., *Preserving presidential library websites*. Technical report [**en ligne**], San Diego Supercomputer Center, 2001, San Diego. Disponible sur : <http://www.sdsc.edu/TR/TR-2001-03.pdf> (Consulté le 26/01/2004)

KAHLE B., *"Editors" interview: The Internet Archive*. RLG DigiNews [**en ligne**]. 2002, vol.6, n°3. Disponible sur : <http://www.rlg.org/preserv/diginews/diginews6-3.html#interview>

(Consulté le 12/02/2004)

KIMPTON M., *Internet Archive Consortium Presentation*. **In** : AGOSTI M., THANOS C. Eds. Research and advanced technology for digital libraries : 6th European conference, ECDL 2002, Rome, Italy, September 16-18, 2002, 664 p.

KRESH D.N., *Courting Disaster: Building a Collection to Chronicle 9/11 and its Aftermath ; LC September 11 Web archive*. Library of Congress Information Bulletin, 2002, vol.61, n°9, p.151-5.

LIBRARY OF CONGRESS, *The national digital information infrastructure preservation program* [**en ligne**]. Disponible sur: <http://www.digitalpreservation.gov/ndiipp> (Consulté le 14/02/2004)

LIBRARY OF CONGRESS, *Election 2000, as it happened: Library and Alexa announce election Web archives*. Library of Congress Information Bulletin, 2001, vol.60, n°7/8, p.161.

LIBRARY OF CONGRESS, *September 11 Web archive: Library and partners announce digital project ; LC, Internet Archive, Pew Internet & American Life Project, webArchivist*. Library of Congress Information Bulletin, 2001, vol.60, n°10, p.215.

NATIONAL LIBRARY OF AUSTRALIA, *PANDORA Archive: PANDAS manual* [**en ligne**]. Disponible sur: <http://pandora.nla.gov.au/manual/pandas/general.html> (Consulté le 01/02/2004)

NATIONAL LIBRARY OF AUSTRALIA, *Digital library of Australia* [**en ligne**]. Disponible sur: <http://www.nla.gov.au/dsp/> (Consulté le 01/02/2004)

NATIONAL LIBRARY OF AUSTRALIA, *National strategy for provision access to australian elcetronic publications: A national library of Australia position paper* [**en ligne**]. Disponible sur : <http://www.nla.gov.au/policy/paep.html#eleven> (Consulté le 01/02/2004)

NESBEITT S.L., *The Internet Archive Wayback Machine*. Online Information Review, 2002, vol.26, n°2, p.128-129.

NOTESS G. R., *The Wayback Machine: the Web's archive*. Online, 2002, vol.26, n°2, p.59-61.

O'LEARY M., *Internet Archive joins history's great libraries*. Information Today, 2003, vol.20, n°10, p.41-46.

PANOS P., *The Internet archive: An end to the digital dark age*. Journal of Social Work Education, 2003, vol.39, n°2, p.343-347.

REILLY B. et **al.**, *Political Communications Web Archiving : Adressing Typology and Timing for Selection, Preservation and Access*, **In** : KOCH T., SØLVBERG I.T. Eds. Research and advanced technology for digital libraries : 7th European conference, ECDL 2003 [**en ligne**], Trondheim, Norway, August 17-22, 2003, p.94-102. Disponible sur : <http://bibnum.bnf.fr/ecdl/2003/index.html> (Consulté le 02/12/2003)

SCHNEIDER S.M., **FOOT** K., **KIMPTON** M.**, JONES** G., *Building Thematic Web Collections: Challenges and Experiences from the September 11 web Archive and the Election 2002 Web Archive*. **In** : KOCH T., SØLVBERG I.T. Eds. Research and advanced technology for digital libraries : 7th European conference, ECDL 2003 [**en ligne**], Trondheim, Norway, August 17-22, 2003, p.77-93. Disponible sur : <http://bibnum.bnf.fr/ecdl/2003/index.html> (Consulté le 02/12/2003)

STATA R., *Presentation of the Internet Archive*. **In** : AGOSTI M., THANOS C. Eds. Research and advanced technology for digital libraries : 6th European conference, ECDL 2002, Rome, Italy, September 16-18, 2002, 664 p.

2.3. En France et dans le reste de l'Europe

ABITEBOUL S., **COBENA** G., **MASANES** J., **SEDRATI** G., *A first experience in archiving the French Web*. **In** : AGOSTI M., THANOS C. Eds. Research and advanced technology for digital libraries : 6th European conference, ECDL 2002, Rome, Italy, September 16-18, 2002, 664 p.

