Webarchivierung an der Französischen Nationalbibliothek

Bert Wendland
Bibliothèque nationale de France
bert.wendland@bnf.fr

{BnF

Wer ich bin / wer wir sind

- Bert Wendland
  - crawl engineer in der IT-Abteilung der BnF
- gemeinsame Arbeitsgruppe
  - Legal-Deposit-Abteilung, Service „Dépôt légal numérique (DLnum)“
    - 1 Service-Leiter
    - 2 Bibliothekare
    - 2 Dokumentare
  - IT-Abteilung
    - 1 Projektkoordinatorin
    - 1 Software-Entwickler
    - 2 Crawl-Ingenieure
- ein Komitee von 80 Bibliothekaren aus den Fachabteilungen

BnF

Kontext

Die BnF und Web-Archivierung
als Teil ihres Sammelauftrags

Die BnF

- Bibliothèque *nationale* de France
- ca. 30 Millionen Bücher, Periodika und anderes
  - 10 Mio. am neuen Standort
  - jährlich 60.000 neue Medien
  - ca. 550 TB Daten im Web-Archiv
    - jährlich 100 TB neue Daten
- zwei Standorte
  - alter Standort « Richelieu » im Pariser Zentrum
  - neuer Standort « François-Mitterand » seit 1996
- zwei Ebenen am neuen Standort
  - Studienbibliothek (« Haut-de-jardin »): Freihandbestand
  - Forchungsbibliothek (« Rez-de-jardin »): Zugriff zum Gesamtbestand, einschließlich Web-Archiv

Frankreich und die Pflichtexemplare

1368	königliche Manuskriptsammlung von Karl V. im Louvre
1537	Franz I.: Pflichtabgabe – alle Herausgeber müssen ein Exemplar ihrer Produktion der königlichen Bibliothek überlassen
1648	Pflichtabgabe erweitert auf Karten und Pläne
1793	Musikpartituren
1925	Fotos und Schallplatten
1975	Videoaufnahmen
1992	CD-ROMs und elektronische Ausgaben
2002	Websites (experimental)
2006	Websites (produktiv)

Erweiterung des DADVSI am 1. August 2006

- « Droit d'auteur et droits voisins dans la société de l'information »
  - Französisches Urheberrechtsgesetz
- Umfang (Artikel 39)
  - *« Sont également soumis au dépôt légal les signes, signaux, écrits, images, sons ou messages de toute nature faisant l'objet d'une communication au public par voie électronique. »*
- Bedingungen (Artikel 41 II)
  - *« Les organismes dépositaires procèdent à la collecte des signes, signaux, écrits, images sons ou messages de toute nature mis à la disposition du public ou de catégories de public, … Ils peuvent procéder eux-mêmes à cette collecte selon des procédures automatiques ou en déterminer les modalités en accord avec ces personnes. »*
- Zuständigkeiten (Artikel 50)
  - INA (Institut national de l'audiovisuel) für Websites von Radio und TV
  - BnF für alles andere
- Webharvests ohne besondere Erlaubnis, aber Zugriff zum Archiv beschränkt auf Lesesäle
- Ziel: nicht das ganze frz. Internet, auch nicht „das Beste vom Web", sondern einen repräsentativen Querschnitt zu einem bestimmten Datum

