Des outils pour archiver le Web : l'expérience de la Bibliothèque nationale de France

Sara Aubry

Bibliothèque nationale de France

Colloque CECO qui archive le Web et comment ?

Archives fédérales suisses, Berne, 27 mai 2013

Cadre général

- Cadre juridique : le dépôt légal (loi de 2006 puis Code du Patrimoine)
  - « signes, signaux, écrits, sons ou messages de toute nature faisant l'objet d'une communication au public par voie électronique »
  - deux dépositaires : INA (sites de radio et télévision), BnF (tout le reste !)
  - la collecte remplace le dépôt
  - la représentativité remplace l'exhaustivité
  - l'accès est restreint aux enceintes de la BnF

Cadre général

- Cadre documentaire : les contenus édités ou les éditeurs doivent avoir un lien avec le territoire français
  - sites en .fr
  - sites édités par des personnes physiques ou morales domiciliées en France
  - publications produites sur le territoire
  - pas de jugement de valeur
- Quelques chiffres :
  - 15,3 milliards d'URL, 245 To (déc 2012)
  - 9 ETP (bibliothécaires et ingénieurs), des réseaux de partenaires interne, national et international

La coopération internationale

- Consortium international pour la préservation d'Internet (IIPC)
- Existe depuis 2003, 44 membres, 25 pays
- Objectifs :
  - développer en commun **outils**, normes, bonnes pratiques
  - assurer une couverture du Web au niveau mondial
  - promouvoir l'archivage du Web
- Fort investissement de la BnF dans ces travaux

IIPC netpreserve.org INTERNATIONAL INTERNET PRESERVATION CONSORTIUM

« L'ensemble des actions, outils et méthodes mis en œuvre pour **réunir, identifier, sélectionner, classer et conserver…**dans le but de…les **rendre accessibles…»**

Pour collecter : un robot de collecte

- Logiciel aussi appelé « aspirateur », « araignée » ou « moissonneur » de sites
- Simule un internaute utilisant un navigateur mais de manière répétitive et parallélisée
  - part d'une URL de départ (« graine »)
  - envoie une requête au serveur web, reçoit une réponse
  - parcourt le code de la page pour identifier des références vers d’autres pages / objets
  - enregistre le contenu et les nouvelles URL à traiter
- Ce processus est répété tant que les URL trouvées font partie du périmètre de la collecte (profondeur, budget, temps)

1 page, 48 URL

1 HTML

- 1 text/css
- 4 javascript
- 17 image/png
- 5 image/jpeg
- 21 image/gif

HERITRIX

View logs

Status as of May. 23, 2013 09:42:08 GMT Alerts: 3 (3 new)

CRAWLING JOBS RUNNING job: *5993-28*

0 jobs pending, 0 completed 220371 URIs in 1h36m53s (42.05/sec)

Console | Jobs | Profiles | Logs | Reports | Setup | Help

View: crawl.log
- local-errors.log
- progress-statistics.log
- runtime-errors.log
- uri-errors.log

