



Le dépôt légal face aux mégadonnées

L'exemple de l'archivage du web à
la Bibliothèque nationale de France

Big Humanities

Journée d'étude : Recherche et Internet

8 décembre 2014

Ange Aniesa

Dépôt légal numérique

Bibliothèque nationale de France

{BnF



Quel big Data?

{BnF



Le dépôt légal numérique de la BnF

Mise en contexte

Un cadre: le dépôt légal



Une mission en perpétuelle extension :

De 1537 à 2006

Des livres à l'internet

2006 Loi Dadvsi - 2011 Décret d'application : une mission inscrite au code du Patrimoine

« Sont également soumis au dépôt légal les signes, signaux, écrits, sons ou messages de toute nature faisant l'objet d'une communication au public par voie électronique »



Le périmètre concerné

Deux institutions dépositaires

INA

BnF

Une délimitation de l'internet français, du « domaine français »

- ce qui est produit en France
- ce qui est enregistré en France
- ce qui est édité par un producteur résidant en France.

Pas de jugement de valeur !



Des évolutions fondamentales

Une collecte plus qu'un dépôt

De l'exhaustivité à la « représentativité »

Des permanences

Les missions du dépôt légal:


Collecter

Préserver

Signaler

Communiquer

La continuité des collections



Statut juridique et accès aux collections

Des conditions d'accès strictement réglementées

Droit d'auteurs et propriété intellectuelle

Droit à la vie privée et protection des données personnelles

Accessible par des chercheurs accrédités, dans les emprises de la BnF

(BnF

Une infrastructure technique
Un circuit complet d'archivage



Accès

Sélection



Préservation

Collecte



Indexation

HERITRIX



Un robot de collecte

HERIRIX

Status as of Dec. 2, 2014 15:46:12 GMT Alerts: 16 (16 new)

CRAWLING JOBS

RUNNING job: BnF-12087-47

View logs

0 jobs pending, 0 completed 2210172 URIs in 1d12h22m39s (14.75/sec)

[Console](#) [Jobs](#) [Profiles](#) [Logs](#) [Reports](#) [Setup](#) [Help](#)

View: crawl.log

By: [Line number](#)Refresh time: [local-errors.log](#)[Time stamp](#)Lines to show: [progress-statistics.log](#)[Regular expression](#)[runtime-errors.log](#)

Tail

[uri-errors.log](#)