{BnF

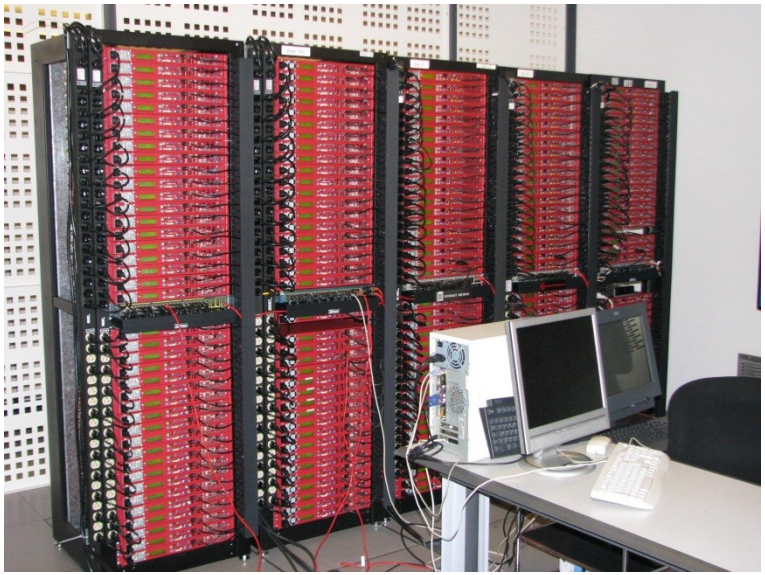
Konzept

Wie wir das französische Web speichern

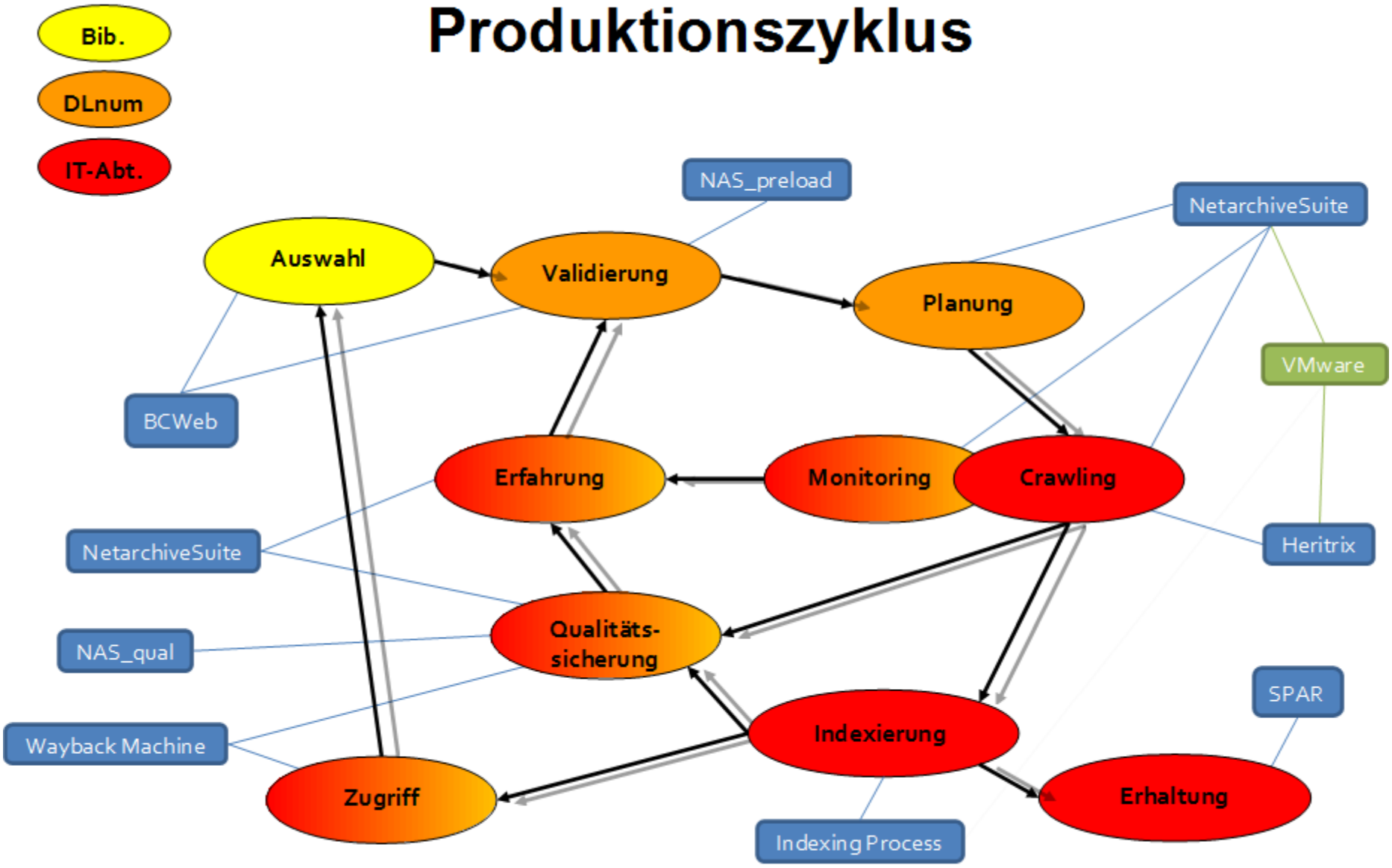
{BnF

Partnerschaft BnF – Internet Archive

- fünf Jahre zwischen 2004 und 2008
  - danach weitere Zusammenarbeit im Rahmen des IIPC (International Internet Preservation Consortium)
- Daten
  - 5 broad crawls und 2 focused crawls im Auftrag der BnF
  - Extraktion von historischen Alexa-Daten zu .fr rückwirkend bis 1996
- Technologie
  - Heritrix
  - Wayback Machine
  - 5 Petaboxen
- Know-how
