

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

1- Introduction

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

2

Mise en place du processus et passage à l'échelle

3

Un projet dans une perspective plus large

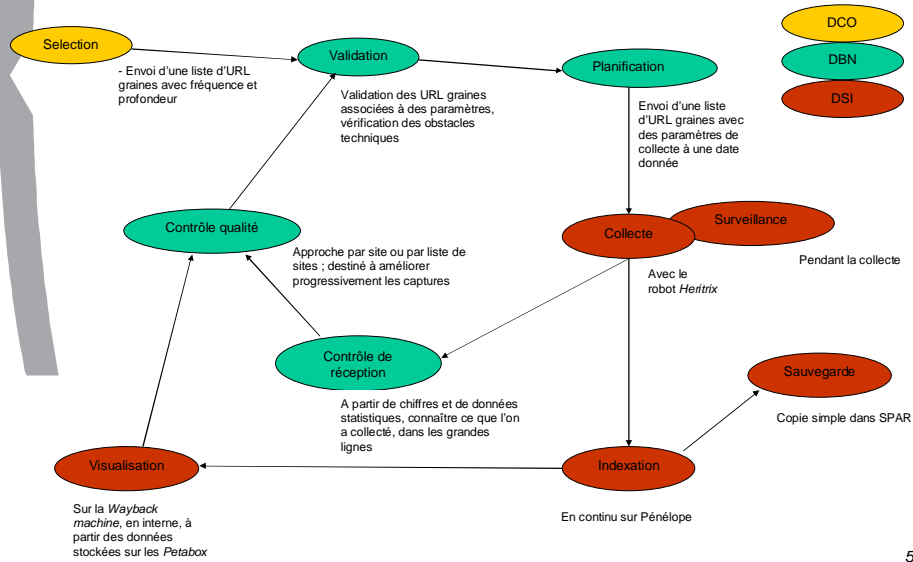
- ◆ Pourquoi collecter les sites électoraux ?

- ◆ 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

4

Le processus de collecte



Outil de saisie – la sélection

:: <http://desirdavenir-mosaïque.over-blog.com> ::

Les champs marqués d'une étoile sont obligatoires

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

Enregistrer la proposition

Outil de saisie – la validation

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

* Fréquence

Date de capture précise
(JJ/MM/AAAA)

* Profondeur

URL à inclure (un par ligne)

URL à exclure (un par ligne,
troncature possible)