DHERENT C., *L'archivage à long terme des documents électroniques en France*, [**en ligne**]. 2000. Disponible sur : <http://www.archivesdefrance.culture.gouv.fr/fr/notices/archi2.html> (Consulté le 14/12/2003)

GOMES D., **SILVA** M.J., *A Characterization of the Portuguese Web*. **In** : KOCH T., SØLVBERG I.T. Eds. Research and advanced technology for digital libraries : 7th European conference, ECDL 2003 [**en ligne**], Trondheim, Norway, August 17-22, 2003, p.63-76. Disponible sur : <http://bibnum.bnf.fr/ecdl/2003/index.html> (Consulté le 02/12/2003)

HAKALA J., *Collecting and preserving the web : developping and testing the NEDLIB harvester*. RLG DigiNews [**en ligne**]. 2001, vol.5, n°2. Disponible sur : <http://www.rlg.org/preserv/diginews/diginews5-2.html#feature2> (Consulté le 20/01/2004)

JENKINS C., *Presentation of the CEDARS project*. [**en ligne**]. Disponible sur: <http://leeds.ac.uk/cedars/> (Consulté le 23/01/2004)

MASANES J., *The BnF project for Web archiving*. **In** : CONSTANTOPOULOS P., SØLVBERG I.T.E. Eds. Research and advanced technology for digital librairies : 5th European conference, ECDL 2001, Darmstadt, Germany, September 4-9, 2001, 462 p. [**en ligne**]. Disponible sur : <http://bibnum.bnf.fr/ecdl/2001/france/slg001.htm> (Consulté le 13/01/2004)

RAUBER A., **ASCHENBRENNER** A., **WITVOET** O., *Austrian Online Archive Processing: analyzing archives of the World Wide Web*. **In** : AGOSTI M., THANOS C. Eds. Research and advanced technology for digital libraries : 6th European conference, ECDL 2002, Rome, Italy, September 16-18, 2002, 664 p.

SHENTON H., *From talking to doing: Digital preservation at the British Library* [**en ligne**]. 2000. Disponible sur: <http://www.rlg.org/events/pres-2000/shenton.html> (Consulté le 03/02/2004)

STEENBAKKERS J.F., *NEDLIB guidelines for setting up a deposit system for electronic publications*. Zeitschrift Fur Bibliothekswesen Und Bibliographie, 2001, vol.48, n°3-4, p.181-183.

VOERMAN G., **DEN HOLLANDER** F., **DRUIVEN** H., *Archiving the Web: Political party Web sites in the Netherlands*. Information services & use, 2003, vol.23, n°1, p. 1-7.

ZABICKA P., *Archivig the Czech Web: Issues and Challenges*. **In** : KOCH T., SØLVBERG I.T. Eds. Research and advanced technology for digital libraries : 7th European conference, ECDL 2003 [**en ligne**], Trondheim, Norway, August 17-22, 2003, p.111-117. Disponible sur : <http://bibnum.bnf.fr/ecdl/2003/index.html> (Consulté le 04/12/2003)

3. Sites fédérateurs et métadonnées

AMEROUALI Y., *Métadonnées basées sur l'association d'éléments de description de ressources et d'éléments de profil d'utilisateur*. Thèse de doctorat, Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques (ENSSIB),Villeurbanne 2001, 243 p. [**en ligne**]. Disponible sur : <http://www.enssib.fr/bibliotheque/document/these/amerouali/amerouali.pdf> (Consulté le 28/12/2003)

ARMSTRONG C., *Metadata, PICS and quality*. Ariadne [**en ligne**] 1997, n°9. Disponible sur: <http://www.ariadne.ac.uk/issue9/pics/> (Consulté le 07/01/2004)

AYRES M.L., **KILNER** K., **FITCH** K., **SCARVELL** A., *Report on the successful Auslit: Australian Literature Gateway implementation of the FRBR and INDEC event models, and implications for other FRBR implementations*. **In** 68th IFLA General Conference and Council, Glasgow, 2002. [**en ligne**]. Disponible sur:<http://www.ifla.org/IV/ifla68/papers/054-133e.pdf> (Consulté le 10/01/2004)

BARRUECO J.M., **COLL** I.S., *Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH): description, functions and applications of a protocol*. El Profesional de la Informacion, 2003, vol.12, n°2, p.99-106.

