Le dépôt légal de l'Internet :

Principes et réalisations

Gildas ILLIEN
Chef de projet Dépôt légal d'Internet
Département de la Bibliothèque numérique, Bibliothèque nationale de France
gildas.illien@bnf.fr

Introduction

- L 'archivage d 'Internet : un enjeu de société et de mémoire;
- A la BnF, il s 'inscrit dans la continuité juridique et patrimoniale des missions de dépôt légal;
- Après une phase d 'expérimentation, le projet de la BnF est entré en phase de production.

Plan

- **1. Les principes directeurs**
  - **Le cadre juridique**
  - **Les principes techniques**
  - **Les principaux enjeux**
- **2. Bilan d'étape et perspectives**
  - **Mise en pratique du « modèle intégré »**
  - **Les collections en chiffres**
  - **Les projets en cours et à venir**

1.1 Le cadre juridique du dépôt légal d'Internet

Titre IV de la Dadvsi (1er août 2006)

Bref historique du dépôt légal :

- **1537** : les livres
- **1648** : les estampes, dont les cartes et plans
- **1793** : les partitions musicales
- **1925** : les photographies, arts graphiques de toute nature
- **1938** : les phonogrammes
- **1941** : les affiches
- **1975** : les vidéogrammes et les documents multimédias
- **1977** : les œuvres cinématographiques
- **1992** : les documents audiovisuels de la radio télévision, l'édition électronique sur support (progiciels, bases de données, systèmes experts).

- **2006 : Internet et les sites web**

5

Qu'est-ce que le dépôt légal?

- **CODE DU PATRIMOINE 20/02/2004 - Chapitre I : Objectifs et champ d'application**

- **Article L131-1 - Le dépôt légal a pour objet :**
  - la **collecte et la conservation** des documents;
  - la constitution et la diffusion de **bibliographies nationales**;
  - la **consultation** des documents, sous réserve des secrets protégés par la Loi, dans les conditions conformes à la législation sur la propriété intellectuelle et compatibles avec leur conservation.
- **Article L131-2 - Le dépôt légal concerne :**
  - Les documents imprimés, graphiques, photographiques, sonores, audiovisuels, multimédias, dès lors qu'ils sont mis à la disposition d'un public
  - Les progiciels, les bases de données, les systèmes experts et les autres produits de l'intelligence artificielle,dès lors qu'ils sont mis à la disposition du public par la diffusion d'un support matériel

6

Le dépôt légal d'Internet dans la Dadvsi :

- Le titre IV de la loi DADVSI votée le 1er août 2006 prévoit l'extension du dépôt légal à tous " **les signes, signaux, écrits, sons ou messages de toute nature qui font l'objet d'une communication au public par voie électronique** ". Les sanctions pénales pour non respect de cette obligation n'entreront toutefois pas en vigueur avant un délai de 3 ans. Un **décret d'application** viendra préciser les conditions de sélection et de consultation des informations collectées.

- **l'Institut national de l'Audiovisuel** collectera les sites du domaine de la communication audiovisuelle (en particulier ceux de la radio et de la télévision) et la **Bibliothèque nationale de France** tous les autres.

- L'obligation de dépôt légal pèse sur les personnes qui éditent et produisent des sites Internet sur le territoire français. Contrairement à ce qui est pratiqué pour les autres supports, **elle n'implique pas de démarche particulière de leur part** car la collecte est principalement effectuée par le biais de collectes automatiques réalisées par des robots que pilotent les institutions dépositaires.

- La seule obligation qui incombe aux producteurs est de **fournir les codes et les informations techniques susceptibles de faciliter l'archivage de leurs sites en cas de difficulté**. Une procédure de dépôt pourra en outre être mise en œuvre dans les cas où l'architecture d'un site sélectionné ou les formats utilisés rendraient impossible la collecte automatique.

- Le décret devrait autoriser la **consultation** des archives de la Toile par des chercheurs dûment accrédité, **dans les seules emprises de la BnF** (salles de recherche), comme pour les autres collections issues du dépôt légal.