Motif de rejet éventuel

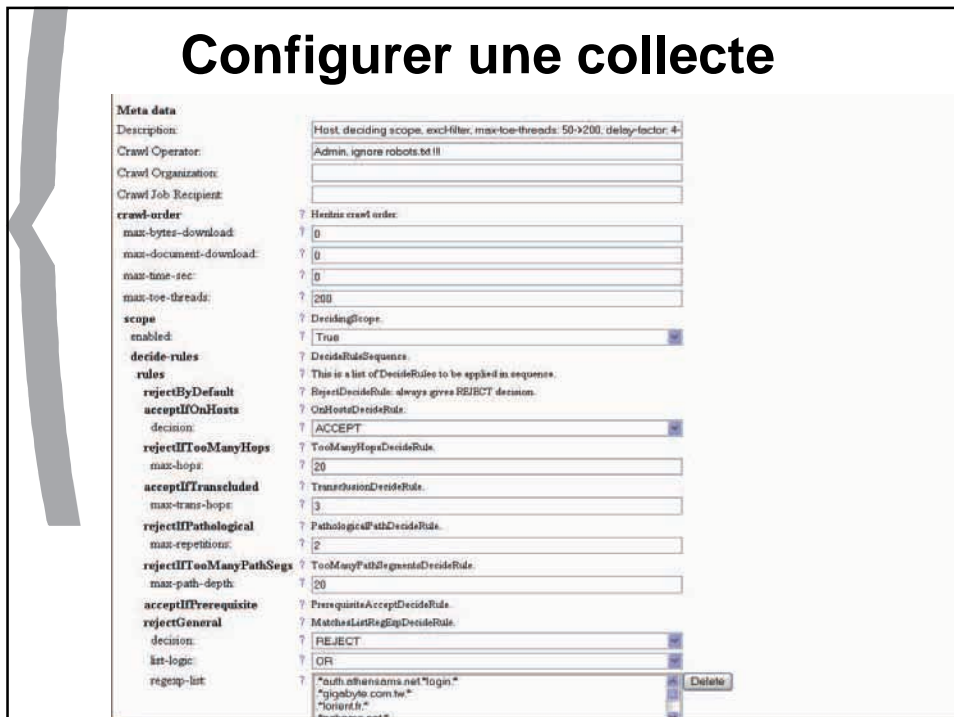
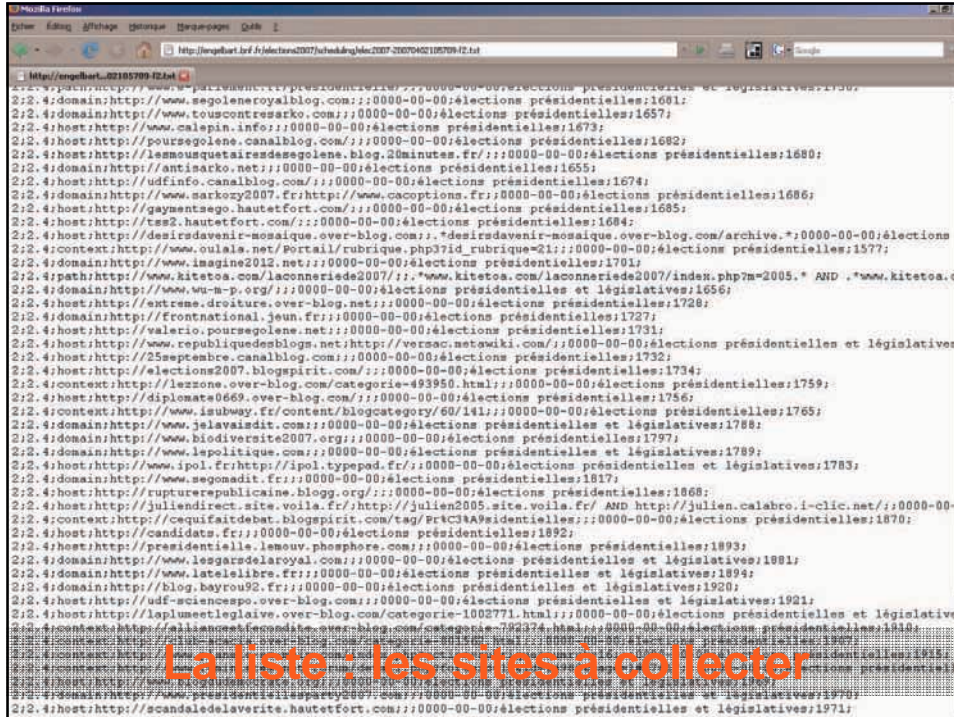
Commentaires ADMINISTRATEUR