BECKETT D., *IAFA Templates in use as Internet Metadata* [**en ligne**]. Disponible sur: <http://www.w3.org/Conferences/WWW4/Papers/52/>
(Consulté le 20/01/2004)

BREEDING M., *Understanding the protocol for metadata harvesting of the Open Archives Initiative*. Computers in Libraries, 2002 vol. 22, n°8, p. 24-29.

CAMPBELL D., *Australian subject gateways -metadata as an agent of change*. **In** : Books and bytes: technologies for the hybrid library : proceedings, 10th biennial conference and exhibition, 16-18 February, 2000, Melbourne Convention Centre / Victorian Association of Library Automation Inc.p.421-430. ISBN : 0908478143. [**en ligne**]. Disponible sur :
<http://www.vala.org.au/vala2000/2000pdf/campbell.pdf>
(Consulté le 15/01/2004)

CHAPMAN A. D., *Metadata: cataloguing practice and Internet subject-based information gateways*. Ariadne, 1998, n°18.

COLE T.W., *Qualified Dublin Core Metadata for Online Journal Articles*. OCLC System & Services, 2002, vol.18, n°2, p.79-87.

CUNNINGHAM A., *Dynamic descriptions: recent developments in standards for archival description and metadata*. Canadian Journal of Information and Library Science, 2000, vol.25, n°4, p.3-17.

CEDARS, *Metadata for digital preservation: The CEDARS project outline specification* [**en ligne**]. Disponible sur:
<http://www.leeds.ac.uk/cedars/colman/metadata/metadataspec.html>
(Consulté le 06/12/2003)

CUNDIFF M., *Using METS for Web Archiving*. **In** : AGOSTI M., THANOS C. Eds. Research and advanced technology for digital libraries : 6th European conference, ECDL 2002, Rome, Italy, September 16-18, 2002, 664 p.

DAY Y.M., *Resource discovery, interoperability and digital preservation : some aspects of current metadata research and development*. Kingdom Journal: VINE. Very informal newsletter on library automation, 2000, n°117, p.35-48.

DAY M., *Metadata for digital preservation: a review of recent developments*. **In** : CONSTANTOPOULOS P., SØLVBERG I.T.E. Eds. Research and advanced technology for digital librairies : 5th European conference, ECDL 2001, Darmstadt, Germany, September 4-9, 2001, p.171-72.

DHERENT C., *Une DTD pour la description des fonds d'archives et collections spécialisées, l'EAD* [**en ligne**]. 2002. Disponible sur : <http://www.archivesdefrance.culture.gouv.fr/fr/archivistique/pr%E9sentationEAD.html> (Consulté le 21/12/2003)

DUFF W.M., *Evaluating metadata on a metalevel*. Archival Science, 2001, vol.1, n°3, p. 285-94.

DUVAL E. & **al.**, *Metadata principles and practibilities*. D-Lib Magazine [**en ligne**]. 2002, vol.8, n°4. Disponible sur: <http://www.dlib.org/dlib/april02/weibel/04weibel.html> (Consulté le 18/01/2004)

EMBLEY D.W. & **al**., *Conceptual-model-based data extraction from multiple-record Web pages*. Data and Knowledge Engineering, 1999, n°31, p.227-251.

FAST K.V., **CAMPBELL** D.G., *The ontological perspectives of the Semantic Web and the metadata harvesting protocol: applications of metadata for improving Web search*. Canadian Journal of Information and Library Science, 2001, vol.26, n°4, p. 5-19.

GILLILAND-SWETLAND A. J., *Popularizing the finding aid: exploiting EAD to enhance online discovery and retrieval in archival information systems by diverse user groups*. Journal of Internet Cataloging, 2001, vol.4, n°3-4, p.199-225.

HAKALA J., **HANSEN** P., **HUSBY** O., **KOCH** T., *The Nordic metadata project: Final report* [**en ligne**]. Helsinki Universitary Library, 1998. Disponible sur: <http://linnea.helsinki.fr/meta/nmfinal.htm> (Consulté le 17/12/2003)

HENSEN S.L., *Archival cataloging and the Internet: the implications and impact of EAD*. Journal of Internet Cataloging, 2001, vol.4, n°3-4, p.75-95.

HUTHWAITE A., *AACR2 and other metadata standards: the way forward*. Cataloging & Classification Quarterly, 2003, vol. 36, n°3/4, p. 87-100.