  - Installation der Petaboxen durch Ingenieure des IA
  - Anwesenheit eines crawl engineers einen Tag pro Woche für 6 Monate



Produktionszyklus



{BnF

Herausforderungen

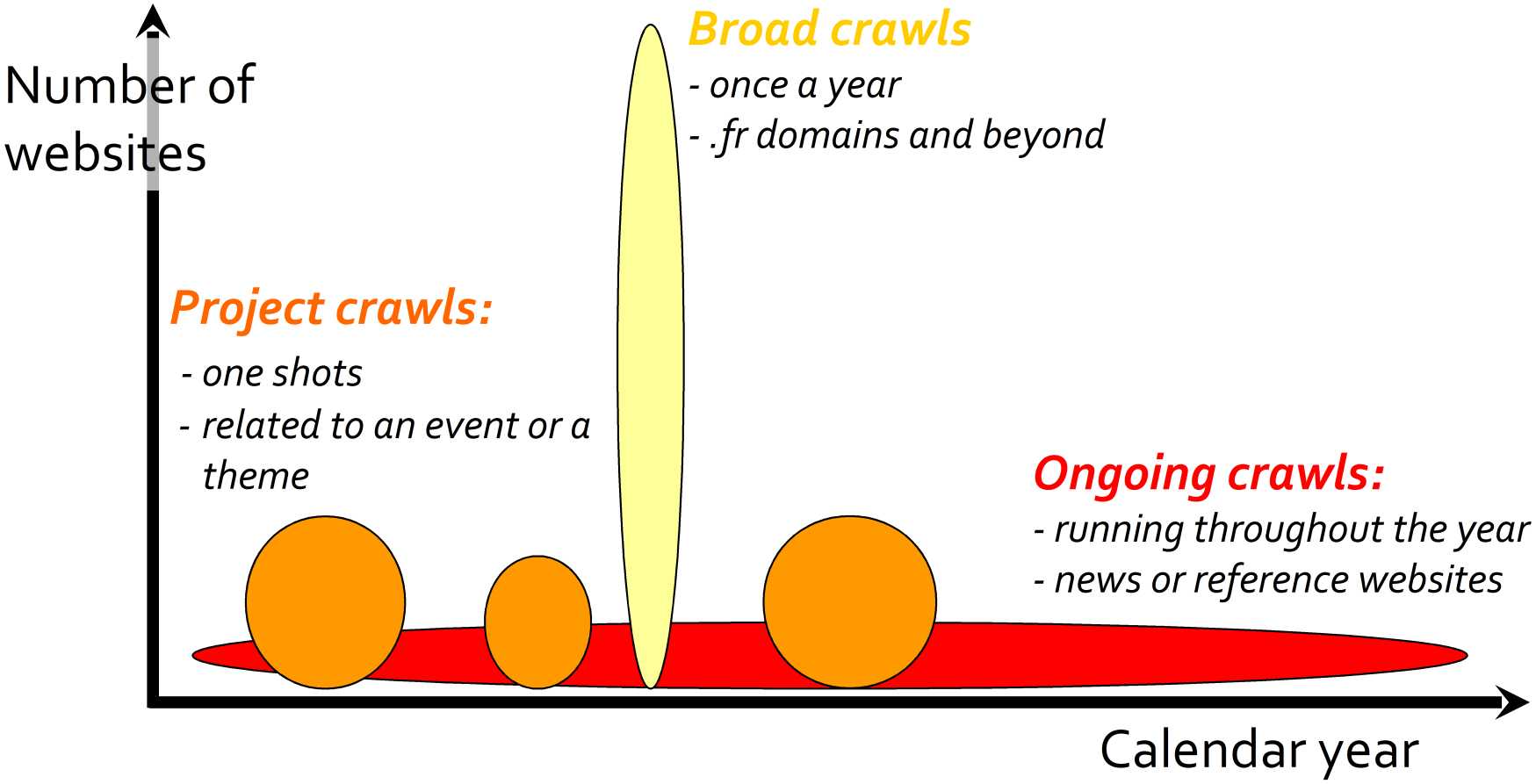
- Was ist das *französische* Web?
  - nicht nur .fr/.re/.nc, auch .com und .org
- einige Daten schwierig zu harvesten
  - Streaming, Datenbanken, Videos, JavaScript
  - dynamische Webseiten
  - Paßwort-geschützte Inhalte
  - Erfolge durch komplexe Konfigurationen, u.a.
    - Dailymotion
    - Online-Ausgaben im Abonnentenbereich von Zeitungen