By: Line number
- Time stamp
- Regular expression
- Tail

Refresh time: No refresh

Lines to show: 50

crawl.log for 5993-28

```
2013-05-23T09:42:08.014Z   301        0 http://television.telerama.fr/tele/programmes-tv/la-cuisine-est-un-jeu-d-enfants,437416
2013-05-23T09:42:08.022Z   200     6321 http://www.spectacles.carrefour.fr/item/243/2438612251285479683/Visuel_Petit/iris731233
2013-05-23T09:42:08.027Z   200    31009 http://i.huffpost.com/gen/1128528/thumbs/s-MORT-JEANNE-COOPER-large300.jpg LEE http://w
2013-05-23T09:42:08.048Z   200     6453 http://www.youtube.com/embed/oq2y9ODSEEM LE http://www.martinique.franceantilles.fr/act
2013-05-23T09:42:08.071Z   200    23519 http://www.ultimedia.com/deliver/musique/iframe/mdtk/01692309/article/3luxm/zone/4/auto
2013-05-23T09:42:08.072Z   404     1097 http://i4.ytimg.com//vi//oq2y9ODSEEM//hqdefault.jpg LEX http://www.youtube.com/embed/oq
2013-05-23T09:42:08.095Z   200        9 http://jx-stream3.dginfra.net//e98b2345ba3d51dd5bf954cfab3f7446//c3BlZWQ9MTUwO3VzZXI9an
2013-05-23T09:42:08.099Z   200     1483 http://www.leprogres.fr/fr/images/DCA18C5D-D819-4673-BD27-386A06B3866C/LPR_08/un-specta
2013-05-23T09:42:08.110Z   200     5111 http://www.dna.fr/fr/images/644F9937-0595-4AD5-A975-3B8AB10FD9F1/DNA_02/photo.jpg LE ht
2013-05-23T09:42:08.115Z   200    50678 http://www.touteleurope.eu/fileadmin/CIEV2/frise_multimedia/130306_infog_budget_UE_150_
2013-05-23T09:42:08.124Z   200     4647 http://www.lanouvellerepublique.fr/var/nrv2/storage/images/contenus/articles/2013/05/22
2013-05-23T09:42:08.142Z   404    22395 http://www.lesechos.fr/economie-politique/regions/h_normandie/lesechos.fr LX http://www
2013-05-23T09:42:08.185Z   200    12877 http://cache.20minutes.fr/illustrations/2013/05/22/interieur-centrale-nucleaire-sud-ira
2013-05-23T09:42:08.188Z   404        0 http://www.leberry.fr/page-2/a.niv_3:visible LX http://www.leberry.fr/page-2/accueil.ht
2013-05-23T09:42:08.198Z   200     3533 http://www.lest-eclair.fr/media_lesteclair/imagecache/article-miniature-1/protec/2013-0
2013-05-23T09:42:08.209Z   200     6133 http://www.courrierdelouest.fr/sites/default/files/imagecache/liste/2013/05/18/photo_9.
2013-05-23T09:42:08.242Z   200      993 http://www.ledauphine.com/haute-savoie/2013/05/09/click.modalMail LX http://www.ledauph
2013-05-23T09:42:08.243Z   200    24005 http://www.lexpress.fr/medias/2156/1103906.jpg LEX https://plus.google.com/104989737591
2013-05-23T09:42:08.265Z   404        0 http://www.lamontagne.fr/page-3/idle.idleTimer LX http://www.lamontagne.fr/page-3/accue
2013-05-23T09:42:08.275Z   200    73026 http://mondoweiss.net/images/2012/11/Gaza4.jpg LXE http://blogs.mediapart.fr/blog/james
2013-05-23T09:42:08.311Z   302        0 https://graph.facebook.com/oauth/authorize?client_id=289531604408142&redirect_uri=http%
2013-05-23T09:42:08.316Z   200     2069 http://www.iookaz.com/static/images/std/actsup.gif LXE http://www.iookaz.com/Saisie/Dep
2013-05-23T09:42:08.317Z   200     3483 http://www.lepoint.fr/images/2013/05/21/1472158-jpg_1371205_113x76.jpg LE http://www.le
2013-05-23T09:42:08.330Z   200    66429 http://static.mediapart.fr/files/imagecache/770_pixels/jamesinparis/qosmos.jpg LE http:
2013-05-23T09:42:08.332Z   200     4634 http://www.arretsurimages.net/images/blocs/fonds/cont-emi.png EE http://www.arretsurima
2013-05-23T09:42:08.358Z   404        0 http://www.slateafrique.com/sites/default/files/imagecache/vignette_list_v2/20130430/rs
2013-05-23T09:42:08.393Z   200   138899 http://www.leparisien.fr/insolite/concept-car-renault-s-amuse-avec-la-twizy-renault-spo
```

