Conserver la mémoire de l'internet

La préservation du dépôt légal de la Toile à la Bibliothèque nationale de France

Clément OURY
Chargé de projet pour le Dépôt légal de l'internet
Département de la Bibliothèque numérique, Bibliothèque nationale de France
clement.oury@bnf.fr

1- Introduction

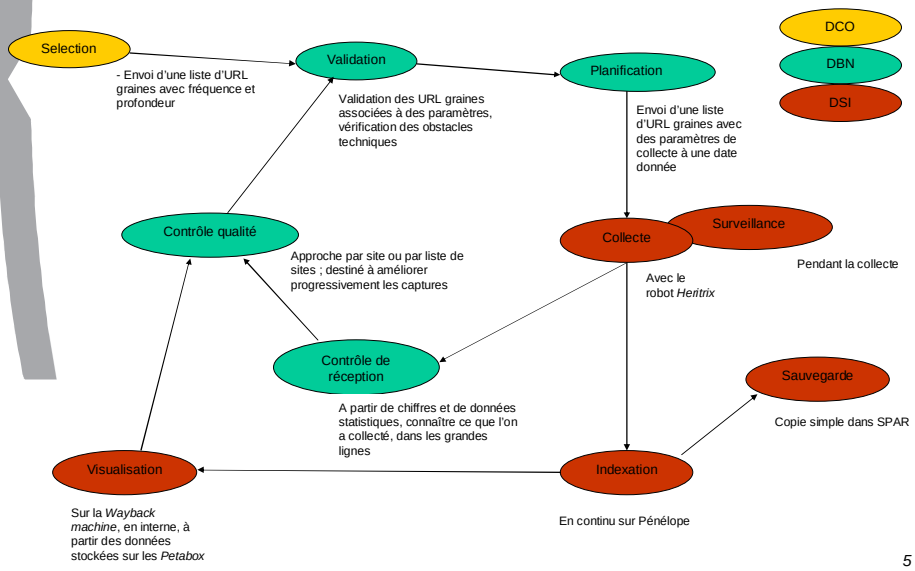
- Etat des lieux du processus
- Quels formats pour l'archivage de la Toile ?
- Les perspectives d'archivage à long terme

Mise en place du processus et passage à l'échelle

Un projet dans une perspective plus large

- Pourquoi collecter les sites électoraux ?
- Objectif recherché: l'internalisation des collectes larges
  - Réaliser un circuit complet de collecte
  - Identifier les problèmes et les besoins
- Le Web électoral, petite image de la Toile
  - Modes de publication variés…
  - … et format complexes

Le processus de collecte



5

Outil de saisie – la sélection

:: http://desirdavenir-mosaique.over-blog.com ::

Les champs marqués d'une étoile sont obligatoires

* Type d'élection	élections présidentielles
Notes	
* Type de site	1.3 autres organisations de soutien
Candidat (Nom, Prénom)	Royal, Ségolène
Formation politique	PS - Parti socialiste
Adresse e-mail ou URL d'abonnement à la lettre d'information	
* Fréquence	2 - une fois par mois
Date de capture précise (JJ/MM/AAAA)	
* Profondeur	hôte (host)

Enregistrer la proposition

Outil de saisie – la validation

Adresse e-mail ou URL d'abonnement à la lettre d'information

* Fréquence: 4 - une seule fois

Date de capture précise (JJ/MM/AAAA)

* Profondeur: domaine

URL à inclure (un par ligne)

URL à exclure (un par ligne, troncature possible)