Valider la proposition

Rejeter la proposition

Calendrier des collectes

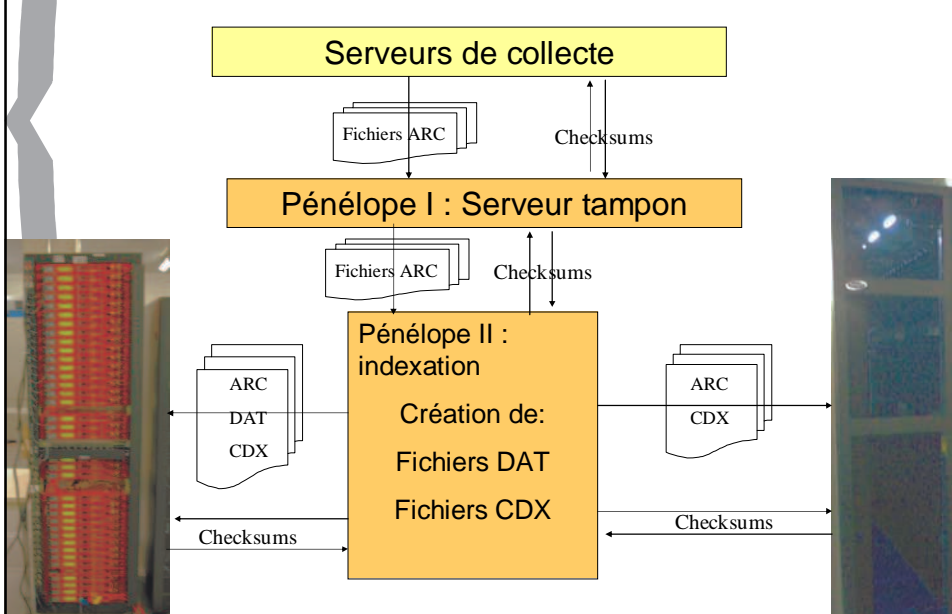
| ELECTIONS PRESIDENTIELLES : planning des collectes | | | | | | | | | | | |
|--|-----------------|-------------|------------------|----------|-------------|------------|------------|-------------|-----------------|-------------|----------------|
| ELECTIONS LEGISLATIVES 2007 : planning des collectes | | | | | | | | | | | |
| Mars | Presid | Avril | Presid | Legis | Mai | Presid | Legis | Juin | Legis | Juillet | Legis |
| Jeu 1 | F2 presid- | Dimanche 1 | | | Mai 1 | Faire | Faire | 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 | Jeu 3 | F2 presid+ | | Mardi 3 | Dimanche 3 | Mardi 3 | F1 legis+ |
| Dimanche 4 | | Mercredi 4 | | | Vendredi 4 | F3 presid+ | | Lundi 4 | F2 legis+ | Mercredi 4 | |
| Lundi 5 | F1 presid- | Jeu 5 | | | Samedi 5 | | | Mardi 5 | F1 et F3 legis+ | Jeu 5 | |
| Mardi 6 | | Vendredi 6 | | | Dimanche 6 | | | Mercredi 6 | | Vendredi 6 | |
| Mercredi 7 | | Samedi 7 | | | Lundi 7 | F1 presid- | F3 presid+ | F1 legis | Jeu 7 | | Samedi 7 |
| Jeu 8 | | Dimanche 8 | | | Mardi 8 | Faire | Faire | Vendredi 8 | | Dimanche 8 | |
| Vendredi 9 | | Lundi 9 | Faire | Faire | Mercredi 9 | | F2 legis | Samedi 9 | | Lundi 9 | CI devant pres |
| Samedi 10 | | Mardi 10 | F1 presid+ | F1 legis | Jeu 10 | | | Dimanche 10 | | Mardi 10 | F1 legis+ |
| Dimanche 11 | | Mercredi 11 | | | Vendredi 11 | | | Lundi 11 | F1 legis+ | Mercredi 11 | |
| Lundi 12 | F1 presid- | Jeu 12 | | | Samedi 12 | | | Mardi 12 | F3 legis+ | Jeu 12 | |
| Mardi 13 | | Vendredi 13 | | | Dimanche 13 | | | Mercredi 13 | F2 legis+ | Vendredi 13 | |
| Mercredi 14 | date de capture | Samedi 14 | | | Lundi 14 | F1 presid+ | | Jeu 14 | | Samedi 14 | |
| Jeu 15 | | Dimanche 15 | | | Mardi 15 | F1 presid+ | F1 legis | Vendredi 15 | | Dimanche 15 | |
| Vendredi 16 | | Lundi 16 | F2 presid+ | F1 legis | Mercredi 16 | | | Samedi 16 | | Lundi 16 | |
| Samedi 17 | | Mardi 17 | F1 presid+ | | Jeu 17 | Faire | Faire | Dimanche 17 | | Mardi 17 | |
| Dimanche 18 | | Mercredi 18 | | F2 legis | Vendredi 18 | | | Lundi 18 | F1 legis+ | Mercredi 18 | |
| Lundi 19 | F1 presid- | Jeu 19 | | | Samedi 19 | | | Mardi 19 | F3 legis+ | Jeu 19 | |
| Mardi 20 | | Vendredi 20 | | | Dimanche 20 | | | Mercredi 20 | | Vendredi 20 | |
| Mercredi 21 | lançamento | Samedi 21 | | | Lundi 21 | F1 presid+ | | Jeu 21 | | Samedi 21 | |
| Jeu 22 | Relatório | Dimanche 22 | | | Mardi 22 | | F1 legis | Vendredi 22 | | Dimanche 22 | |
| Vendredi 23 | Introdução | Lundi 23 | F3 presid+ | | Mercredi 23 | | | Samedi 23 | | Lundi 23 | |
| Samedi 24 | de | Mardi 24 | F1 presid+ | F1 legis | Jeu 24 | | | Dimanche 24 | | Mardi 24 | |
| Dimanche 25 | collecte | Mercredi 25 | | | Vendredi 25 | | | Lundi 25 | F4 legis+ | Mercredi 25 | |
| Lundi 26 | F1 presid- | Jeu 26 | seal daltymotion | | Samedi 26 | | | Mardi 26 | F1 legis+ | Jeu 26 | |
| Mardi 27 | | Vendredi 27 | | | Dimanche 27 | | | Mercredi 27 | | Vendredi 27 | |
| Mercredi 28 | | Samedi 28 | | | Lundi 28 | Faire | Faire | Jeu 28 | | Samedi 28 | |
| Jeu 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 | | Jeu 31 | | | Jeu 31 | | | Mardi 31 | | Mardi 31 | |