Heritrix (http://crawler.archive.org), logiciel libre, développé depuis 2003 par Internet Archive, avec le soutien du Consortium IIPC

Pour collecter : Heritrix

- Atouts
  - « simple » d'utilisation
  - archivage sélectif ou grands volumes
  - tous formats
  - utilisable sur Internet ou sur des intranets/extranets
  - s'adapte à différentes architectures de sites, recompose les pages vues par les internautes
- Limites, obstacles
  - pages fictives
  - liens complexes
  - Web profond, contenus payants, contenus en flux

Pour identifier : le format ARC

- Un format-valise pour enregistrer les données et les métadonnées (fichiers, collectes)
- Caractéristiques :
  - auto-descriptible
  - taille : 100 Mo
  - extensible
  - compressé (gzip)

http://www.republicain-lorrain.fr/ 145.226.50.12 20130106110321 text/html 158747
HTTP/1.1 200 OK
Date: Sun, 06 Jan 2013 11:03:21 GMT
Server: Apache
Set-Cookie: .XCONNECT=QvZH25t4jPTgr_8d1S3pjQayli53Pkxy2yPhTwGDoBKDAnEs3vu5O6lfF1dFuoDfD3nsMCPGCBU6hqVgNEwyXYC4
6nKBJ86xe3LI58H7JKXP7FTNwe5_SmxjgF4IWFLHd8kuMWuURwREZYN6p1y1kv4LeG31v9p9_8aT-cLi6w2W8FJ9-_DVvdiMDxagh_9X1ITJe6
MIcerE84y6B1CErttdi90YsN1myRpmLBn3VbahPz_qd2fIEBTXSTVFFbDns; path=/
Vary: User-Agent,Accept-Encoding
Connection: close
Content-Type: text/html; charset=ISO-8859-1

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transit
<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="fr" lang="fr">
        <head>
                <title>Actualit&#233; | Une - Le R&#233;publicain Lorrain</title>
                                <meta name="Keywords" content="Une,Actualit&#233;" />
                                                <meta name="google-site-verification" content="RzeCV2SkboNWSvL

<link type="text/css" rel="stylesheet" href="/fr/css/style.css" media="screen"/>
<link type="text/css" rel="stylesheet" href="/fr/css/jquery-ui-1.8.5.custom.css" />
<link type="text/css" rel="stylesheet" href="/fr/css/style_impr.css" media="print" />
<link type="image/x-icon" rel="shortcut icon" href="/fr/images/favicon.ico" />
```

Pour identifier : bientôt le format WARC

- Un format standardisé (ISO 28500:2009) conçu pour la préservation
- 8 types d’enregistrement
  - warcinfo
  - response
  - resource
  - request
  - metadata
  - conversion
  - revisit
  - continuation

```
WARC/1.0
WARC-Type: response
WARC-Target-URI: http://www.republicain-lorrain.fr/robots.txt
WARC-Date: 2012-11-08T11:52:37Z
WARC-Payload-Digest: sha1:YLMPNF4774VISBNE2FN55EMOPLRYSMBT
WARC-IP-Address: 145.226.50.12
WARC-Record-ID: <urn:uuid:fb1c7106-781b-46bc-96b4-f4c066f04c63>
Content-Type: application/http; msgtype=response
Content-Length: 1102

HTTP/1.1 200 OK
Date: Thu, 08 Nov 2012 11:52:38 GMT
Server: Apache
Last-Modified: Thu, 14 Jun 2012 08:10:32 GMT
Accept-Ranges: bytes
Content-Length: 842
Vary: Accept-Encoding,User-Agent
Connection: close
Content-Type: text/plain; charset=ISO-8859-1

User-agent: *
Disallow: /fr/GPRS_EnvoiAmi_Form.aspx
Disallow: /fr/search/GPRS_Search.aspx
Disallow: /fr/recherche/recherche.html
```