BnF

Datenakquisition

Unser gemischtes Modell
des Web harvesting

BnF "mixed model" of harvesting



Broad crawl
Aggregation verschiedener Quellen

- 2013:
  - 2,7 Mio. domains in .fr und .re, geliefert von AFNIC (*Association française pour le nommage Internet en coopération* – französische Registrierungs-organisation)
  - 3500 domains in .nc, geliefert von OPT-NC (*Office des postes et télécommunications de Nouvelle-Calédonie* – Registrierungsorganisation von Neukaledonien)
  - 1,6 Mio. domains, geliefert von OVH (privater frz. Internetdienstleister)
  - 3,2 Mio. domains aus der Produktions-Datenbank
  - 11000 domains durch URL-Selektion von Bibliothekaren der BnF
  - 7800 domains aus anderen Geschäftsgängen der Bibliothek, die URLs als Teil der Metadaten enthalten: Verlagsdeklarationen für Bücher und Periodika, der Katalog der BnF, Identifikation neuer Periodika durch Bibliothekare, Online-Publizierung ehemaliger Print-Periodika, …
- nach Deduplikation eine Liste von 4,0 Mio. eindeutigen domains

{BnF

Der broad crawl 2013 in Zahlen

- Start-Domains: 4.014.595
- max. URLs pro domain: 2300
- Dauer: 10 Wochen
- parallele Crawler-Instanzen: 50
- threads pro crawler: 200
- gespeicherte URLs: 1.703.800.806
- ARC-Files (à ca. 100 MB): 551.177
- Datenvolumen (komprimiert): 56,2 TB

- TLDs: 299, davon 43,6 % .fr, 39,9 % .com, …, 0,8 % .de
- URLs pro domain: 49,8 % < 10, 45,5 % 10..2300, 4,7 % > 2300
- Statuscodes: 80,4 % 2xx, 11,4 % 4xx