7

1.2 Principes techniques pour l'archivage du Web

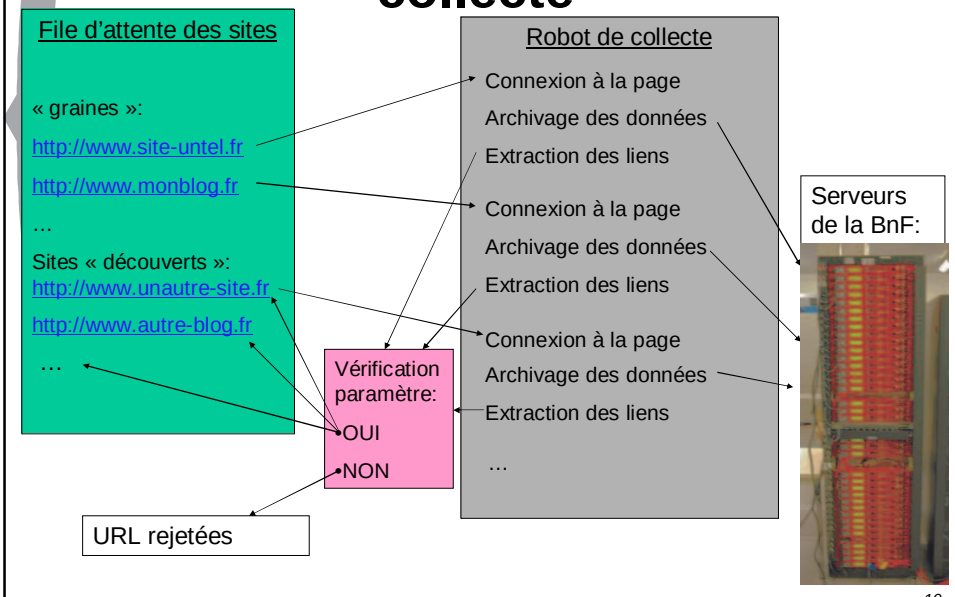
8

Pour archiver le Web, il faut :

- Une **liste de sites** (URL) ou « graines » qui servent de point de départ à la collecte;
- Un **robot de collecte** qui fonctionne comme un internaute automatique (sans carte bleue ni mot de passe) et qui aspire le Web de lien en lien, en profondeur et en largeur… tant qu'il ne rencontre pas d'obstacles;
- De la **bande passante** et des **serveurs** de forte capacité pour la collecte, l'indexation, le stockage;
- Des **compétences spécialisées** et une organisation dédiée, structurée comme une chaîne de production numérique.

9

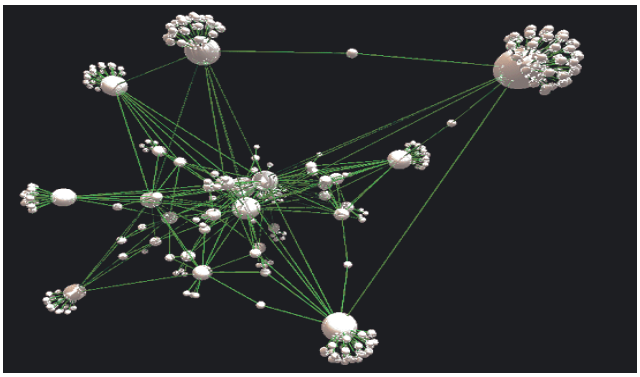
Fonctionnement du robot de collecte



10

Important : on archive le Web comme un tissu de liens

Les unités documentaires sont des sites
Les liens font partie des contenus des sites.

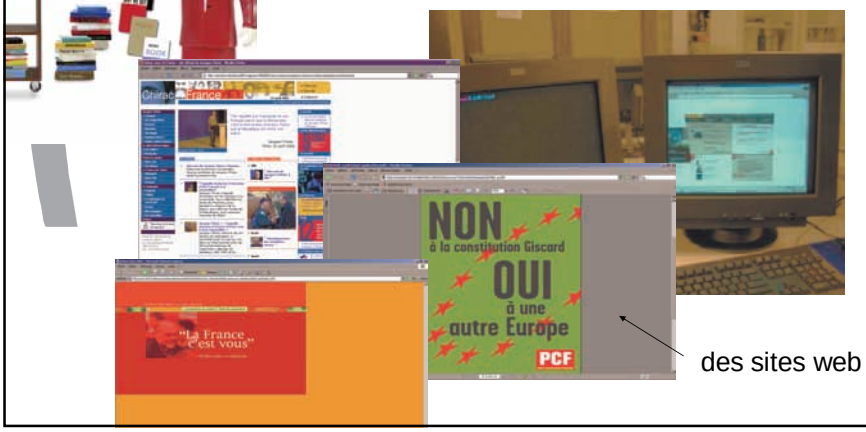


11

L'archivage du Web en pratique (1)