.*desirsdavenir-mosaique.over-blog.com/archive.*

Motif de rejet éventuel

Commentaires ADMINISTRATEUR

Valider la proposition | Rejeter la proposition

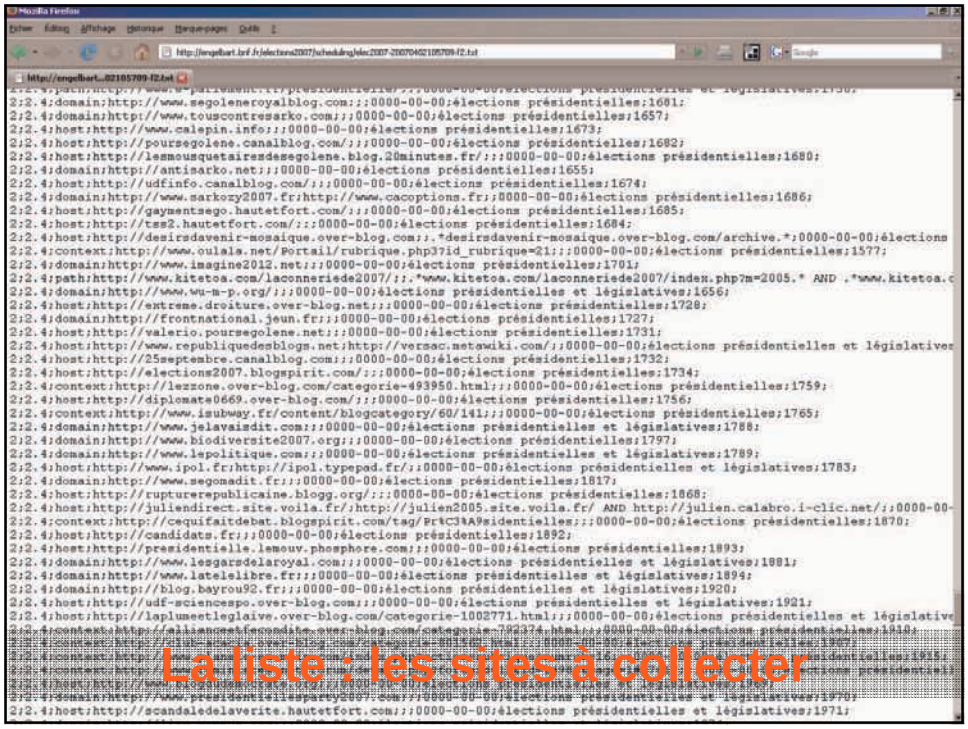
Calendrier des collectes

ELECTIONS PRESIDENTIELLES : planning des collectes

ELECTIONS LEGISLATIVES 2007 : planning des collectes

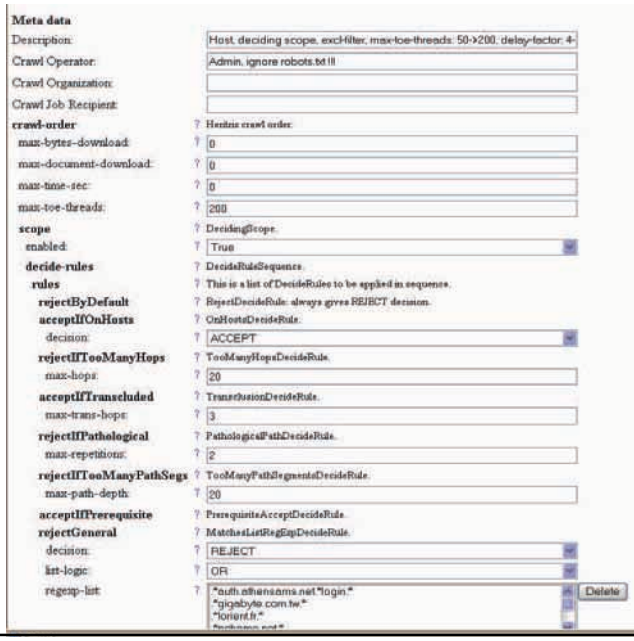
Mars	Presid	Avril	Presid	Legis	Mai	Presid	Legis	Juin	Legis	Juillet	Legis
Jeudi 1	F2 presid+	Dimanche 1			Mardi 1	Férié	Férié	Vendredi 1		Dimanche 1	
Vendredi 2		Lundi 2	F2 presid +		Mercredi 2	F1 presid+	F1 legis	Samedi 2		Lundi 2	F2 legis+
Samedi 3		Mardi 3	F1 presid+	F1 legis	Jeudi 3	F2 presid +		Dimanche 3		Mardi 3	F1 legis+
Dimanche 4		Mercredi 4			Vendredi 4	F3 presid +		Lundi 4	F2 légis+	Mercredi 4	
Lundi 5	F1 presid+	Jeudi 5			Samedi 5			Mardi 5	F1 et F3 legis +	Jeudi 5	
Mardi 6		Vendredi 6			Dimanche 6			Mercredi 6		Vendredi 6	
Mercredi 7		Samedi 7			Lundi 7	F1 presid+ F3 presid+	F1 legis	Jeudi 7		Samedi 7	
Jeudi 8		Dimanche 8			Mardi 8	Férié	Férié	Vendredi 8		Dimanche 8	
Vendredi 9		Lundi 9	Férié	Férié	Mercredi 9		F2 legis	Samedi 9		Lundi 9	Ci-devant pres
Samedi 10		Mardi 10	F1 presid+	F1 legis	Jeudi 10			Dimanche 10		Mardi 10	F1 legis+
Dimanche 11		Mercredi 11			Vendredi 11			Lundi 11	F1 legis+	Mercredi 11	
Lundi 12	F1 presid+	Jeudi 12			Samedi 12			Mardi 12	F3 legis+	Jeudi 12	