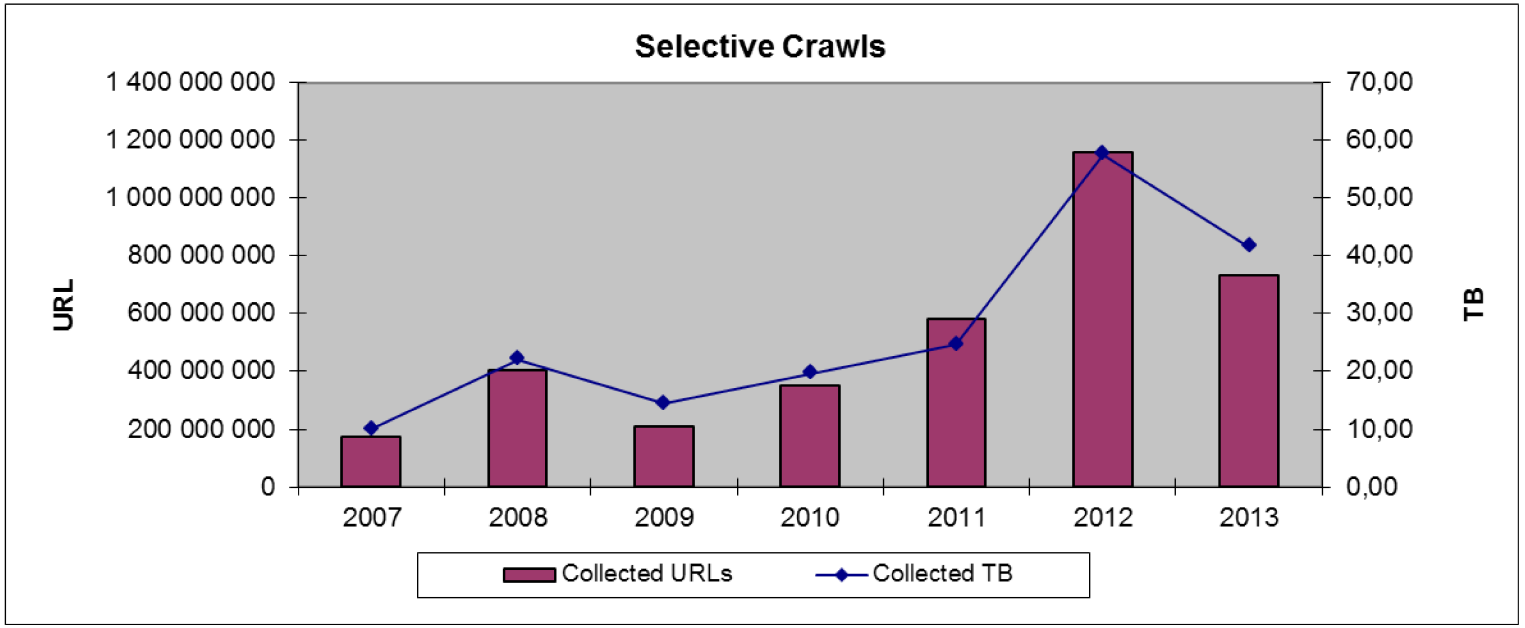
{BnF

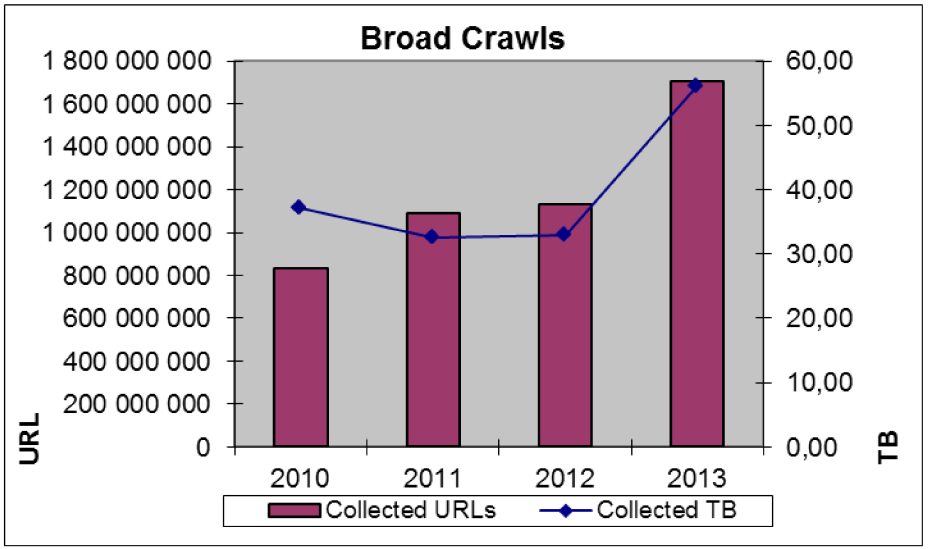
Datenvolumen

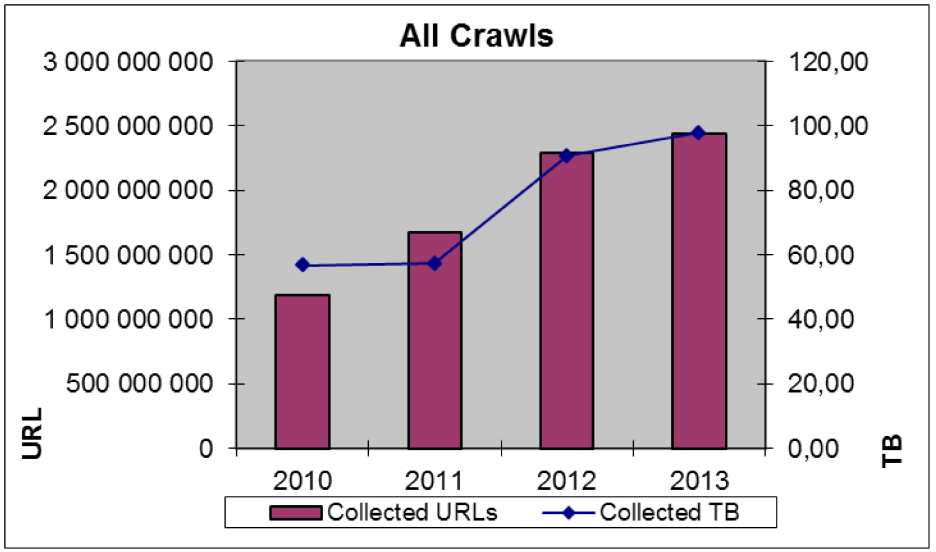
- Neun broad crawls seit 2004
- einige Zehntausend focus-crawled websites seit 2002
- 1996-2005 Internet Archive, seit 2006 BnF
- Gesamtgröße
  - 25 Milliarden URLs
  - 550 Terabyte