Surveiller la collecte

| URL | Score | Snippet |
|-------------------------|-------|---|
| 2007-04-2716103104.201E | 200 | 7453 http://dmdm.typepad.com/au_fai_de_mes_journees/WindowsLiveWriter/UnepartiedeHist_9F44/voix/nap/582445B.gif EEE |
| 2007-04-2716103104.290E | 404 | 1339 http://downloads.thepingbox.com/robots.txt EEEF http://downloads.thepingbox.com/web/wrapper.php?file=BSN10Bem |
| 2007-04-2716103104.460E | 302 | 214 http://www.marianne2007.info/robots.txt EEEF http://www.marianne2007.info/docs/images/stic-marianne-blog.gif text/ |
| 2007-04-2716103104.981E | 200 | 1837 http://code.com/lang/ndqge-senl-me-small-black.gif EEE http://www.nuesblog.com/7617/Asous-Begag-par-Christophe-Car |
| 2007-04-2716103106.176E | 200 | 62997 http://www.estvideo.com/dew/index/2007/12 EEE http://www.estvideo.com/dew/ text/html #017 2007042160204977+617 Q- |
| 2007-04-2716103106.524E | 200 | 14004 http://downloads.thepingbox.com/web/wrapper.php?file=BSN10Bem.dew EEE http://www.nuesblog.com/7617/Asous-Beg |
| 2007-04-2716103106.590E | 200 | 10231 http://dmdm.typepad.com/au_fai_de_mes_journees/styles.css EEE http://dmdm.typepad.com/text/cse #050 20070421 |
| 2007-04-2716103106.645E | 200 | 31945 http://www.estvideo.com/dew/index/2007/03/18/700-chronique-Bus-bautrage EEE http://www.estvideo.com/dew/ text/html # |
| 2007-04-2716103106.900E | 200 | 54693 http://blpwebline.blogspot.com/pollitcshw/images/ct/home.jpg E http://blpwebline.blogspot.com/pollitcshw/ image/jpeg # |
| 2007-04-2716103107.452E | 302 | 218 http://dmdm.typepad.com/au_fai_de_mes_journees/atom.xml EEE http://dmdm.typepad.com/ text/html #009 20070421 |
| 2007-04-2716103107.453E | 200 | 34618 http://www.estvideo.com/dew/index/2007/03/31/793-blindtest-0-classeique EEE http://www.estvideo.com/dew/ text/html # |
| 2007-04-2716103107.472E | 404 | 11869 http://methique.info/robots.txt EEEF http://methique.info/images/methique_web_buttons.gif text/html #048 2007042160 |
| 2007-04-2716103108.376E | 302 | 380 http://nuesblog.com/robots.txt EEEF http://nuesblog.com/preunes.jpg text/html #00 20070421603031714996 0F08ue |
| 2007-04-2716103108.379E | 404 | 272 http://www.f017d.com/robots.txt EEEF http://www.f017d.com/small/addicted.png text/html #001 200704216030387+49 |
| 2007-04-2716103108.492E | 200 | 70857 http://www.estvideo.com/dew/index/2004/07 EEE http://www.estvideo.com/dew/ text/html #09 200704216030793+472 3II |
| 2007-04-2716103109.487E | 200 | 0 http://post-it-npides.com/robots.txt EEEF http://post-it-npides.com/2/raloonoee5.jp text/plain #10 2007042160300- |
| 2007-04-2716103109.492E | 200 | 531 http://methique.info/images/methique_web_buttons.gif EEE http://www.nuesblog.com/7617/Asous-Begag-par-Christophe-Car |