Mardi 13		Vendredi 13			Dimanche 13			Mercredi 13	F2 légis+	Vendredi 13	
Mercredi 14	date de capture	Samedi 14			Lundi 14	F4 presid +		Jeudi 14		Samedi 14	
Jeudi 15		Dimanche 15			Mardi 15	F1 presid+	F1 legis	Vendredi 15		Dimanche 15	
Vendredi 16		Lundi 16	F3 presid +	F1 legis	Mercredi 16			Samedi 16		Lundi 16	
Samedi 17		Mardi 17	F1 presid +		Jeudi 17	Férié	Férié	Dimanche 17		Mardi 17	
Dimanche 18		Mercredi 18		F2 legis	Vendredi 18			Lundi 18	F1 legis+	Mercredi 18	
Lundi 19	F1 presid+	Jeudi 19			Samedi 19			Mardi 19	F3 legis+	Jeudi 19	
Mardi 20		Vendredi 20			Dimanche 20			Mercredi 20		Vendredi 20	
Mercredi 21	lancement	Samedi 21			Lundi 21	F1 presid+		Jeudi 21		Samedi 21	
Jeudi 22	Européens	Dimanche 22			Mardi 22		F1 legis	Vendredi 22		Dimanche 22	
Vendredi 23	Interdiction	Lundi 23	F3 presid +		Mercredi 23			Samedi 23		Lundi 23	
Samedi 24	de	Mardi 24	F1 presid+	F1 legis	Jeudi 24			Dimanche 24		Mardi 24	
Dimanche 25	collecter	Mercredi 25			Vendredi 25			Lundi 25	F4 legis+	Mercredi 25	
Lundi 26	F1 presid+	Jeudi 26	essai dailymotion		Samedi 26			Mardi 26	F1 legis+	Jeudi 26	
Mardi 27		Vendredi 27			Dimanche 27			Mercredi 27		Vendredi 27	
Mercredi 28		Samedi 28			Lundi 28	Férié	Férié	Jeudi 28		Samedi 28	
Jeudi 29		Dimanche 29			Mardi 29	F1 presid+	F1 legis	Vendredi 29		Dimanche 29	
Vendredi 30		Lundi 30			Mercredi 30			Samedi 30		Lundi 30	
Samedi 31					Jeudi 31					Mardi 31	