IFLA Study Group on the Functional Requirements for Bibliographic Records, *Functional Requirements for Bibliographic Records report* [**en ligne**]. 1997, Francfort, 144p. ISBN : 3-598-11382-X. Disponible sur : <http://www.ifla.org/VII/s13/frbr/frbr.pdf> (Consulté le 10/02/2004)

LAZINGER S. S., *Digital Preservation and Metadata: History, Theory, Practice*. Librairies Unlimited, Englewood, 2001, 359p. ISBN: 1-56308-777-4.

Library of Congress, *Encoded archival description (EAD) Official EAD Version 2002 Web Site* [**en ligne**]. Disponible sur: <http://www.loc.gov/ead/> (Consulté le 23/01/2004)

Library of Congress, *METS Standard* [**en ligne**]. Disponible sur: <http://www.loc.gov/standards/mets/> (Consulté le 23/01/2004)

Library of Congress, *MODS Standard* [**en ligne**]. Disponible sur: <http://www.loc.gov/standards/mods/> (Consulté le 23/01/2004)

LIDDLE S.W., **YAU** S.H., **EMBLEY** D.W., *On the automatic extraction of data from the hidden web* [**en ligne**]. Disponible sur : <http://www.deg.byu.edu/papers/daswis01.pdf> (Consulté le 11/12/2003)

LUPOVICI C., *Identification des ressources sur Internet et des métadonnées : la diversité des standards*. Documentaliste, 1999, vol.36, n°6, p.321-25.

LUPOVICI C., **MASANES** J., *Metadata for long-term preservation* [**en ligne**]. Disponible sur : <http://www.kb.nl/coop/nedlib/results/D4.2/D4.2.htm> (Consulté le 05/12/2003)

MEDEIROS N., *A craftsman and his tool: Andy Powell and the DC-dot metadata editor*. OCLC Systems & Services, 2001, vol.17, n°2, p.60-64.

MEDEIROS N., *Report from the trenches – the 8th International Dublin Core Metadata Initiative Workshop*. OCLC Systems & Services, 2001, vol.17, n°1, p.15-18.

MEDEIROS N., *Peering over the fortress walls: the medatata invasion begins*. OCLC Systems & Services, 2001, vol.17, n°4, p.154-56.

POWELL A., *DC-DOT RDF editor* [**en ligne**]. 2000. Disponible sur : <http://www.ukoln.ac.uk/metadata/dcdot/> (Consulté le 27/12/2003)

REMIZE M., *Archives et bibliothèques : tous pour l'EAD*. Archimag, 2003, n°160, p.18-19.

SAIDI S., *Utilisation des métadonnées dans les SBIG (Suject-Based Information Gateways) ou Sites Fédérateurs de Qualité Contrôlée*. Rapport de recherche bibliographique, Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques (ENSSIB), Villeurbanne, 2003, 55 p.

SAITO E., **ONODERA** N., *The use of metadata for science resources on the Web*. Journal of Information Processing and Management, 2001, vol.44, n°3, p. 174-83.

THIRION B., **LOOSLI** G., **DOUYERE** M., **DARMONI** S. J., *Metadata Element Set in a Quality-controlled Subject Gateway: A Step to a Health Semantic Web*.**In** :Medical informatics Europe; The new navigators: from professionals to patients - International conference : Studies In Health Technology And Informatics, 2003, vol. 95, p.707-712.

THORNELY J., *Metadata and the deployment of Dublin Core at State Library of Queensland and Education Queensland, Australia*. OCLC Systems & Services, 2001, vol.16, n°3, p.118-19.

WEIBEL S., *The evolving metadata architecture for the World Wide Web: Bringing together the semantics, structure and syntax of ressource description* [**en ligne**]. Disponible sur:<http://www.dl.ulis.ac.jp/ISDL97/proceedings/weibe.html> (Consulté le 30/12/2003)

World Wide Web Consortium (W3C), *Introduction to DTD* [**en ligne**]. Disponible sur : <http://www.w3schools.com/dtd/dtd_intro.asp> (Consulté le 11/01/2004)

4. Les sites fédérateurs : typologies, exemples de sites et de projets.

ARDOE A., **BERGGREN** M., **KOCH** T., **KRINGSTAD** R., *Nordic Interconnected Subject Based Information Gateways (NISBIG). Final report*. Nordinfo-Nytt, vol.23, n°3, 2000, p.7-33.

BAWDEN D., **ROBINSON** L., *Internet subject gateways revisited*. International Journal of Information Management, 2002, vol.22, n°2, p. 157-62.