| 2007-04-2716103109.891E | 200 | 220997 http://dmdm.typepad.com/au_fai_de_mes_journees/WindowsLiveWriter/LePassepoteDeWaltidreyesucbay_KO0F/wedpassp |
| 2007-04-2716103109.900E | 200 | 97694 http://www.estvideo.com/dew/index/medias EEE http://www.estvideo.com/dew/ text/html #114 200704216030949+445 QUR |
| 2007-04-2716103110.181E | 200 | 99377 http://blpwebline.blogspot.com/pollitcshw/images/2007/03/14/bove1.jpg E http://blpwebline.blogspot.com/pollitcshw/ image/jpeg # |
| 2007-04-2716103111.111E | 200 | 33510 http://www.marianne2007.info/docs/images/stic-marianne-blog.gif EEE http://www.nuesblog.com/7617/Asous-Begag-par-CI |
| 2007-04-2716103111.666E | 200 | 202074 http://www.estvideo.com/dew/index/2006/05 EEE http://www.estvideo.com/dew/ text/html #19 200704216031061171 60- |
| 2007-04-2716103111.731E | 200 | 4040 http://wp.postitexpres.fr/pollitcshw.js E http://blpwebline.blogspot.com/pollitcshw/ application/x-javascript #04 |
| 2007-04-2716103111.792E | 1 | 178 dms/www.google.fr EEEF http://www.google.fr/custom text/dms #050 20070421603127342 3141H3H8F028V27X2TKATXK3508 |
| 2007-04-2716103111.763E | 200 | 373 http://log93.xiti.com/bit-xiti=4537174p-pollitcshw EEE http://wp.postitexpres.fr/pollitcshw.js image/gif #032 |
| 2007-04-2716103111.841E | 200 | 4148 http://www.estvideo.com/dew/index/developpement-midi_mel_mel-http://d3w.free.fr/medias/q78.midi/evo1image-50 EEE http://www |
| 2007-04-2716103111.103E | 200 | 5833 http://blpwebline.blogspot.com/photos/perso/moblibygyy.jpg E http://blpwebline.blogspot.com/pollitcshw/ image/jpeg #04 |
| 2007-04-2716103113.277E | 200 | 2355 http://www.google.fr/robots.txt EEEF http://www.google.fr/custom text/plain #038 200704216031144+11 065530M17FAI- |
| 2007-04-2716103113.795E | 200 | 60840 http://www.google.com/dew/images/2007_03_27_inlanc.jpg EEE http://www.estvideo.com/dew/ image/jpeg #003 20070421 |
| 2007-04-2716103113.888E | 200 | 4481 http://www.google.fr/custom EX http://WP.postitexpres.fr/pollitcshw.js text/html #023 200704216031191480 15MR4- |
| 2007-04-2716103113.971E | 200 | 40603 http://www.estvideo.com/dew/images/2007_05_23_frechugs.jpg EEE http://www.estvideo.com/dew/ image/jpeg #038 200704 |
| 2007-04-2716103114.438E | 200 | 5669 http://www.google.fr/img/arc/images/2007/03/27/arc.gif EEE http://www.google.fr/custom image/gif #089 200704216031437- |
| 2007-04-2716103115.277E | 200 | 4763 http://www.estvideo.com/dew/index/2007 EEE http://www.estvideo.com/dew/ text/html #068 2007042160314871+311 RW- |
| 2007-04-2716103116.156E | 200 | 31823 http://www.estvideo.com/dew/index/2004/0 EEE http://www.estvideo.com/dew/ text/html #04 2007042160315044+55 2A |