{BnF

Datenvolumen







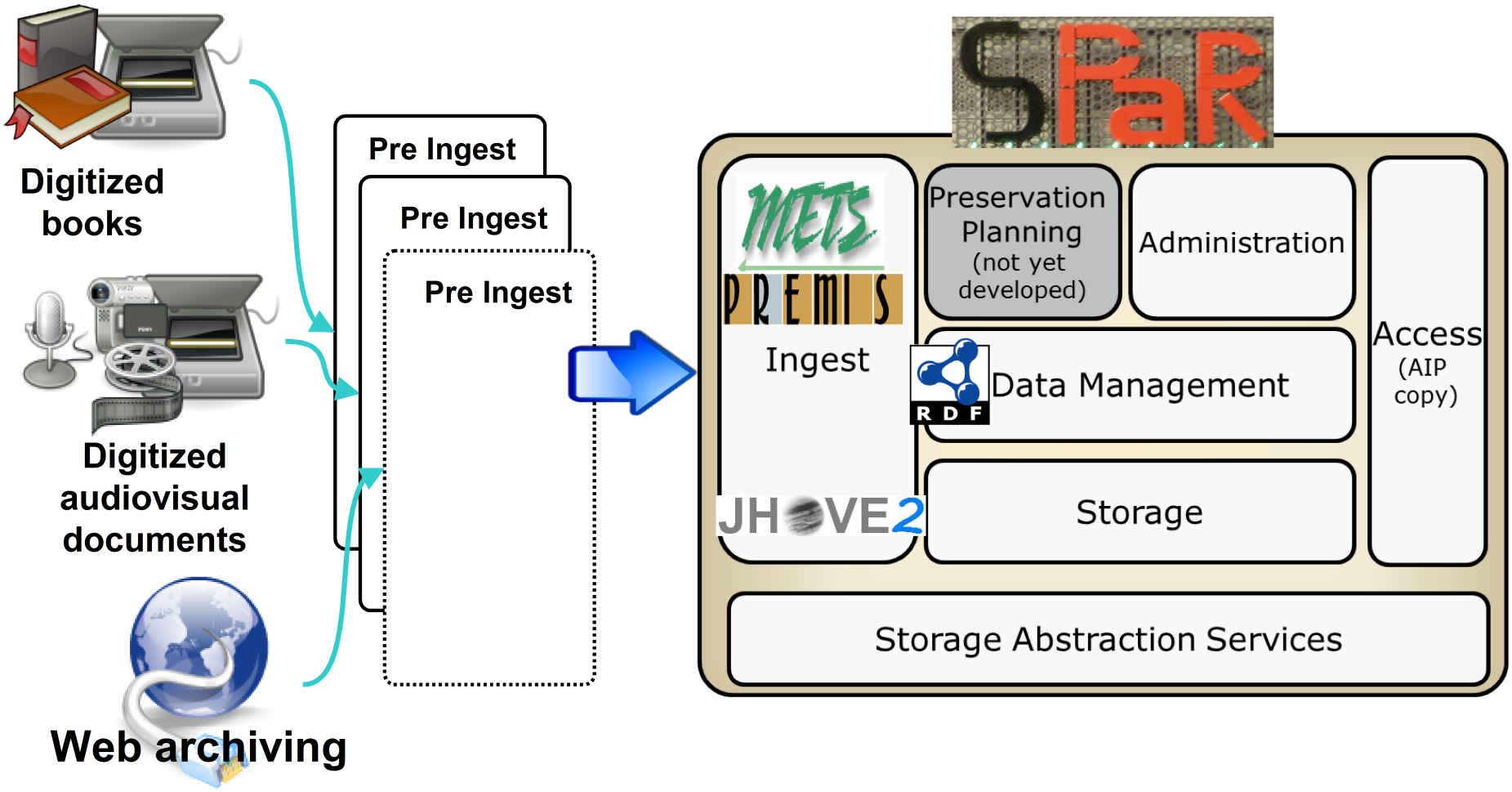
{BnF

Archivierung und Zugriff

Wie wir die Daten erhalten
und die Nutzer darauf zugreifen

SPAR
Système de Préservation et d'Archive Réparti

Langzeitarchivierungssystem für digitale Objekte, konform dem OAIS-Standard (Open Archival Information System), ISO 14721



SPAR
Das digitale Langzeitarchiv der BnF

Nutzerzugang zum Archiv

- angepaßte Version der open-source Wayback Machine
- drei Zugangswege:
  - URL-Suche
  - experimentelle Volltext-Suche über NutchWAX
    - im Moment für 10% des Archivs
  - parcours guidés (geführte Touren)
    - Auswahl aus dem Archiv, redaktionell aufbereitet durch Bibliothekare der BnF und externe Partner
    - nutzerfreundliches Interface zur Erschließung des Archivinhalts
    - Sichtbarkeit der thematischen harvests

BnF

Archives de l'internet

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 ...