BOMEKE M., *The Engineering Subject Gateway (ViFaTec)*. IATUL Proceedings, New Series [**en ligne**], 2002, vol.12. Disponible sur : <http://educate.lib.chalmers.se/IATUL/proceed.html> (Consulté le 30/01/2004)

CALHOUN K., *From information gateway to digital library management system: a case analysis*. Library Collections, Acquisitions, & Technical Services, 2002, vol.26, n°2, p.141-50.

CALHOUN K., *CORC and collaborative Internet resource description: A new partnership for technical services, collection development and public services*. Journal of Internet Cataloguing, 2001, vol.4, n°1-2, p.131-42.

CAMPBELL D., *Australian subject gateways: political and strategic issues*. Online Information Review, 2000, vol.24, n°1, p.73-77.

CAMPBELL D., *An overview of subject gateway activities in Australia*. Ariadne [**en ligne**], 1999, n°21. Disponible sur: <http://www.ariadne.ac.uk/issue21/subject-gateways> (Consulté le 20/12/2003)

DAWSON H., *Building a virtual library for the social sciences: SOSIG (the Social Science Information Gateway) and the Library of the London School of Economics and Political Science*. Behavioral & Social Sciences Librarian, 2002, vol.20, n°2, p.17-27.

DEMPSEY L., *The subject gateway: experiences and issues based on the emergence of the Resource Discovery Network*. Online Information Review, 2000, vol.24 , n°1, p.8-23.

DESIRE, *Subject Gateway Community : Imesh* [**en ligne**]. Disponible sur: <http://www.desire.org/html/subjectgateways/community/imesh/> (Consulté le 10/02/2004)

DOYLE C., **KJELLBERG** S., *Renardus: an example of co-operation between subject gateways*. Tidskrift for Dokumentation (Nordic Journal of Documentation) [**en ligne**]. 2002, vol.57, n°1, p.11-20. ISSN: 0040-6872. Disponible sur : <http://www.tls.se/publikationer/tdnew_eng.lasso> (Consulté le 14/02/2004)

FISCHER TH., **NEUROTH** H., *SSG-FI--special subject gateways to high quality Internet resources for scientific users*. Online Information Review, 2000, vol.24, n°1, p.64-68.

GILL T., **GROUT** C., *Finding and preserving visual arts resources on the Internet ; ADAM: Art, Design, Architecture & Media Information Gateway and VADS: Visual Arts Data Service*. Art Libraries Journal, 1997, vol.22, n°3, p.19-25.

GILLES A., **JUPILE** B., **NIESZKOWSKA** E., **PERICARD** C., *Evaluation des sites web fédérateurs, Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques* (ENSSIB), Villeurbanne, 2002, 48p.

GRAY L., **MANNING**, P., **WARD** R., *OMNI - quality filtered access to the healthy Internet*. **In** : PATEL V.L., ROGERS R., HAUX R. Eds., Medinfo, Proceedings of the 10th World Congress on Medical Informatics (Londres), vol.2, 2001, p.1495. ISBN: 1 58603 194 5.

HEERY R., **CARPENTER** L., **DAY** M., *Renardus project developments and the wider digital library context*. D-Lib Magazine [**en ligne**]. 2001, vol.7, n°4. Disponible sur : <http://www.dlib.org/dlib/april01/heery/04heery.html> (Consulté le 05/02/2004)

HEERY R., *Information gateways: collaboration on content*. Online Information Review, 2000, vol.24, n°1, p.40-45.

HILL C. *Building gateways: a case study of the Australasian virtual engineering library*. LASIE, 2000, vol.31, n°1, p.4-10.

HIOM D., *SOSIG: an Internet hub for the social sciences, business and law*. Online Information Review, 2000, vol. 24, n°1, 2000, p.54-58.

HUXLEY L., **PLACE** E., **BOYD** D., **CROSS** P., *Planet SOSIG - a spring-clean for SOSIG: a systematic approach to collection management*. Ariadne, 2002 n°33.

HUXLEY L., **CARPENTER** L., **PEEREBOOM** M., *The Renardus broker service: collaborative frameworks and tools*. Electronic Library, 2003, vol.21, n°1, p. 39-48.

HUXLEY L., *Renardus: Following the fox from project to service Research and advanced technology for digital libraries*. **In** : AGOSTI M., THANOS C. Eds. Research

and advanced technology for digital libraries : 6th European conference, ECDL 2002, Rome, Italy, September 16-18, 2002, 664 p.