```
WARC/1.0
WARC-Type: request
WARC-Target-URI: http://www.republicain-lorrain.fr/robots.txt
WARC-Date: 2012-11-08T11:52:37Z
WARC-Concurrent-To: <urn:uuid:fb1c7106-781b-46bc-96b4-
     f4c066f04c63>
WARC-Record-ID: <urn:uuid:7297596b-9406-4b04-a1f6-3241c0bd9615>
Content-Type: application/http; msgtype=request
Content-Length: 359

GET /robots.txt HTTP/1.0
User-Agent: Mozilla/5.0 (compatible; bnf.fr_bot;
     +http://www.bnf.fr/fr/outils/a.dl_web_capture_robot.html)
From: robot@bnf.fr
Connection: close
Accept: text/xml, application/xml, application/xhtml+xml,
     text/html; q=0.9, text/plain; q=0.8, image/png, */*; q=0.5
Accept-Language: fr, en; q=0.3
Host: www.republicain-lorrain.fr
```

Pour rendre accessible : Wayback

- Donner la possibilité de retrouver le site « tel qu'il était »
  - navigateur web
  - logiciel de restitution : Wayback
- Logiciel libre développé par Internet Archive pour permettre la recherche par URL, l’affichage des pages archivées et la navigation
- Ne pas confondre avec la « Wayback Machine » d’Internet Archive

accueil | aide

BnF

Archives de l'Internet

Outils : Recherche par URL | Recherche par mot | Parcours guidés |

Les archives sont constituées de sites internet du domaine français archivés de 1996 à aujourd'hui.

à propos des archives de l'Internet...

Recherche par URL

Retrouver un site, une page, un fichier en indiquant son adresse internet (exemple : http://www.ihtp.cnrs.fr).

http://www.lemonde.fr/

Remonter le temps Recherche avancée

Option

Limiter la recherche à cette année : -

Recherche par mot

Retrouver ces mots dans la partie indexée des archives (environ 5%, documents archivés en nov-déc 2006 et 2007).

Rechercher Recherche avancée

Possibles :

- une expression : "Louis XIV"
- un mot sur un site : site:www.francegenweb.org Bretagne

Parcours guidés

Tous les parcours ...

Découvrir le contenu des archives et se familiariser avec les outils de recherche et de consultation.

Images amateurs, amateurs d'images

Le développement du Web, associé à celui des équipements de prise de vue numérique et des logiciels de post-traitement, a offert aux simples particuliers des facilités sans précédent pour publier des photographies et des vidéos. Les archives de l'Internet constituées par la BnF fournissent la matière d'un regard rétrospectif sur une décennie d'usages et de réalisations d'amateurs dans le domaine des images fixes et animées.

La révolution tunisienne à travers le Web