HERIRIX

crawl.log for BnF-12087-47

2014-12-02T15:46:00.241Z	404	46831	http://www.teamlemac.fr/index.php/forum-243/35-entraide-faq/14413-probleme-graphique/list LLR http://www.teamlemac.fr/
2014-12-02T15:46:00.264Z	200	3040	http://team-marmules.fr/wp-content/mingle-forum-skins/ArtofPic-Elegant-Premiumskin-1.1/images/bbc/u.png LLRRLRLRL
2014-12-02T15:46:00.309Z	302	213	http://www.tecan.com/handler.asp?Menu=1&Item=33&Id=2090 RRLRLRL http://www.tecan.com/platform/apps/product/index.a
2014-12-02T15:46:00.320Z	200	5185	http://www.avatava.org/pics100/abstraction/avatar-82.jpg LLLXLLLLLE http://yhaqaji.hol.es/4243.html image/jpeg #017
2014-12-02T15:46:00.383Z	200	3401	http://www.tech2roo.com/2505-home_default/leviers-synto-yamaha-tdm-900-2002-2003.jpg RLE http://www.tech2roo.com/10
2014-12-02T15:46:00.511Z	200	50265	http://www.techni-plongee-lyon.fr/chaussure-nike-blazer-low-or-blanc-ly0939-p-939.html LL http://www.techni-plongee
2014-12-02T15:46:00.976Z	200	125769	http://www.techdeli.nl/montage/bouten/zeskantbout-deeldraad/bout-zeskant-bouten.html?lengte=73 RLX http://www.techd
2014-12-02T15:46:00.976Z	200	188211	http://stats.team-kzn.com/hlstats.php?mode=playerinfo&player=350158&game=dods LRRLL http://stats.team-kzn.com/hlsta
2014-12-02T15:46:00.999Z	200	72	http://img.photobucket.com/robots.txt LLLLEP http://img.photobucket.com/albums/v442/iftk/05/1176375035_84184_92135.
2014-12-02T15:46:01.006Z	200	5927	http://www.teatower.com/img/p/957-teatower-home.jpg LEE http://www.teatower.com/lu/category.php?id_category=1234607
2014-12-02T15:46:01.015Z	200	19498	http://www.laposte.fr/design/laposte/images/bienvenue/fond_home_bienvenue.jpg LLLLEEE http://www.laposte.fr/design/
2014-12-02T15:46:01.024Z	301	0	http://legiteam.xooit.eu/p2061.htm RLLL http://legiteam.xooit.eu/t264-et-bonjours.htm text/html #019 20141202154556
2014-12-02T15:46:01.034Z	404	214	http://www.tebubio-data.com/owl/08929511.pdf RLLX http://www.tebubio-data.com/owl/browse.php?sess=0&parent=688&exp
2014-12-02T15:46:01.071Z	200	50860	http://www.technologiemaileva.fr/design/laposte/images/bienvenue/julia_300x600.swf LLLLE http://www.technologiemail
2014-12-02T15:46:01.202Z	200	750	http://technica.fr/index.php/component/search/?Itemid=264&c=jsjobs&layout=list_jobs&jobcat=61&format=opensearch LLE
2014-12-02T15:46:01.411Z	200	23709	http://laruche.ecole-eme.com/calendar/view.php?view=month&course=1&cal_d=1&cal_m=8&cal_y=2015 RRRRLRLRLRLRL http://
2014-12-02T15:46:01.417Z	404	1372	http://www.technal.fr/fr/Accueil-Particuliers/bon-a-savoir/Sys.Services.ProfileGroup RLEX http://www.technal.fr/Sc
2014-12-02T15:46:01.461Z	200	563849	http://www.tech-brest-iroise.fr/files/2118/fiche_developpement_durable.pdf LL http://www.tech-brest-iroise.fr/Les-3-
2014-12-02T15:46:01.462Z	301	0	http://www.cityplay.fr/faqs/category/equipements/cityplay.fr RRLRLX http://www.cityplay.fr/faqs/category/equipemen
2014-12-02T15:46:01.492Z	200	27826	http://www.teavana.com/retail/united-states/ohio/beachwood/beachwood-place(62) LLL http://www.teavana.com/retail/un