HUXLEY L., *Renardus: fostering collaboration between academic subject gateways in Europe*. Online Information Review, 2001, vol.25, n°2, p.121-7.

KOCH T., *Quality-controlled subject gateways: definitions, typologies, empirical overview*. Online Information Review, 2000, vol.24, n°1, p.24-34.

KOCH T., *Subject gateways – Introduction*. Online Information Review, 2000, vol.24, n°1, p.6-7.

LAROUK O., *Classifications et Méta-données pour le Web sémantique: Typologie et évaluation des sites fédérateurs*. **In** : Conférence internationale SETIT : Sciences Electroniques, technologies de l'Information et des Télécommunications, Ecole Nationale Supérieure des Télécommunications de Bretagne, ENST-Brest, 2003, p.235-251.

MACLEOD R., *Promoting a subject gateway: a case study from EEVL (Edinburgh Engineering Virtual Library)*. Online Information Review, 2000, vol.24, n°1, p.59-63.

MONOPOLI M., **NICHOLAS** D., *A user evaluation of subject based information gateways: case study ADAM*. Aslib Proceedings-New Information Perspectives, 2001, vol.53, n°1, p.39-52.

MONOPOLI M., **NICHOLAS** D., *A user-centred approach to the evaluation of subject based information gateways: case study SOSIG*. Aslib Proceedings-New Information Perspectives, 2000, vol.52, n°6, p.218-31.

NEUROTH H., *Suche in verteilten "Quality-controlled Subject Gateways"Entwicklung eines Metadatenprofils (Searching in "Quality controlled Subject Gateways". Development of Metadata profiles)*. Bibliothek, 2002, vol.26, n°3, p.275-296.

PEEREBOOM M., *DutchESS: Dutch Electronic Subject Service - a Dutch national collaborative effort*. Online Information Review, 2000, vol.24, n°1, p.46-48.

PLACE E., *International collaboration on Internet subject gateways*. IFLA Journal, 2000 vol.26, n°1, p.52-56.

THELWALL M., *Subject gateway sites and search engine ranking*. Online Information Review, 2002, vol.26, n°2, p.101-7.

VAN HALM J., *Information gateways*. New Library World, 2002, vol.10, p. 222-224.

5. Le Web Invisible

BERGMAN M.K., *The Deep Web : Surfacing Hidden Value*. The Journal of Electronic Publishing [**en ligne**]. 2001, vol.7, n°1. Disponible sur : <http://www.press.umich.edu/jep/07-01/bergman.html> (Consulté le 13/12/2003)

EDOLS L., *Uncovering the invisible Web*. Online Currents, 2000, vol.15, n°8, p. 6-8.

LACKIE R.J., *Those dark hiding places: The Invisible Web revealed* [**en ligne**]. 2001. Disponible sur: <http://library.rider.edu/scholarly/rlackie/Invisible/Inv_Web.html> (Consulté le 15/12/2003)

LARDY J.P., *Le Web invisible, compte-rendu à la Commission Internet de l'Addnb le vendredi 22 mars 2002* [**en ligne**]. Disponible sur : <http://addnb.org/fr/docs/webinvisible.htm> (Consulté le 13/12/2003)

LAWRENCE S., **LEE GILES** C., *Accesibility of information on the Web*. Nature, 1999, n°400, p.107-109.

LAWRENCE W.G. & **al.**, *Risk management of digital information : A file format investigation* [**en ligne**]. 2000. Disponible sur: <http://www.clir.org/pubs/reports/pub93/contents.html> (Consulté le 20/12/2003)

OUF R., *Le dynamisme du Worl Wide Web : taille, croissance, visibilité, distribution et accessibilité de l'information*. Rapport de recherche bibliographique, Ecole Nationale Supérieure des Sciences de l'Information et des Bibliothèques (ENSSIB), Villeurbanne, 2001, 58p.

PRICE G., **SHERMAN** C., *The Invisible Web: Uncovering information sources search engine can't see*, CyberAge Books, Medford, 2001, 439p. ISBN : 091096551X.

RAGHAVAN S., **GARCIA-MOLINA** H., *Crawling the hidden Web* [**en ligne**], Technical Report 200-36, Computer Science Departement, Stanford University, 2000. Disponible sur : <http://dbpubs.stanford.edu/pub/2000-36> (Consulté le 06/01/2003)

RAUBER A., **ASCHENBRENNER** A., **WITVOET** O., **BRUCKNER** R. & **Al**., *Uncovering information hidden in Web archives: a glimpse at Web analysis building on data warehouses*. D-Lib Magazine [**en ligne**]. 2002. vol.8, n°12. Disponible sur : <http://www.dlib.org/dlib/december02/rauber/12rauber.html> (Consulté le 10/01/2003)

6. Sites web et listes de discussions relatifs aux archives

Archipol Netherlands : <http://www.archipol.nl> (archives politiques des Pays-Bas).