des bibliothécaires

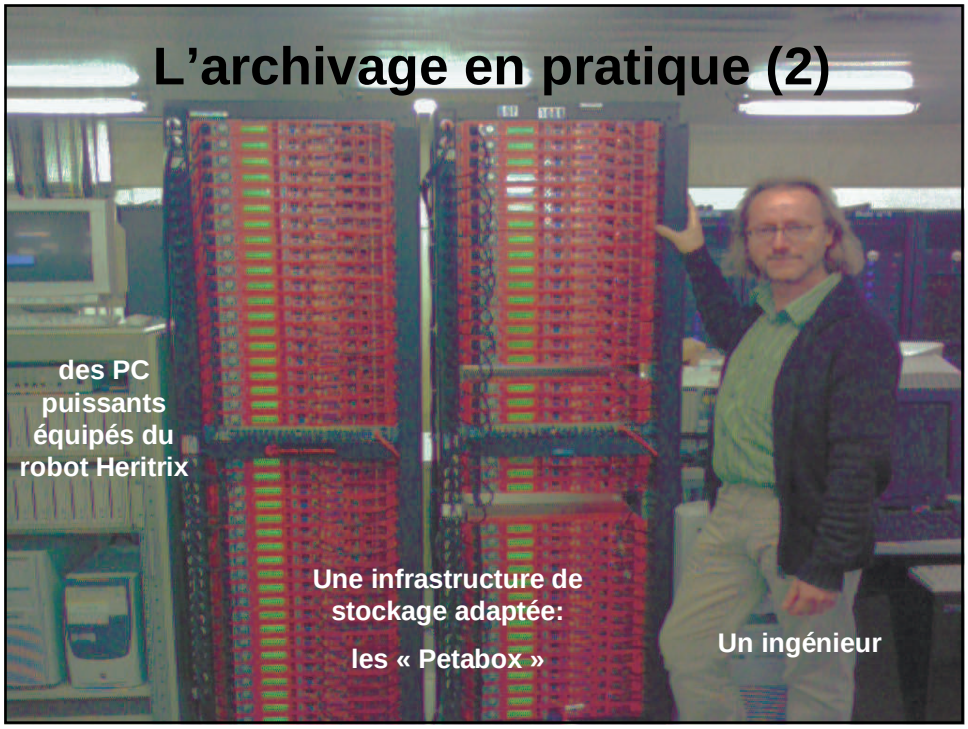
Une interface de saisie



des sites web

12

L’archivage en pratique (2)