2014-12-02T15:46:01.502Z	200	22747	http://www.teamutecia.fr/category/actu/a.zoombox/a.zoombox/a.zoombox/5/page2 LLLXLLXLLL http://www.teamutecia.fr
2014-12-02T15:46:01.518Z	200	28424	http://www.team-ttk.fr/coin-pastis/souvenir-souvenir-sous-venir-t392.html LLL http://www.team-ttk.fr/zone-technique
2014-12-02T15:46:01.614Z	200	2523	http://www.team-outdoor.fr/2022-medium_default/ah13-pointes-cross-xcs-4-femme.jpg LLE http://www.team-outdoor.fr/ch
2014-12-02T15:46:01.810Z	200	27689	http://www.teamlog.fr/actualites/modules/system/sites/all/themes/open_corporate/css/misc/jquery.once.js LXX
2014-12-02T15:46:01.810Z	200	166877	http://editions.lavoisier.fr/nouveautes-chimie-analytique.asp RL http://editions.lavoisier.fr/nouveautes-tec-et-doc
2014-12-02T15:46:01.851Z	200	19721	http://www.technicmanager04.com/default/catalog/seo_sitemap/product/anotherlevel//anotherlevel//anotherlevel//nivea
2014-12-02T15:46:02.050Z	200	33665	http://www.technic.com/apac/equipment/capabilities/turnkey-solutions?t=1&s=small XLLL http://www.technic.com/apac/e
2014-12-02T15:46:02.332Z	200	93978138	http://tebeo.static.e-declic.tv/ee948745a722b5d6419910d2e8fcec115/video.mp4 LLLX http://www.tebeotv.fr/video/6601-li
2014-12-02T15:46:02.439Z	302	0	http://catalogue.techniparts.fr/fr-default-home/vernis-antirouille-penetrant-rustol-300ml.jpg LLE http://catalogue
2014-12-02T15:46:02.590Z	200	18454	http://img.webnode.com/kb/fr/domaines-et-e-mails/modifier-le-mot-de-passe-de-boite-mail/3.png LRLRLRLRRE http://www
2014-12-02T15:46:02.597Z	1	105	dns:l2.yimg.com LLEP http://l2.yimg.com/bt/api/res/1.2/UJn4XwkWjpeVSK43RqM7QQ--/YXBwaWQ9eW51d3M7Zmk9aW5zZXQ7aD0zNT
2014-12-02T15:46:02.942Z	200	1910	http://www.technifor.com/media/system/images/closebox.png REE http://www.technifor.com/media/system/css/modal.css i
2014-12-02T15:46:03.375Z	200	12886016	http://www.techmart.fr/FurEineHandvoll/audio/jacques%20perconte%20-%20f%fc%r%20eine%20Handvoll%20%8uro%20(09)%20eat
2014-12-02T15:46:03.957Z	200	39246	https://www.electrogeoz.com/le-groupe/les-filiales/networks/22-filiales RXXX http://www.electrogeoz.com/le-groupe
2014-12-02T15:46:04.232Z	200	4215	http://qzonestyle.gtimg.cn/qzone_v6/proj_qzonelogin_sprite/qzonelogin-jan140808174545.png LXXE http://qzonestyle.g
2014-12-02T15:46:05.028Z	404	31152	http://surfcastingdream83.over-blog.com/tiny mce.dom.StringWriter LREX http://surfcastingdream83.over-blog.com/_tiny
2014-12-02T15:46:05.208Z	200	25795	http://www.techinter.fr/actualites/system/lightbox2/sites/techinter.fr/files/sites/techinter.fr/files/system/lightb
2014-12-02T15:46:05.261Z	404	770	http://clubhalieutis.e-monsite.com/a.smiley LREX http://clubhalieutis.e-monsite.com/themes/content.js?v=5_139394358
2014-12-02T15:46:05.939Z	200	57548	http://mezohir.technikboerse.com/view/gebrauchtmaschine/traktor/1700468/fendt-102.html LE http://www.technikboerse.
2014-12-02T15:46:06.537Z	200	15062	http://www.techage.fr/picto/indextests/ecrans/Samsung%20XL2370.jpg LLE http://www.techage.fr/viewtopic.php?f=105&t=
2014-12-02T15:46:06.565Z	200	4197	http://img.clubic.com/008C008C06167200-cl-photo-portable.jpg ELEE http://www.clubic.com/video/le-debug/video-google