Archive nationale tchèque : <http://www.webarchiv.cz>

Archives relatives aux attentats du 11 septembre et aux élections fédérales de 2002 : <http://loc.gov> ; <http://www.webarchivist.org>

Bibliothèque Nationale de France, Département de la bibliothèque numérique : <http://bibnum.bnf.fr>

Bibliothèque Nationale de France, Département de la bibliothèque numérique : Présentation des conférences de l'ECDL (Powerpoint) : <http://bibnum.bnf.fr/ecdl/index.html>

FAST Search & Transfer ASA : <http://fastserach.com> (solutions logicielles)

Internet Archive : <http://www.archive.org>

Imesh, *The Imesh Toolkit : An architecture and toolkit for Distributed Gateways* : <http://www.imesh.org/toolkit/>; <imesh@mailbase.ac.uk>

Kulturarw3 : <http://www.kb.se/kw3/ENG/Default.htm>

LANIC (*Latin American Network Information Center*) <http://www.lanic.utexas.edu>

Liste de discussions de la BnF sur l'archivage du web, web-archive@cru.fr. Disponible sur : <http://listes.cru.fr/wws/info/web-archive>

Networked European Digital LIBrary (NEDLIB) : <http://wwwkb.nl/coop/nedlib/>

Nordic Web Archiv : <http://nwa.nb.no>

Occasio: digital social history archive : <http://www.iisg.nl/occasio/index.html>

Webarchivist : <http://www.webarchivist.org>

Welsh political archive : <http://www.llgc.org.uk/lc/awg_s_awg.htm> (archive politique du Pays de Galles)

Table des annexes

Annexe : principaux acteurs, projets et sites fédérateurs.

Internet Archive et Webarchivist:

Comme nous l'avons vu, la Bibliothèque du Congrès a collaboré dans ses projets d'archivage du web (notamment pour les archives des attentats du 11 septembre 2001 et des élections fédérales de 2002) avec Internet Archive et Webarchivist. La Bibliothèque du Congrès est la plus grande bibliothèque du monde, c'est aussi la plus ancienne institution fédérale culturelle des Etats-Unis. Sa mission est de rendre disponible des ressources au Congrès et au peuple américain.

Pionnière dans l'archivage du web, Internet Archive est une organisation à but non-lucratif, elle détient la plus grande collection numérique, plus de 150 tera octet des données. Ses collections sont accessibles gratuitement aux chercheurs, historiens et autres spécialistes. Pour y accéder, il faut utiliser le moteur de recherche conçu par Alexa, société qui appartient à Amazon.com et qui est le principal bailleur de fonds d'Internet Archive. Ce moteur est appelé la Wayback Machine. Le siège d'Internet Archive se trouve à San Francisco.

Privilégiant un archivage exhaustif du web, son objectif est de capturer le maximum de ressources du web mondial, par contre, on n'y trouve pas de ressources du Web Invisible.

Webarchivist.org est un groupe composé de chercheurs dirigé par S. Schneider du SUNY Institute of Technology et K. Foot de l'Université de Washington. Elle développe des systèmes pour identifier, collecter, cataloguer et analyser à grande échelle les archives web. Ces chercheurs travaillent souvent en collaboration avec Internet Archive.

Pour revenir sur les archives du 11 septembre et des élections fédérales de 2002, voici un tableau qui résume bien cet archivage :

Table 1. Collection Overview for the September 11 Web Archive and the Election 2002 Web Archive

	September 11 Web Archive	Election 2002 Web Archive
URLs	30,000+	3,000+
Crawl Dates	September 11, 2001 - December 1, 2001	July 1, 2002 – November 30, 2002
Crawler	Internet Archive/ALEXA crawler	Internet Archive/ALEXA crawler
Crawl Periodicity	Daily	Varied
Unique URLs	332,000	82,000
HTML/TEXT/RTF	193K (60%)	48K (59%)
IMAGES	127K (38%)	29K (35%)
Size of collection	5 TB	1 TB

In : **SCHNEIDER** S.M., **FOOT** K., **KIMPTON** M.**, JONES** G., *Building Thematic Web Collections: Challenges and Experiences from the September 11 web Archive and the Election 2002 Web Archive*. **In** : KOCH T., SØLVBERG I.T. Eds. Research and advanced technology for digital libraries : 7th European conference, ECDL 2003, Trondheim, Norway, August 17-22, 2003, p.79

Les collaborations communes pour l'archivage.