Une archive du Web, c’est plein de trous…

Ex : archive du site du Figaro du 12/01/2004

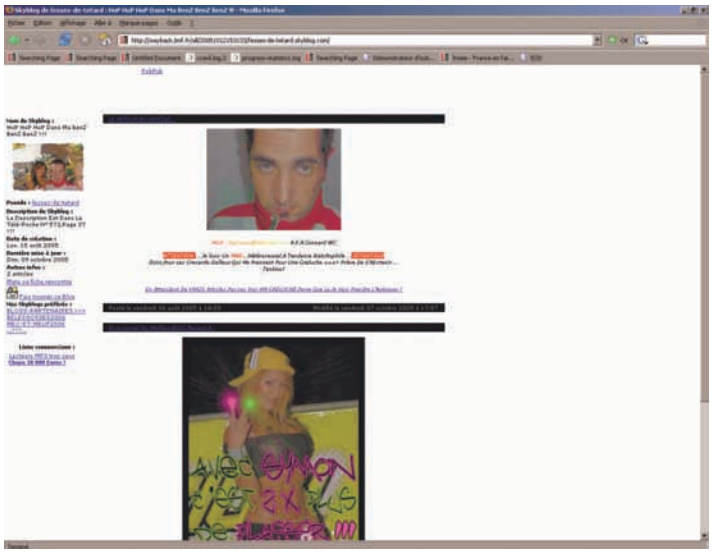
14

Une archive du Web, ce n'est pas seulement des revues savantes…

Ex : http://bubulle-attitude.skyblog.com

15

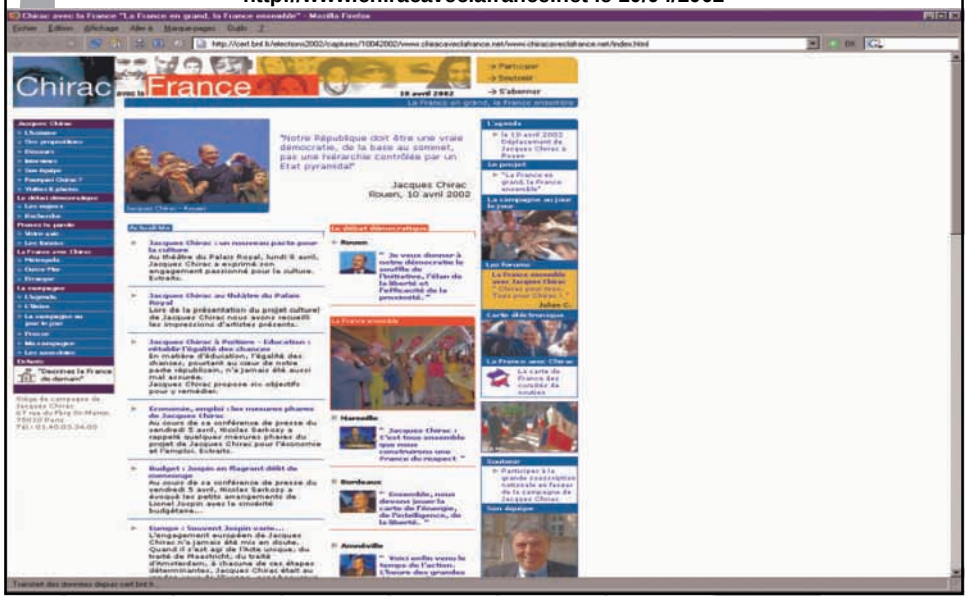
HoP HoP HoP Dans Ma BenZ BenZ BenZ !!!, *12/10/2005*



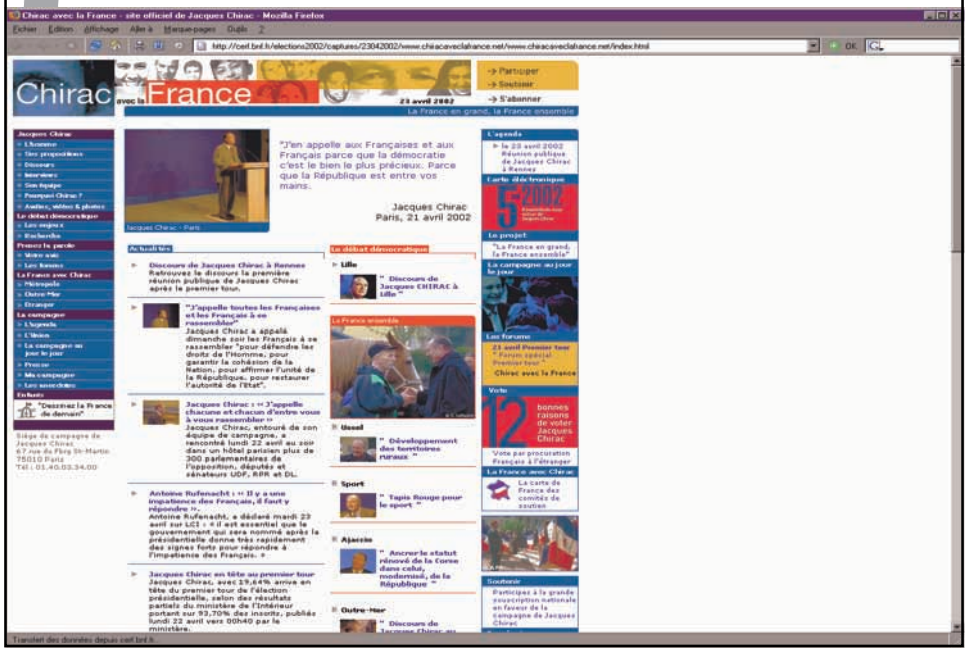
16

Un site Web, c'est souvent éphémère…

Ex : Site de campagne de Jacques Chirac
http://www.chiracaveclafrance.net le 10/04/2002