8



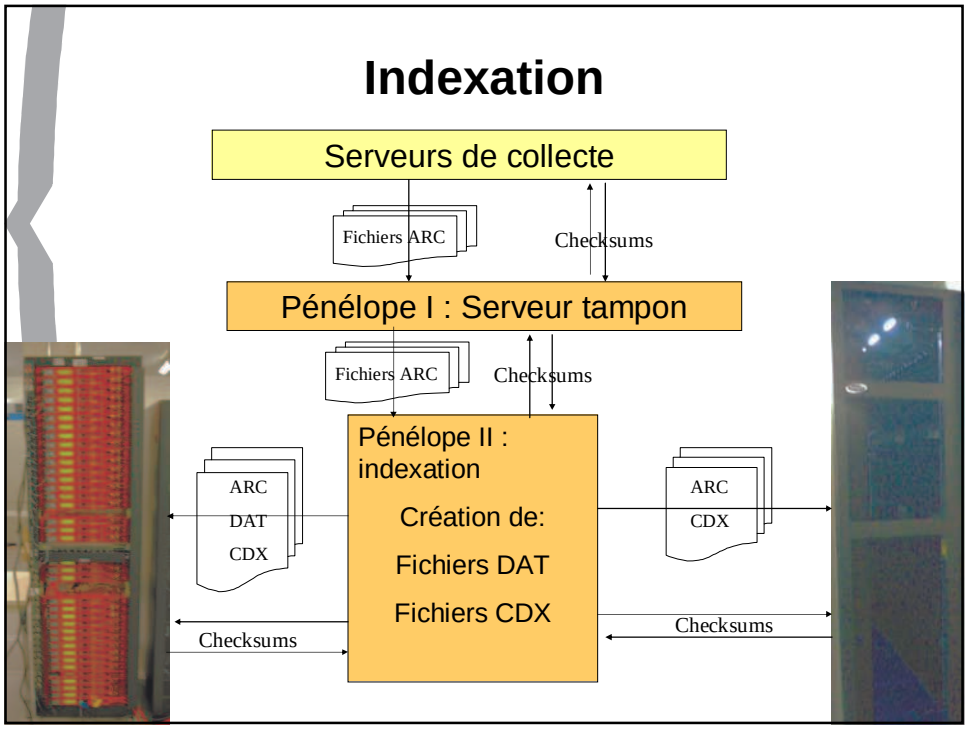
La liste : les sites à collecter

Configurer une collecte



Surveiller la collecte

Indexation



Sauvegarde

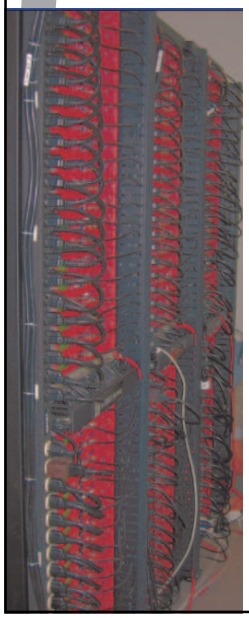
SPAR :

Entrepôt numérique

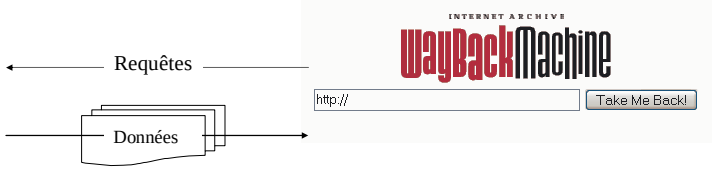
Préservation du train de bits

13

Accès



Une architecture de stockage à grande échelle : les Petabox



Accès avec la Wayback Machine

14

La problématique des formats

15

Des formats spécifiques : le format ARC

- Un fichier ARC, un fichier *container*
  - Un fichier ARC regroupe de multiples enregistrements ARC
  - Un enregistrement ARC est le fichier collecté sur le Web, accompagné de métadonnées
- Autres caractéristiques du format ARC
  - format auto descriptible
  - format extensible
  - format compressé
  - taille limitée arbitrairement à 100 Mo