{BnF



Des formats de stockage normalisés

ARC

WARC

Deux formats normalisés à l'international

{BnF

● ● ● | Ordres de grandeur

MASSE

Les archives de l'internet
représentent

21,2 milliards de fichiers
soit 470 To de données

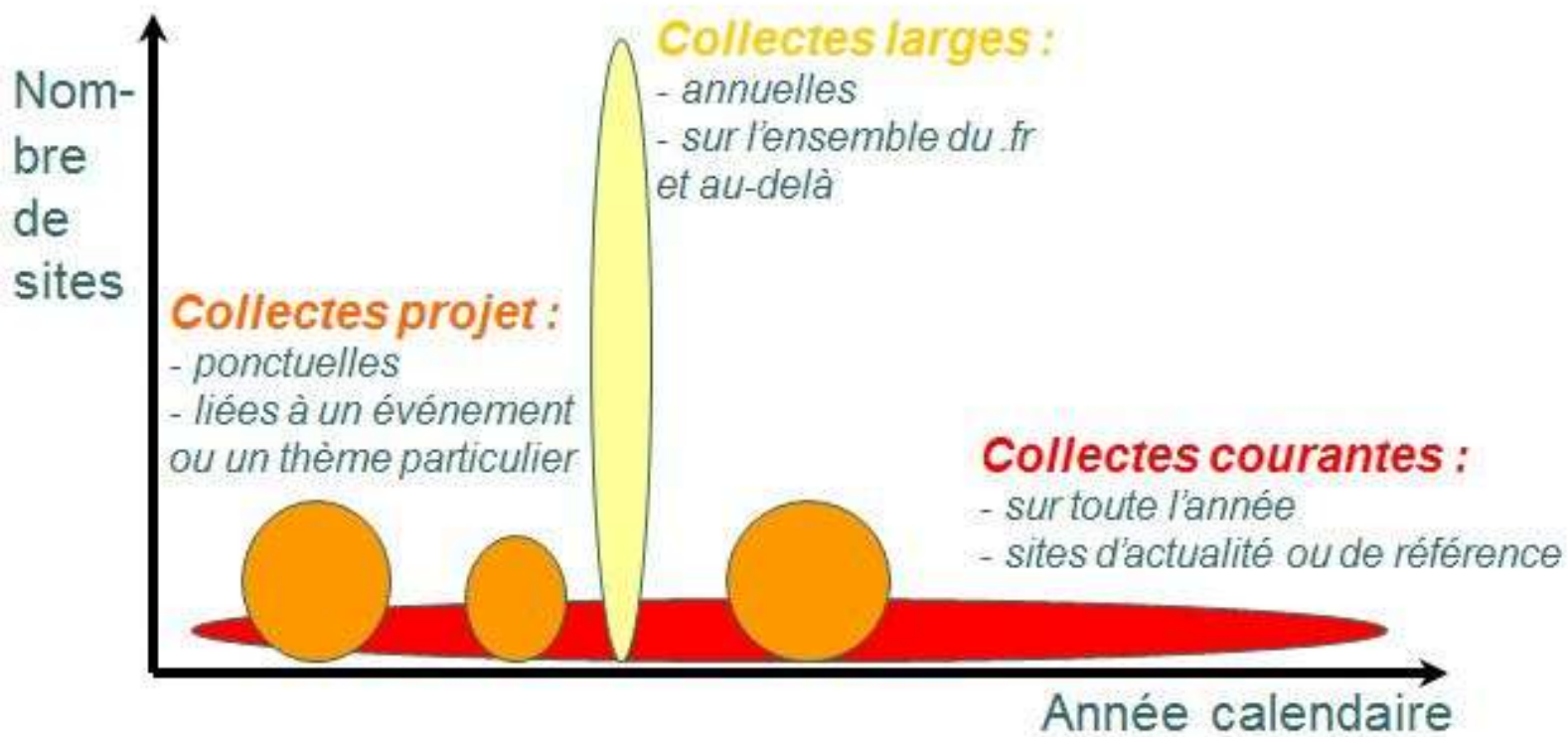
Ordres de grandeur

File type	Documents	Data
text/html	675924	24 GB
image/jpeg	587553	15 GB
image/png	89700	2.7 GB
image/gif	43173	
text/css	18870	
application/javascript	14787	
application/x-shockwave-flash	13826	
application/octet-stream	12642	
no-type	11926	
video/mp4	10480	
text/plain	10230	
text/dns	9771	
text/javascript	8582	
text/xml	8081	
application/x-javascript	6070	
image/jpg	4310	
image/svg+xml	2727	
application/xml	2114	
application/rss+xml	1841	
application/atom+xml	1442	
application/json	1342	
application/vnd.ms-sstr+xml	1149	
application/vnd.ms-fontobject	1110	
image/x-icon	888	
audio/mpeg	760	
audio/x-mpegurl	721	
application/pdf	653	
application/x-font-woff	508	
text/vtt	496	
application/x-font-ttf	436	
font/ttf	425	
application/font-woff	383	
application/vnd.apple.mpegurl	345	
image/vnd.microsoft.icon	255	
image/bmp	246	
video/x-ms-wmv	235	
application/zip	153	794 MB
font/truetype	192	9.5 MB

Hétérogénéité

HTML, PDF,
 ZIP, CSS, MP4,
 PNG, XML,
 GIF, EPUB,
 SWF, MPEG,
 WAV, AVI,
 RAR, FLV ...

Le « modèle intégré » de la BnF





Deux filières d'entrée pour les données

Collectes larges

Une collecte de surface sur quatre millions de domaines, soit 60% des sites français

Des listes d'URL obtenues auprès des agences et bureaux d'enregistrement des extensions.

Avec un nombre limités d'URL pour chaque sites web



Deux filières d'entrée pour les données

Collectes ciblées

Des collectes plus complètes et régulières

Sélectionnés par un réseau de bibliothécaires spécialisés

Avec des paramètres techniques modulables (fréquence, nombre d'URL...)



Les collectes projets

Mise en place de collectes ciblées en
« mode projet »

Un modèle thématique et coopératif
aisément reproductible

Archivage du web à l'international



INTERNATIONAL
INTERNET
PRESERVATION
CONSORTIUM



Convergence des outils et des collections

IIPC

Des outils communs et interopérables

Coopération entre différents types d'institutions patrimoniales

Des activités documentaires similaires (collectes projets, collectes collaboratives)



Focus: 12 ans de
collecte du « web
électoral » à la BnF
2002-2014

12 ans de collecte du « web électoral » à la BnF 2002-2014



Valerie Trierweiler

@valtrier

Abonnements



Courage à Olivier Falorni qui n'a pas démerité, qui se bat aux côtés des rochelais depuis tant d'années dans un engagement désintéressé.

← Répondre ↻ Retweeter ★ Favori

Page archivée le 02 avril 2007 à 09:14 GMT
Formuler
<http://archive.winter.net/mb/2007/04>

Recherche rapide
<http://www.cartouz.fr/pagine/index> [FR]

accueil | **ENSEMBLE TOUT DEVIENT POSSIBLE** | aujourd'hui 21 TR

ME TV DÉBATTE/ PARTICIPER GAGNER ENSEMBLE MON PROJET CE QUE JE CROIS DEVOIR DE REPONDRE NEWS EDITO PRESSE BONSIS & ACCUEIL

ACCUEIL / GAGNER ENSEMBLE /

Ils me soutiennent

VOS MESSAGES

28-03-07
J'ai toujours été de droite, et ces dernières années j'ai été déçu. J'ai écouté vos discours et sous m'avez redonné confiance. J'espère que vous serez élu. Vous êtes le seul à pouvoir sauver la France, à redonner le courage aux gens qui travaillent. Car il y en a assez de l'assistanat, de voir tous ces gens qui ne travaillent pas et qui ne veulent pas travailler et moins galérer que ceux qui se lèvent le matin pour aller au boulot. J'ai particulièrement apprécié votre prestation sur Canal+. J'espère que vos engagements seront réalisés.
Par: *Cristine Froulet*

18-03-07
J'espère de tout coeur que vous serez notre président. Je vous soutiens et suis d'accord avec vos projets. J'espère que la droite vaincra encore une fois la gauche !!! Merci de nous donner un espoir.
Par: *Sacré L*

PERSONNALITÉS **TOUT AFFIRMER**

Carlos Gonzalez, soutien de Nicolas Sarkozy:
Mon nom est Carlos Gonzalez, je suis député du Parlement ...