La presse en ligne est une source d'informations en plein développement. Si la plupart des sites présentent des informations produites dans le cadre d'un projet

http://archive-access.sourceforge.net/projects/wayback/

{BnF

Archives de l'Internet

Retrouver ce site ou cette page :

http://www.lemonde.fr/

Remonter le temps

Option

Limiter à :

- - -

Recherche par URL

pour : http://www.lemonde.fr/

3 867 résultats

du 1 jan. 1996 au 23 mai 2013

ouvrir tout | fermer tout

2013	234 résultats
2012	1 679 résultats
2011	1 173 résultats
2010	369 résultats
2009	59 résultats
2008	67 résultats
2007	206 résultats
2006	13 résultats
2005	10 résultats
2004	2 résultats
2003	10 résultats

BnF

Archives de l'Internet

Retrouver ce site ou cette page :

http://www.lemonde.fr/

Remonter le temps

Option

Limiter à :

- - -

Recherche par URL

pour : http://www.lemonde.fr/

3 867 résultats

du 1 jan. 1996 au 23 mai 2013

ouvrir tout | fermer tout

2013

234 résultats

2012

1 679 résultats

jan.	fév.	mar.	avr.	mai	juin	jui.	août	sep.	oct.	nov.	déc.
1 jan.	1 fév.	1 mar.	1 avr.	1 mai	1 juin	1 jui.	1 août	1 sep.	1 oct.	1 nov.	3 déc.
1 jan.	1 fév.	1 mar.	1 avr.	1 mai	1 juin	1 jui.	1 août	1 sep.	2 oct.	1 nov.	3 déc.
1 jan.	1 fév.	1 mar.	1 avr.	1 mai	1 juin	1 jui.	1 août	1 sep.	3 oct.	1 nov.	3 déc.
1 jan.	1 fév.	1 mar.	1 avr.	1 mai	1 juin	2 jui.	2 août	2 sep.	4 oct.	2 nov.	3 déc.
1 jan.	1 fév.	1 mar.	2 avr.	2 mai	1 juin	3 jui.	2 août	3 sep.	4 oct.	3 nov.	4 déc.
1 jan.	2 fév.	1 mar.	3 avr.	2 mai	2 juin	4 jui.	3 août	4 sep.	4 oct.	4 nov.	4 déc.
2 jan.	2 fév.	2 mar.	4 avr.	2 mai	2 juin	5 jui.	3 août	5 sep.	5 oct.	5 nov.	5 déc.
2 jan.	3 fév.	2 mar.	4 avr.	2 mai	3 juin	6 jui.	4 août	6 sep.	6 oct.	6 nov.	5 déc.
3 jan.	3 fév.	2 mar.	4 avr.	3 mai	3 juin	7 jui.	4 août	12 sep.	6 oct.	7 nov.	6 déc.
3 jan.	4 fév.	3 mar.	5 avr.	3 mai	3 juin	8 jui.	5 août	12 sep.	7 oct.	8 nov.	6 déc.
3 jan.	4 fév.	3 mar.	5 avr.	4 mai	4 juin	9 jui.	5 août	12 sep.	8 oct.	9 nov.	7 déc.
4 jan.	5 fév.	3 mar.	6 avr.	4 mai	4 juin	9 jui.	6 août	12 sep.	9 oct.	10 nov.	7 déc.
4 jan.	5 fév.	4 mar.	7 avr.	5 mai	5 juin	9 jui.	6 août	12 sep.	9 oct.	11 nov.	8 déc.
5 jan.	6 fév.	4 mar.	8 avr.	5 mai	5 juin	9 jui.	7 août	12 sep.	10 oct.	12 nov.	8 déc.
5 jan.	6 fév.	4 mar.	9 avr.	6 mai	6 juin	9 jui.	7 août	13 sep.	11 oct.	13 nov.	9 déc.
6 jan.	7 fév.	4 mar.	10 avr.	6 mai	6 juin	10 jui.	8 août	13 sep.	11 oct.	13 nov.	9 déc.
6 jan.	7 fév.	5 mar.	11 avr.	7 mai	6 juin	11 jui.	8 août	14 sep.	11 oct.	13 nov.	10 déc.