- En Europe septentrionale : Nordic Web Archiv.

Les bibliothèques nordiques sont d'accord pour dire qu'il est impossible pour le moment de faire une loi sur un dépôt légal du Web et qui plus est d'harmoniser les législations de plusieurs pays. Le Nordic Web Archiv est la réunion de plusieurs projets d'archivage du web pour les pays nordiques. Ces groupes de bibliothèques veulent développer des outils d'archivage communs et rationaliser les initiatives locales dans un ensemble plus vaste.

Les Bibliothèques Nationales du Danemark, de Finlande, de Norvège, d'Islande et de Suède ont coopéré pour développer le projet Nordic Web Archive (NWA). Le but est de développer une solution logicielle pour accéder aux archives du Web. Il s'agit d'exporter le contenu des archives obtenues vers un format commun, d'indexer toute cette production et d'utiliser un moteur de recherche pour permettre une recherche efficace. Une interface Web pour la recherche et la navigation sera développée. Le logiciel sera disponible en qualité de logiciel open source. Le NWA réfléchit également au problème du dépôt de ces sites et tente de trouver une solution pour harmoniser leur législation dans ce sens.

- En Europe Occidentale : NEDLIB (Networked European Deposit Library).

Ce projet regroupe les grandes bibliothèques nationales (BnF, Norvège, Finlande, Allemagne, Pays-Bas, Portugal, Suisse, Italie), le chef de projet est la Bibliothèque Royale des Pays-Bas. NEDLIB a été crée en 1998, son objectif est de trouver une solution au dépôt légal des documents électronique par mise en place d'un système fondé sur l'OAIS.

Les projets concernant les sites fédérateurs.

Il s'agit ici surtout de projets qui visent à fournir un meilleur accès à des ressources scientifiques et culturelles et à harmoniser le référencement des sites fédérateurs. Les principaux projets en Europe sont nordiques comme NISBIG (*Nordic interconnected Subject Information Gateway*), et anglo-saxons tels DESIRE (*Development of a European Service for Information on Research and Education*) ou Renardus. DESIRE vise à faciliter l'accès à de l'information validée pour les chercheurs en promouvant le modèle des points d'accès par sujet (comme pour Bubl)[68]. Le projet Renardus s'inscrit aussi dans la même logique avec cette réflexion sur la création de métadonnées pour rendre les sites fédérateurs cohérents entre eux.

Les sites fédérateurs.

Les sites généralistes.

Bubl : <http://www.bubl.ac.uk>
NISS (*National Information Services and System*): <http://www.niss.ac.uk>
Open Directory Project :
<http://dmoz.org/Computers/Internet/WWW/Directories/Open_Directory_Project>
Signets de la BnF: <http://www.portail.culture.fr/sdx/pic/culture/int/index.htm>
RDN (*Ressource Discovery Network*) : <http://www.rdn.ac.uk>
Librarian's Index to the Internet (Lii): <http://www.lii.org>

Les sites thématiques.

ADAM (Art Design Architecture and Media Information Gateway) : <http://adam.ac.uk>

Agrigate (agriculture) : <http://www.agrigate.edu.au/>

BIOME (médecine) : <http://biome.ac.uk/biome.html>

BizEd (économie) : <http://bized.ac.uk>

Chemdex (pharmaceutique): <http://www.chemdex.org>

CisMef (médecine) : <http://www.cismef.org>

EEVL (*Edinburgh Engineering Virtual Library*, ingéniérie) : <http://www.eevl.ac.uk>

Geoguide (géographie) : <http://www.geo-guide.de>

History Guide (histoire) : <http://www.historyguide.de>

Mathguide (mathématiques) : <http://www.mathguide.de>

OMNI (*Organising Medical Networked Information*): <http://omni.ac.uk>

RIME (économie) : <http://www4.ccip.fr/_rime/>

SIBEL (Sciences de l'Information et Bibliothèque en Ligne) : <http://sibel.enssib.fr>

SOSIG (sciences sociales) : <http://www.sosig.ac.uk/>

[68] Des sites fédérateurs comme SOSIG, DutchESS et EELS ont été partenaires au sein de ce projet.