Naouf Cagnat, soutien de Nicolas Sarkozy:
Je suis Naouf Cagnat, d'origine indochinoise, élu de la Marine ...

François Kocciulko-Hortez, soutien de Nicolas Sarkozy:
Je m'appelle François Kocciulko-Hortez, je suis Maire de Sèvres ...

Le « web électoral »



**Une échelle différente
de collecte de
données**

Temporalité courte

Volatilité extrême

**Un Espace
documentaire à
archiver**

**Une collection à
constituer, continuer
et faire exploiter**

{BnF



2002-2014: Des enjeux de
collecte identiques...

Maintenir la continuité des collections

Suivre une temporalité raccourcie

Des enjeux identiques: Saisir une volatilité accrue



Page archivée le : 19 mai 2011 à 13:56 GMT

Permalien

<http://val100archivesinternet.bnf.fr/20>

« Capture 1 sur 4 »

Recherche rapide :

OK

[accueil](#) | [recherche initiale](#)

Les Français de l'étranger avec Dominique Strauss-Kahn pour les primaires PS de 2011

DSK

Blog

Actualités

Participez !

Nous sommes nombreux, parmi les Français établis hors de France, à souhaiter que DSK soit le candidat du Parti socialiste et de la Gauche à l'élection présidentielle de 2012. Les distances qui nous séparent, notre dispersion entre les pays et les villes du monde rendent indispensable un site de dialogue et d'information. C'est la vocation de ce site: il fournit des informations sur les positions défendues par DSK, il prend part aux débats qui seront ceux de la primaire organisée par le PS, il permet à chacun de s'exprimer.

Richard Yung

Claudine Lepage

Sénateurs des Français établis hors de France



{BnF



... une même réponse

Encadrer la sélection

{BnF



Deux principaux émetteurs
identifiés

Sélection via deux grands axes:

Contenu de la classe politique
(candidats et leurs organisations)

Contenu émanant du corps social
(regards et opinions sur la campagne)



Des critères de périmètre

Critère de contenu

Critère d'actualité

Critère de singularité

Critère de langue

Critère de représentativité

Critère de navigabilité

Un collecte complémentaire et reliée à l'ensemble des collectes



Une typologie

- 0 – Sites officiels et institutionnels
- 1 - Les candidats et leurs organisations
 - 1.1 Sites des candidats en campagne
 - 1.2 Sites des formations politiques
 - 1.3 Autres organisations de soutien
- 2 - Regards et opinions sur la campagne
 - 2.1 Annuaire, observatoires et analyses
 - 2.2 Médias traditionnels
 - 2.3 Associations, syndicats et autres organisations
 - 2.4 Expressions individuelles et communautaires sur l'Internet

0 - Sites officiels et institutionnels

- qui veillent à l'organisation des élections et au respect du cadre juridique :
juge électoral, administration centrale, autorités administratives
- qui représentent des institutions, des collectifs :
collectivités territoriales, associations non partisans d'élus

1 - Les candidats et leurs organisations

1.1 Sites des candidats en campagne

sites et blogs officiels ou personnels, permanents ou de campagne, des candidats

1.2 Sites des formations politiques

partis et formations en campagne

1.3 Autres organisations de soutien

uniquement les sites clairement affiliés aux partis et aux candidats (comités de soutien par exemple)

2 - Regards et opinions sur la campagne

2.1 Annuaires, observatoires et analyses

sites de sondages et de tendances, les nouveaux outils d'analyse de type cartographique

2.2 Médias traditionnels

sélection restreinte à un nombre limité de titres de presse d'information disposant d'un site à l'activité significative

2.3 Associations, syndicats et autres organisations

sites relayant la parole d'acteurs collectifs qui n'existent pas uniquement sur l'Internet

2.4 Expressions individuelles et communautaires sur l'Internet

sites représentatifs de l'expression l'opinion publique à travers le média Internet



Un calendrier de fréquence

2012

Janvier 2012 : Lancement des premières collectes, Archivage de la pré-campagne, entrée dans la « phase de production » d'archives

Mars 2012 : Début officiel de la campagne, archivage plus fréquent et étendu aux réseaux sociaux

Avril et mai 2012 : Premier et second tour des élections

Juin 2012 : Premier et second tour des législatives

Juillet 2012 : dernière collecte élections mensuelle

{BnF



Des paramètres de collecte différenciés

Selon la nature du site et sa mise à jour, quatre fréquence différentes :