7 jan.	8 fév.	5 mar.	11 avr.	7 mai	6 juin	12 jui.	8 août	15 sep.	11 oct.	13 nov.	10 déc.
7 jan.	8 fév.	5 mar.	11 avr.	7 mai	7 juin	13 jui.	9 août	16 sep.	11 oct.	13 nov.	10 déc.
8 jan.	8 fév.	6 mar.	12 avr.	7 mai	7 juin	14 jui.	9 août	17 sep.	11 oct.	13 nov.	11 déc.
8 jan.	9 fév.	6 mar.	12 avr.	7 mai	8 juin	15 jui.	10 août	17 sep.	11 oct.	13 nov.	12 déc.
9 jan.	9 fév.	6 mar.	12 avr.	7 mai	8 juin	16 jui.	10 août	17 sep.	11 oct.	13 nov.	13 déc.
9 jan.	9 fév.	7 mar.	13 avr.	7 mai	9 juin	17 jui.	11 août	18 sep.	11 oct.	13 nov.	14 déc.
10 jan.	10 fév.	7 mar.	13 avr.	7 mai	9 juin	17 jui.	11 août	19 sep.	11 oct.	13 nov.	15 déc.
10 jan.	10 fév.	7 mar.	13 avr.	7 mai	10 juin	17 jui.	12 août	20 sep.	11 oct.	13 nov.	16 déc.
11 jan.	11 fév.	7 mar.	13 avr.	7 mai	10 juin	17 jui.	12 août	21 sep.	11 oct.	13 nov.	16 déc.
11 jan.	11 fév.	7 mar.	14 avr.	7 mai	11 juin	17 jui.	13 août	22 sep.	11 oct.	13 nov.	17 déc.
12 jan.	12 fév.	8 mar.	14 avr.	7 mai	11 juin	17 jui.	13 août	23 sep.	11 oct.	13 nov.	18 déc.
12 jan.	12 fév.	8 mar.	14 avr.	7 mai	12 juin	17 jui.	14 août	24 sep.	11 oct.	13 nov.	19 déc.
13 jan.	13 fév.	8 mar.	15 avr.	7 mai	12 juin	17 jui.	14 août	25 sep.	11 oct.	13 nov.	20 déc.

Page archivée le : 06 mai 2012 à 08:05 GMT
Permalien :
http://archivesinternet.bnf.fr/2012050

Capture 2 347 sur 3 867

Recherche rapide :
http://www.lemonde.fr/
accueil | recherche initiale

Le Monde.fr

Mise à jour à 10h04 – Paris

Le second tour de la présidentielle en direct

LIVE 109

Les bureaux de vote en métropole ont ouvert à 8 heures. Une majorité de Français d'outre-mer et de l'étranger a elle déjà voté.
Suivez en direct la journée électorale.

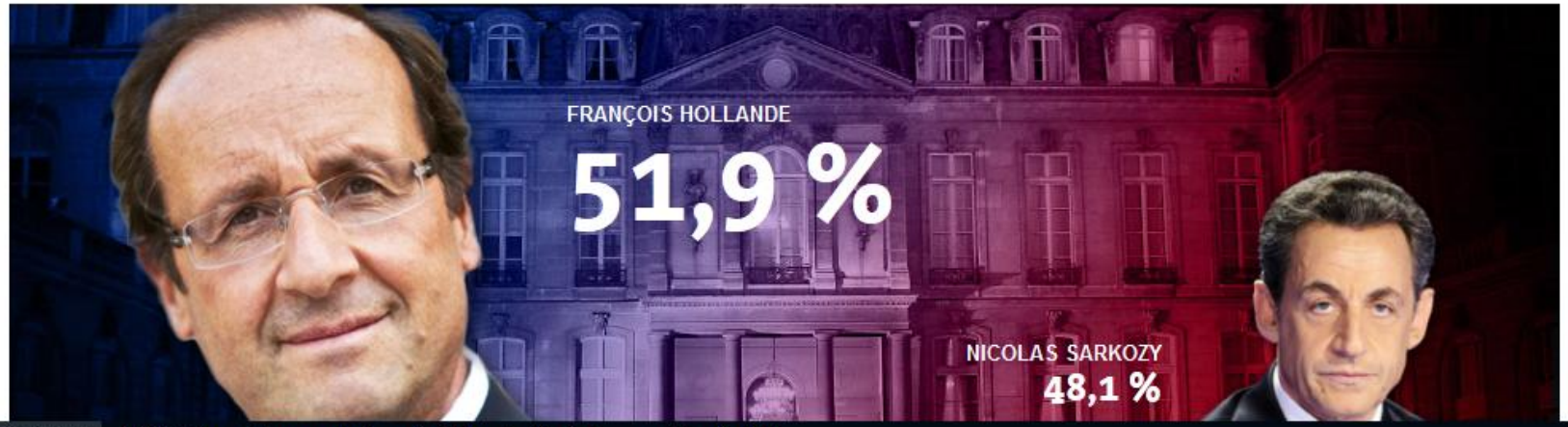
Page archivée le : 06 mai 2012 à 18:09 GMT
Permalien :
http://archivesinternet.bnf.fr/2012050

Capture 2 348 sur 3 867

Recherche rapide :
http://www.lemonde.fr/
accueil | recherche initiale

François Hollande élu président de la République LIVE

Le candidat du PS a devancé Nicolas Sarkozy au second tour de l'élection présidentielle, avec 51,9 % des voix contre 48,1 % au président sortant, selon les estimations Ipsos pour "Le Monde", France télévisions et Radio France à partir des premiers bulletins dépouillés.



EN CONTINU 20:02 Dans le bureau de François Hollande 10 minutes avant les résultats

Page archivée le : 07 mai 2012 à 05:08 GMT
Permalien :
http://archivesinternet.bnf.fr/2012050

Capture 1 sur 26

Recherche rapide :
http://www.lemonde.fr/election-presic

Election présidentielle : les enseignements du second tour