Site de campagne de Jacques Chirac
http://www.chiracaveclafrance.net le 24/04/2002



Site de campagne de Jacques Chirac
http://www.chiracaveclafrance.net aujourd'hui

1.3 Les principaux enjeux de la mise en place du dépôt légal d'Internet par la BnF

20

Les principaux enjeux pour la BnF:

- La **masse** : plus de 700 000 sites rien que pour le .fr début 2007
- L **'hétérogénéité et l'instabilité** : des formats, des fichiers... Comment assurer la conservation pérenne de tous ces documents?
- Les **frontières :** qu'est-ce qu'un Web « national »? qu'est-ce qu'un site public / privé ? qu'est-ce qu'un site « artistique »?
- La **profondeur et la complétude** : il faut pouvoir définir la profondeur, la largeur et les fréquences des collectes, identifier aussi les contenus inaccessibles aux robots pour des raisons techniques ou d'authentification
- de **l 'exhaustivité à la représentativité** : collecter le meilleur et le pire ? Quels critères de sélection? La question des savoirs légitimes au crible des principes du dépôt légal
- **unités documentaires** : qu'est-ce qu'on compte? Les unités possibles : URL (fichiers), Host (serveurs), domaines, ARC, To…

21

Implications et évolution métier :

- Changer d'échelle : l'indispensable automatisation des processus (de collecte, de versement, d'indexation, de stockage, de conservation)
- Appréhender le niveau logique et le niveau physique du document : connaître l'Internet dans ces deux dimensions, comprendre et piloter les robots en conséquence
- Repenser les principes d 'enrichissement des collections : redéfinir les périmètres et les pratiques
- Inventer de nouveaux modèles d 'organisation et de coopération: à la BnF, au niveau national, régional, international
- Nouvelles contraintes et nouvelles opportunités pour la recherche : frustrations du puzzle et nouvelles opportunités d 'analyse (ex : *data mining, link analysis, trend analysis…)* = nécessité de repenser les usages et les accès

22

Les principaux enjeux pour les éditeurs et les producteurs:

- **Le droit à la mémoire?**
- Rien ne garantit plus qu'une œuvre sera collectée et conservée par la BnF
- **Le droit à l'oubli ? Le droit à l'éphémère?**
- La loi ne permet pas aux producteurs de retirer une publication des collections issues du dépôt légal
- **L'inversion du rapport producteur / bibliothèque**
- Ce ne sont plus les artistes qui déposent, c'est la bibliothèque qui collecte, en masse
- **Que faire pour faciliter son propre archivage?**
- Penser à « l'archivabilité » de son site, c'est penser son accessibilité (choisir des formats ouverts et des protocoles conformes aux recommandations du W3C)

23

2. Bilan d'étape et perspectives

24

2.1. la mise en pratique du « modèle intégré »

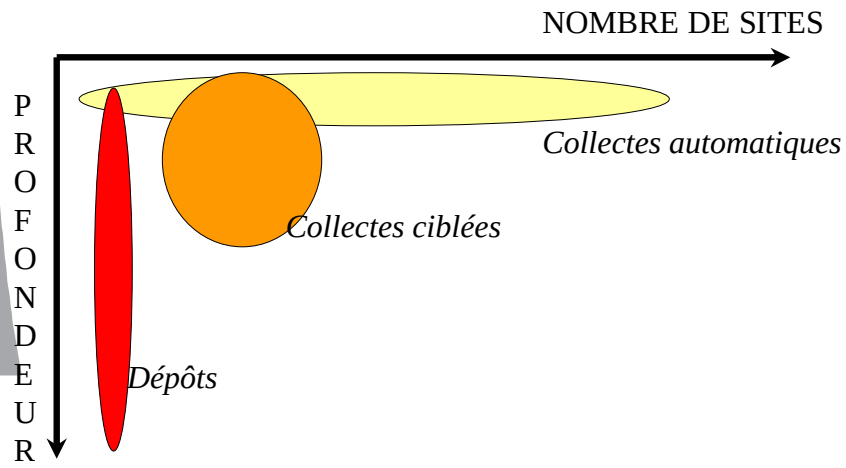
25

Le modèle intégré

- Des **collectes automatiques** à grande échelle (instantanés), réalisées sur le .fr et des domaines génériques (.com, .net…) périodiquement
- Des **collectes ciblées** (campagnes thématiques et événementielles) pilotées par des bibliothécaires.
  Ex : campagnes électorales sur Internet
- Des **dépôts** à l 'unité organisés avec les producteurs, dans des cas exceptionnels
  Ex : la version électronique du *Journal officiel*