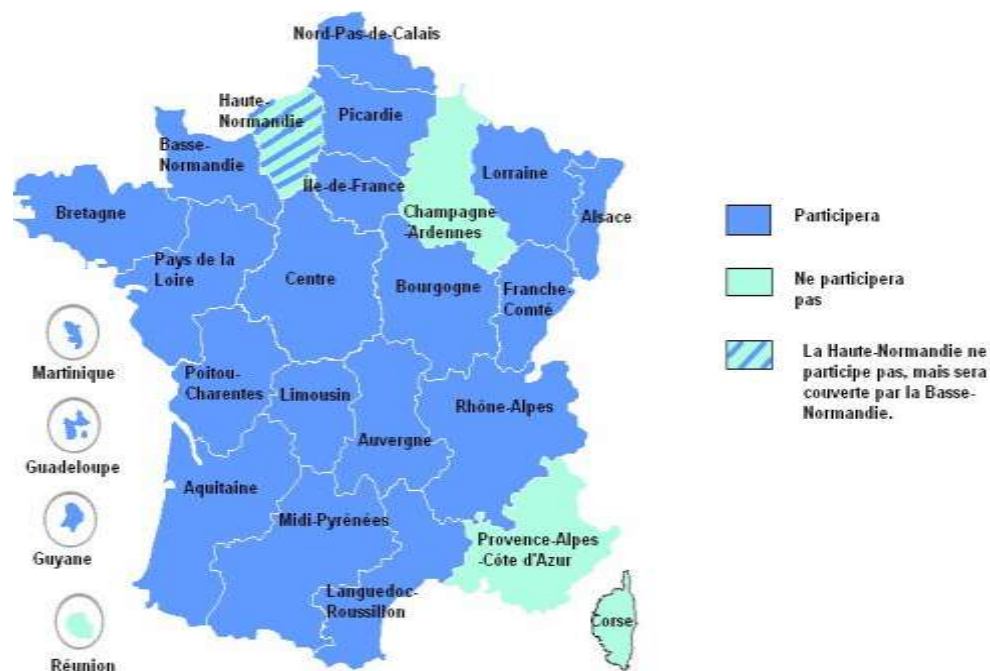
Mensuelle

Hebdomadaire

Quotidienne

Quatre fois par jour

Un dispositif : un réseau de sélectionneurs plus vaste



20 bibliothèques en région pour couvrir les campagnes locales

Des bibliothécaires des départements thématiques de la BnF



Ampleur des collections

Date	Type d'élection	URL proposées	URL collectées	Poids (Go)
2002	Présidentielle et législatives	1 906	14 164 537	332
2004	Européennes et régionales	1 162	8 635 463	202
2007	Présidentielle et législatives	5 813	63 500 000	3 380
2008	Municipales	100	1 322 187	164
2009	Européennes	480	8 890 452	582
2010	Régionales	2 063	20 000 000	1 200

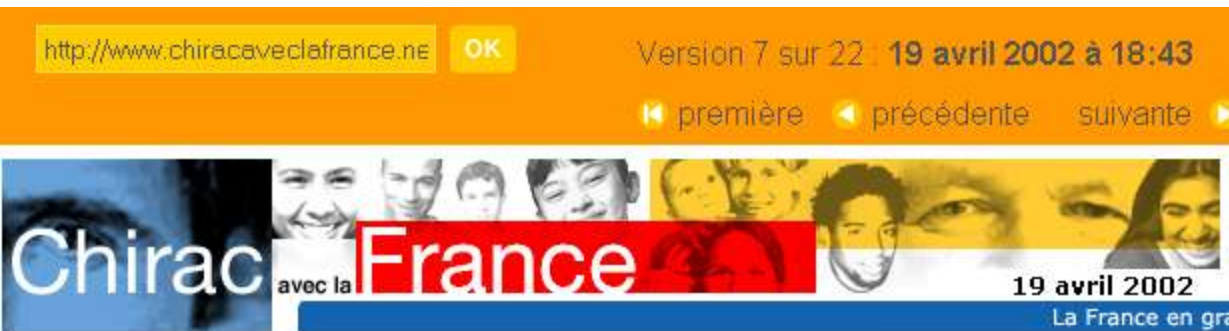


Ampleur des collections

2012: Sur sept mois de collecte, 10 726 URL proposées, 379 millions collectées, soit 11 177 Go de données

2014: 1032 pour les municipales, 463 pour les européennes, 57,2 millions collectées, 1375,81 Go de données

Des grandes tendances observables



- Jacques Chirac
- ▶ L'homme
- ▶ Ses propositions
- ▶ Discours
- ▶ Interviews
- ▶ Son équipe
- ▶ Pourquoi Chirac ?
- ▶ Audios, vidéos & photos
- Le débat démocratique
- ▶ Les enjeux
- ▶ Recherche
- Prenez la parole
- ▶ Votre avis
- ▶ Les forums
- La France avec Chirac
- ▶ Métropole



"Si, comme moi, avec moi, vous voulez une autre politique, alors vous devez en prendre les moyens, vous devez m'en donner les moyens. Unir vos forces, vos énergies. Tendre vos volontés. La victoire est à ce prix. Elle est nécessaire. Je vous invite à vous rassembler."