Le Monde TéléramaLe Monde diplomatiqueLe Huffing

06.05.12 | 17h30 • Mis à jour le 07.05.12 | 06h42

S'abonner au Monde - 15 € / mois

- Services Le Monde
  - **Boutique Le Monde**
  - Accueil
  - DVD
  - CD
  - Livres
  - Hors-séries
  - Accessoires
  - Sélection du mois
  - **Partenaires Le Monde**
  - Annonces auto
  - Annonces immo
  - Annonces emploi
  - Rencontres

François Hollande élu président de la République, avec 51,7 % des voix selon les dernières estimations.

Une archive web : un artefact ?
Problème de cohérence temporelle

Le Monde

- 06h33 "The French are revolting. The Greeks, too" Paul Krugman et les colères européennes : http://t.co/uwPVyIKF lucbronner via twitter
- 06h30 Côté PS, Martine Aubry, qui quittera son poste de première secrétaire, sera impliquée. Mais c'est véritablement François Hollande qu
- 06h30 @Guest Après le discours de Nicolas Sarkozy, on sait que ce ne sera pas le président sortant qui conduire l'UMP. A la tête de la majo Fillon et d'autres ténors. Le Monde
- 06h29 Qui conduira les législatives pour le PS et l'UMP ? Commentaire de la part de Guest
- 06h11 @Guest Les élections législatives auriont lieu les 10 et 17 juin prochains. D'ores et déjà, des sondages sur les intentions de vote, à pr rapport de force : la gauche est créditée de 26 à 35 % des voix, contre 30 à 33 % des voix à l'UMP. Le FN en compte 18 %. Le Monde
- 06h09 Combien de temps avant les législatives, à présent ? Commentaire de la part de Guest
- 06h06 Pour ceux qui le demandaient : le discours de Hollande à la Bastille en vidéo. Le Monde
- 06h03 @Guest Nocture et diurne, puisque que nous continuerons à suivre les suites de l'élections de François Hollande tout au long de la se des législatives. Le Monde
- 06h02 Bonjour Le Monde. Un grand bravo pour votre suivi très nocturne de l?élection. Commentaire de la part de Guest
- 05h56 Et revivez la soirée électorale en images et en vidéos ici. Le Monde
- 05h56 Et n'oubliez pas : retrouvez tous les résultats, commune par commune, les analyses et les reportages de nos journalistes dans notre es

Prochains défis de la BnF

- Maintenance et consolidation des outils
- Collecte de contenus payants (presse, livres électroniques)
- Recherche par mot et indexation plein-texte
- Collecte des réseaux sociaux (Facebook, Twitter, YouTube…)
- Passage au format WARC

Des questions ?

E-mail : sara.aubry@bnf.fr

Site web : http://www.bnf.fr

Twitter : http://twitter.com/DLWebBnF