26

Le modèle intégré: schéma



27

Réalisations

- Collectes larges : instantanés du domaine national « .fr » en partenariat avec Internet Archive = 2004, 2005, 2006, 2007 (6 semaines)
- Collectes ciblées :

  - collectes électorales = 2002, 2004, 2007

  - Le projet « Internet en campagne » : une mobilisation sans précédent avec des partenaires extérieurs : Sciences Po, le Forum des droits sur Internet, 8 bibliothèques de dépôt légal imprimeur en région.

  - collectes multi-sujets = 2005, 2006 : 70 bibliothécaires mobilisés dans tous les départements de collections.
- Dépôts à l'unité :

  - une trentaine d'expérimentations

  - Le Journal officiel électronique

28

2.2 Les collections en chiffres

- Collections rétrospectives (1996-2005) acquises auprès d'Internet Archive = les plus gros volumes
- Instantanés du domaine.fr et collectes ciblées intégrées (2004-2006) = 3 à 4 To pour chaque instantané
- Collectes électorales :

  - Elections 2002-2004 : 3500 sites ; 12617 captures ; 23 millions de fichiers ; 535 Go
  - Elections 2007 : 2700 sites, 35 millions de fichiers, 2,3 To

AU TOTAL :

- **10 milliards de fichiers**
- **130 To de données**

29

2.3 Agenda 2007-2008

- Une nouvelle collecte large sera réalisée par Internet Archive en octobre 2007 (instantané du .fr) depuis la BnF (ingénieur *in situ*). Objectif d'internalisation complète des processus, partenariat possible avec l'AFNIC.
- Couverture de la campagne électorale depuis octobre 2006 jusqu'en juillet 2007: *« Internet en campagne » (présidentielle puis législatives)*, projet pilote pour l'Internalisation de l'infrastructure de collecte (workflow complet) = internalisation de l'ensemble des collectes ciblées.
- Publication en juin 2007 des résultats de l'étude d'usages conduite en partenariat avec Sciences Po d'octobre 2006 à février 2007 : *première rencontre avec les usagers* qui prélude à la mise en place des accès et des outils de consultation dans les salles de lecture fin 2007 / début 2008 selon décret d'application Dadvsi
- Formalisation de la politique documentaire et de l'organisation, travail sur l'évolution des métiers et la formation des personnels (réseau des 70 correspondants DL Internet)
- Versement et gestion des archives Web dans le nouvel entrepôt numérique (*digital repository*) de la BnF à l'horizon 2008-2009
- Coopération nationale et internationale : le réseau des BDLI et celui des pôles associés, le consortium IIPC

30

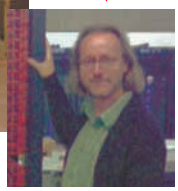
L'équipe BnF
Dépôt légal d'Internet

Younès Hafri (DSI)
Indexation et développement

Igor Ranitovic
Internet Archive

John Lee
Internet Archive

Bert Wendland (DSI)
Exploitation et crawls



France Lasfargue (DBN)
Qualité – Usages

Clément Oury (DBN)
Collections - Préservation

Gildas Illien (DBN)
Chef du projet

Sara Aubry (DBN)
Spécifications - Accès

31

Pour en savoir plus :

- Les dossiers de presse *Les enjeux du dépôt légal de la Toile et « Internet en campagne »* sur le site bnf.fr : http://www.bnf.fr/pages/presse/comm_etabliss.htm
- La rubrique « Informations professionnelles du site bnf.fr : http://www.bnf.fr/pages/infopro/depotleg/dli_intro.htm
- Le texte de loi DADVSI sur le site de Legifrance http://www.legifrance.gouv.fr/WAspad/UnTexteDeJorf?numjo=MCCX0300082L
- Le site d'IIPC (International Internet Preservation Consortium): http://www.netpreserve.org

32


INTERNET ARCHIVE
WayBackMachine
DES QUESTIONS ?
Take Me Back!
33