Jacques Chirac
Lille, 18 avril 2002

2002 :
Les premières
« Web-
campagnes »

Des grandes tendances observables



- Participer
- Débats participatifs
- Autres débats
- Voir les dernières contributions
- Dernières contributions les plus approuvées
- Synthèses des débats
- Soutenir
- Adhérer en ligne
- Messages de soutien



En 2007 :

- La domination des blogs
- Le rôle du Web participatif



Debat-Sarkozy.fr

Les dernières questions - Les plus populaires - Les plus discutées

Connexion

Envie de débats ? Posez une question, c'est simple !

Bienvenue sur le site Debat-Sarkozy.fr ! Nicolas Sarkozy et son équipe répondent en direct du QG de campagne à une question par jour. C'est vous qui décidez de la question du jour en votant pour la question qui vous paraît être la plus importante. Chaque soir nous communiquerons la réponse de Nicolas Sarkozy.

{BnF



Des grandes tendances observables

2010-2014

Flux, Réseaux, et dataviz

(BnF

Des grandes tendances
observables dans la
collection et sa sélection

**Utilisation massive des réseaux
sociaux**

**Évolution de la stratégie de
communication des partis (plate-
formes)**

**Évolution de la forme prise par
l'information (type mime des**

{BnF



Limites observables



Perspectives 2015

Une attention particulière aux supports et avis du corps social (Réseaux sociaux)

Une veille active sur la réforme des régions

{BnF



Solutions d'accès, de
valorisation et
d'utilisation des
collections

{BnF



Deux types d'accès

Via l'interface d'accès « 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>).

[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).

[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.

Presse et actualité

Le web est un vecteur majeur de diffusion de l'information : les sites de presse et d'actualité font par conséquent l'objet d'un traitement spécifique par le dépôt légal de l'internet. Ses archives proposent un accès particulier aux contenus gratuits ou réservés aux abonnés, mis en ligne par les principaux sites nationaux et régionaux. Les collections les plus anciennes permettent d'étudier la naissance du web d'information, tandis que les plus récentes, à partir de 2010, donnent accès à une centaine de titres capturés quotidiennement.



Le web scientifique : de la vulgarisation aux sciences participatives

Accès en salles de lecture

The screenshot displays the 'Archives de l'internet' interface on the BnF website. The search criteria are 'http://elysee.fr' and the date range is 'du 1 jan. 1996 au 5 déc. 2014'. The results are presented as a horizontal bar chart where the length of the orange bar indicates the number of results for each year. The data is as follows:

Année	Nombre de résultats
2014	256
2013	288
2012	847
2011	332
2010	132
2009	22
2008	36
2007	93
2006	11
2005	83



Outils : [Recherche par URL](#) | [Recherche par mot](#) | [Parcours guidés](#) |

Retrouver ce site ou cette page :

[Remonter le temps](#)

Option

limiter à :



Recherche par URL

2 122 résultats

pour : <http://elysee.fr>

du 1 jan. 1996 au 5 déc. 2014

[ouvrir tout](#) | [fermer tout](#)

2014

256 résultats ▾

jan.	fév.	mar.	avr.	mai	juin	jui.	août	sep.	oct.	nov.	déc.
1	1	1	1	1	1	1	1	1	1	1	1
2	3	3	3	2	2	2	2	8	6	2	2
3	10	4	7	5	4	3	3	15	13	3	3
4	17	6	8	8	5	4	4	22	20	4	4
6	18	10	9	12	6	5	5	29	22	5	5
13	20	13	10	13	7	7	6		23	7	7
20	24	17	14	14	8	8	11		24	8	8
27		20	16	15	9	14	13		25	9	9
		24	21	19	10	15	18		26	10	10
		25	24	22	11	17	25		27	11	11
		27	26	26	12	18			28	13	13
		31	28	29	13	19			29	17	17
					14	20			31		
					15	21					
					16	22					
					17	23					
					18	24					
					19	25					
					20	26					
					21	27					
					22	28					
					23	29					
					24	30					
					25	31					
					26						
					27						
					28						
					29						
					30						



Accès en salles de lecture

Reconstitution de la nature hypertextuelle des sites web

Respect des caractères internes : le
« *look and feel* » du document

Respect des caractères externes :