16

Un enregistrement ARC

URL IP-address Archive-date Content-type Archive-length

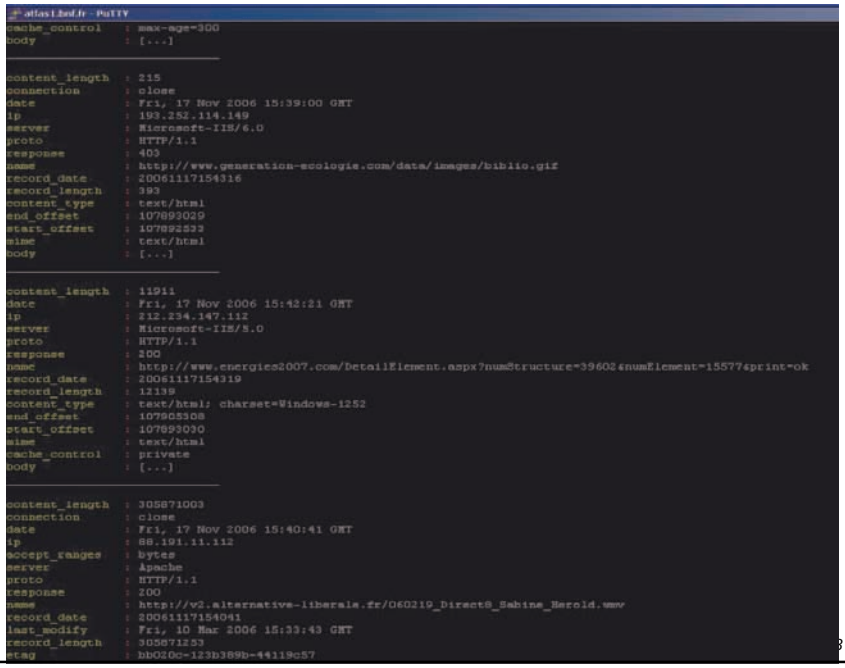
Métadonnées: http://desirsdavenir-mosaique.over-blog.com/ 195.20.15.131 20070416172243 text/html 132721

Objet numérique collecté :

HTTP/1.1 200 OK
Date: Tue, 16 Apr 2007 17:22:56 GMT
Server: Apache/2.0.58 (Unix) mod_ssl/2.0.58 OpenSSL/0.9.7e PHP/4.4.2
X-Powered-By: PHP/4.4.2
Last-Modified: Tue, 16 Apr 2007 17:22:56 GMT
Content-Type: text/html

<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" lang="fr" xml:lang="fr">
<head>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1" />
<title>Comité Mosaïque</title>
<meta name="description" content="desirsdavenir-mosaique.over-blog.com hébergé par over-blog.com Le comité de la diversité riche de ses différences qui soutient Ségolène Royal à la présidence de la République." />
...

Un fichier ARC



Des formats spécifiques

- Les autres formats :
  - Fichiers d'index
  - Fichiers de métadonnées
- Des formats maîtrisés

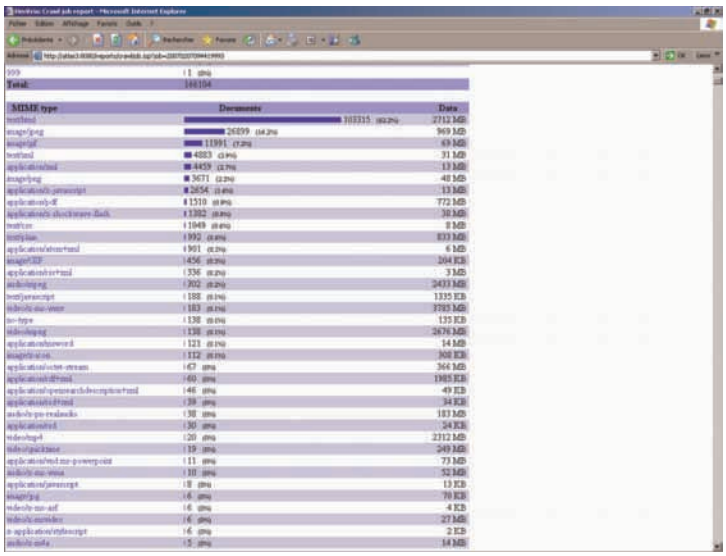
19

Des formats empaquetés

- Des formats multiples…
- … et très mal identifiés
- Mais cinq grands formats dominent : HTML, JPEG, GIF, PDF, plain-text