| 2007-04-2716103117.157E | 200 | 20990 http://www.estvideo.com/dew/images/2007_0_23_laandresau1.jpg EEE http://www.estvideo.com/dew/ image/jpeg #300 200704 |
| 2007-04-2716103117.503E | 200 | 127350 http://www.nuesblog.com/pollitcshw/images/voip1.jpg E http://blpwebline.blogspot.com/pollitcshw/ image/jpeg #049 |
| 2007-04-2716103118.696E | 200 | 16564 http://www.estvideo.com/dew/index/2006/07 EEE http://www.estvideo.com/dew/ text/html #05 2007042160317660 7K- |
| 2007-04-2716103119.790E | 302 | 293 http://nuesblog.com/preunes.jpg EEE http://www.nuesblog.com/7617/Asous-Begag-par-Christophe-Car/poano text/html # |
| 2007-04-2716103120.207E | 200 | 23479 http://www.estvideo.com/dew/index/2007/03/27/793-quadre-ct-sucot-0decodm0 EEE http://www.estvideo.com/dew/ te |
| 2007-04-2716103120.800E | 200 | 44653 http://blpwebline.blogspot.com/pollitcshw/images/2007/03/18/pen_in_black.jpg E http://blpwebline.blogspot.com/pollitcshw |
| 2007-04-2716103121.701E | 200 | 71375 http://www.estvideo.com/dew/index/2003/11 EEE http://www.e |
| 2007-04-2716103121.950E | 200 | 461 http://www.estvideo.com/dew/index/2007/03/27/793-quadre-ct-sucot-0decodm0 EEE http://www.estvideo.com/dew/ image/gif #138 200704- |
| 2007-04-2716103121.951E | 200 | 0 http://blpwebline.blogspot.com/photos/uncategorised/14.jpg E http://blpwebline.blogspot.com/pollitcshw/ image/jpeg #04 |
| 2007-04-2716103121.982E | 200 | 380 http://www.estvideo.com/dew/state/buttons.gif EEE http://w |
| 2007-04-2716103121.499E | 200 | 78715 http://www.estvideo.com/dew/index/2005/04 EEE http://www.e |
| 2007-04-2716103121.975E | 200 | 1099 http://www.estvideo.com/dew/index/default/images/bodybo_1 |
| 2007-04-2716103121.967E | 200 | 181842 http://blpwebline.blogspot.com/pollitcshw/images/2007/03/24/ |
| 2007-04-2716103121.931E | 200 | 4248 http://www.estvideo.com/dew/media/dewplayer-mini.swf?up3=h |
| 2007-04-2716103121.451E | 200 | 4248 http://www.estvideo.com/dew/media/dewplayer-mini.swf?up3=h |
| 2007-04-2716103120.019E | 200 | 203099 http://www.estvideo.com/dew/index/2003/11 EEE http://www.e |
| 2007-04-2716103120.444E | 200 | 4248 http://www.estvideo.com/dew/index/2003/11 EEE http://www.e |
| 2007-04-2716103120.523E | 200 | 110219 http://blpwebline.blogspot.com/pollitcshw/roo.xml E http://blpwebline.blogspot.com/pollitcshw/ application/atom+xml #0 |

Indexation



Sauvegarde

SPAR :
Entrepôt
numérique

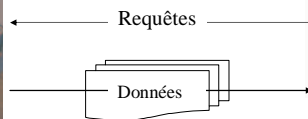
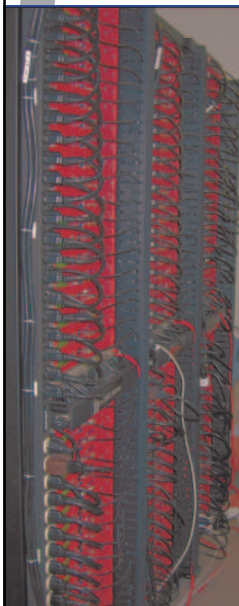
Préservation du
train de bits



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

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 dominant : HTML, JPEG, GIF, PDF, plain-text

20

Décrire et emballer

- ◆ 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