La capacité à naviguer dans l'archive

Citation de l'archive du site web en tant
que source scientifique valide

Accès en salles de lecture

Des liens pérennes et normalisés



Page archivée le : **18 août 2014 à 12:13 GMT**

Permalien :

<http://archivesinternet.bnf.fr/2014081>

Accès rapide : contenu / navigation / recherche

Les actualités / Photos et vidéos / Chronologie

Reliés aux URL collectées

Indépendants de l'URL collectée



Accès en salles de lecture

Limités aux emprises de la BnF

Depuis septembre: un accès distant autorisé dans 26 bibliothèques de région et s'étendant progressivement (Montpellier, Nancy, Strasbourg et à la bibliothèque municipale de Lyon en 2014)

{BnF



Accès par analyse des
métadonnées

**Développements de services et d'outil
pour le Texmining, Datamining,
Linkmining et mapping**

{BnF



Accès par analyse des
métadonnées

**Des formats de fichiers de métadonnées
également normalisés**

A venir, DAT et WAT issus des ARC et
WARC

Signalement partiel des collections



data.gouv.fr

Plateforme ouverte des données publiques françaises

Comment ça marche ? Producteurs Licence Ouverte Métriques Etalab

Connexion / Inscription

Rechercher

Où

Thématiques

PUBLIEZ UN JEU DE DONNÉES !

Collectes du web électoral par la BnF

Ce jeu de données provient d'un service public certifié

Publié le 15 septembre 2013 par Bibliothèque nationale de France

Dans le cadre de sa mission patrimoniale de dépôt légal de l'internet, la Bibliothèque nationale de France collecte régulièrement le web politique français. Depuis 2002, et lors de chaque scrutin, sont archivés sites de l'État et des collectivités territoriales, sites de partis et de candidats en campagne, blogs, presse en ligne, activité publique sur les réseaux sociaux. Il s'agit de documenter le développement et la diversité de la vie politique et du débat citoyen en ligne. Ce travail est mené avec un nombre croissant de bibliothèques de dépôt légal en région (voir documentation jointe). Les sites archivés sont consultables par les chercheurs dans les salles de lecture de la BnF. Leur liste vous est ici proposée par scrutin. Vous y trouverez les URL des sites, leur description (partis rattachés, niveau géographique, typologie utilisée pour la sélection, mots-clés attachés), les paramètres techniques de collecte (profondeur, fréquence de passage des robots, historique de l'URL collectée). La BnF met ces données librement à votre disposition à condition d'en mentionner la source.

Ressources

CSV Collectes du web électoral BnF 2002-2012

PDF Télécharger la liste des établissements de dépôt légal imprimeur associés aux projets

Producteur

{ BnF



La Bibliothèque nationale de France (BnF) a pour mission de collecter, conserver et diffuser le patrimoine documentaire national.



S'ABONNER

Informations

- Licence Ouverte / Open Licence
- 01/01/2002 à 08/08/2012
- ponctuelle
- France
- Pays

ARCHIVAGE DU WWW BIBLIOTHEQUE

Un exemple de coopération BnF-PACTE



Page archivée le : 17 mars 2014 à 11:13 GMT

Permalien :

<http://archivesinternet.bnf.fr/2014031>

Capture 1 sur 1

Recherche rapide

<http://www.olympic.org/fr/news/le-pre>

OK

[accueil](#) | [recherche initiale](#)



Olympic.org

Site officiel du Mouvement olympique

English | Français | Suivez-nous [f](#) [t](#) [You Tube](#)

Trouver un résultat ou médaillé olympique

ou **rechercher**

Go

[Accueil](#)

[Jeux Olympiques](#)

[JOJ](#)

[Sports](#)

[Athlètes](#)

[Pays](#)

[L'Olympisme](#)

[Photos](#)

[Vidéos](#)

[Le CIO](#)

Actualités

[LE MUSÉE OLYMPIQUE](#)

[SHOP](#)

[Accueil](#) > [Actualités](#) > [Médias CIO](#) > Le président du CIO appelle les membres du Mouvement olympique à contribuer directement au débat sur l'Agenda olympique 2020

[partager](#)

ACTUALITÉS OLYMPIQUES

Recherche actualités

Go

OPTIONS

Le président du CIO appelle les membres du Mouvement olympique à contribuer directement au débat sur l'Agenda olympique 2020



À LA UNE

{BnF

Un exemple de coopération

● ● ● | BnF-laboratoire de
recherche

**Passage par un conventionnement de
coopération**

Co-construction de collecte

Mise à disposition de métadonnées

**Développements et mise en open source
d'outils d'analyse**



Merci de votre attention

Questions

Contact

@DLWebBnF

depot.legal.web@bnf.fr

ange.aniesa@bnf.fr