20

La variété des fichiers



MIME type	Documents	Data
text/html	103335 (62.2%)	2712 MB
image/jpeg	26899 (16.2%)	969 MB
image/gif	11991 (7.2%)	69 MB
text/xml	4883 (2.9%)	31 MB
application/xml	4459 (2.7%)	13 MB
image/png	3671 (2.2%)	40 MB
application/x-javascript	2654 (1.6%)	13 MB
application/pdf	1510 (0.9%)	772 MB
application/x-shockwave-flash	1382 (0.8%)	30 MB
text/css	1049 (0.6%)	9 MB
text/plain	992 (0.6%)	833 MB
application/atom+xml	901 (0.5%)	6 MB
image/x-ico	456 (0.3%)	264 KB
application/rss+xml	356 (0.2%)	3 MB
audio/mpeg	302 (0.2%)	2433 MB
text/javascript	188 (0.1%)	1335 KB
audio/x-ms-wma	183 (0.1%)	3785 MB
no-type	138 (0.1%)	135 KB
video/mpeg	128 (0.1%)	2676 MB
application/msword	121 (0.1%)	14 MB
image/x-icon	112 (0.1%)	308 KB
application/octet-stream	67 (0%)	364 MB
application/rdf+xml	60 (0%)	1985 KB
application/opensearchdescription+xml	46 (0%)	49 KB
application/x-rdf+xml	39 (0%)	34 KB
audio/x-pn-realaudio	38 (0%)	183 MB
application/vnd	30 (0%)	24 KB
video/mp4	20 (0%)	2712 MB
video/quicktime	19 (0%)	249 MB
application/vnd.ms-powerpoint	11 (0%)	73 MB
audio/x-ms-wma	10 (0%)	52 MB
application/javascript	8 (0%)	13 KB
image/jpg	6 (0%)	70 KB
video/x-ms-asf	6 (0%)	4 KB
video/x-msvideo	6 (0%)	27 MB
x-application/rtf	6 (0%)	2 KB
audio/aiff	5 (0%)	14 MB

21

Comment assurer la pérennité de la mémoire du Web ?

22

Décrire et empaqueter

- Décrire le processus
- Recenser les données
- Identifier les niveaux de granularité
  - Cadre de la collecte / commande / job / fichier ARC / enregistrement ARC
- Questions ouvertes :
  - Que doit-on conserver ?
  - Comment organiser les informations ?

23

Un format conçu pour la préservation : le format WARC

- Un format bâti sur le modèle du format ARC
- Enrichi de nouvelles fonctionnalités:
  - archivage de l'ensemble des interactions entre le robot de collecte et le serveur sollicité (archivage des requêtes du robot) ;
  - stockage de métadonnées à côté des données collectées ;
  - gestion des doublons ;
  - possibilité de segmenter les fichiers à collecter qui ne tiendraient pas dans un seul fichier WARC ;
  - ajout d'un identifiant pérenne ;
  - gestion des migrations des enregistrements WARC ;
  - archivage de données collectées grâce à d'autres procédures que les collectes automatiques par robot (c'est-à-dire les dépôts à l'unité).
  - taille cible : 1 Go.
- Une normalisation en cours au sein de l'ISO

24

Dans la perspective du SINUM

- Organiser les relations entre le producteur et l'Archive
- Mettre en place les procédures de validation
- Préserver l'information numérique:
  - Perspectives d'émulation
  - Perspectives de migration

25