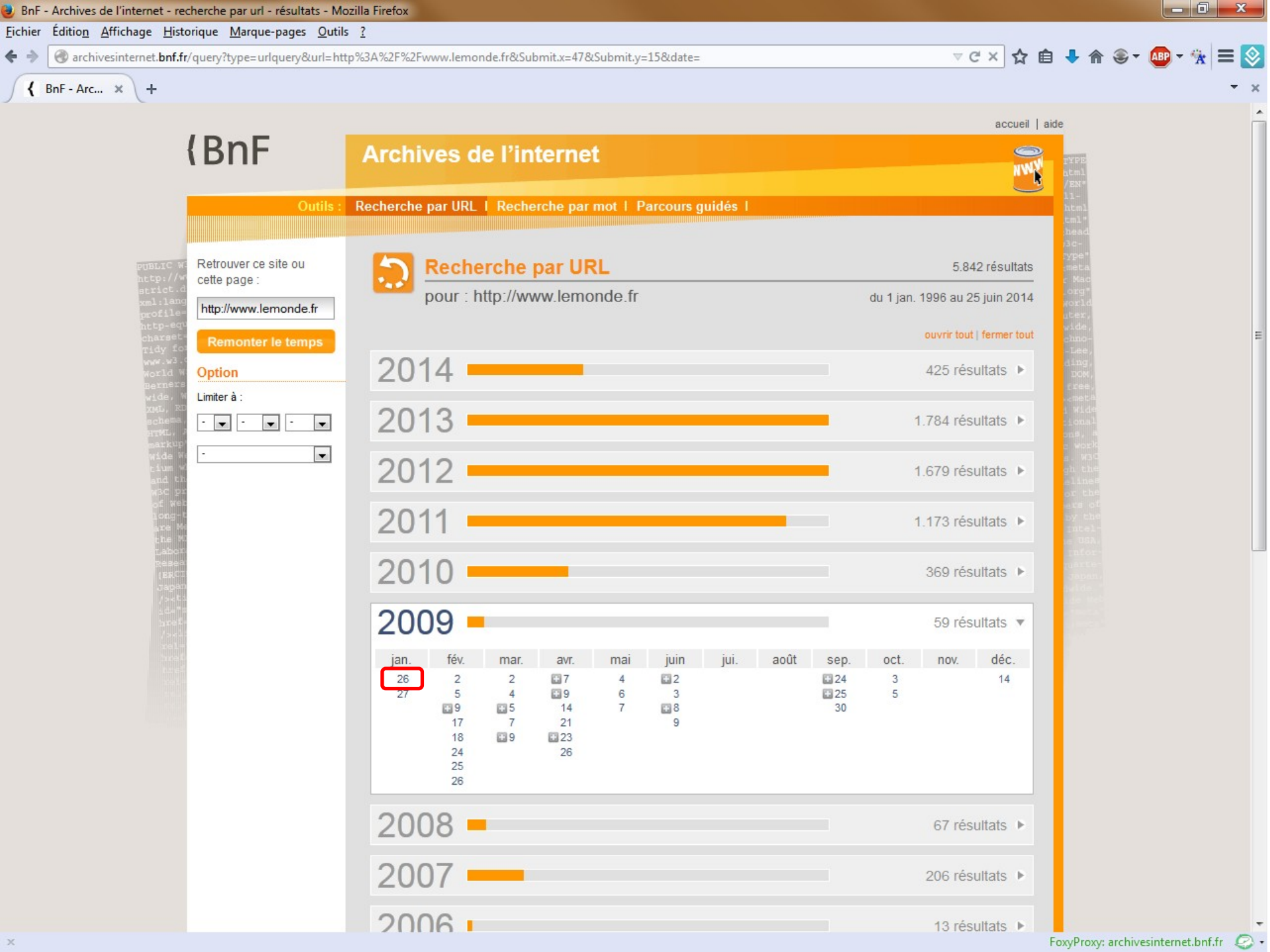
Découvrir le contenu des archives et se familiariser avec les outils de recherche et de consultation.

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.

Carnets de voyage : le monde au bout des doigts

Les carnets de voyage, forme ancienne sur papier, peuvent être aussi divers que les voyages qu'ils racontent. L'internet prolonge les pratiques existantes et apporte de nouvelles


BnF - Archives de l'internet - recherche par url - résultats - Mozilla Firefox
Fichier Édition Affichage Historique Marque-pages Outils ?
archivesinternet.bnf.fr/query?type=urlquery&url=http%3A%2F%2Fwww.lemonde.fr&Submit.x=47&Submit.y=15&date=
BnF - Arc...
accueil | aide
BnF
Archives de l'internet
Outils : Recherche par URL | Recherche par mot | Parcours guidés |
Retrouver ce site ou cette page :
http://www.lemonde.fr
Remonter le temps
Option
Limiter à :
Recherche par URL
5.842 résultats
pour : http://www.lemonde.fr
du 1 jan. 1996 au 25 juin 2014
ouvrir tout | fermer tout
2014 425 résultats
2013 1.784 résultats
2012 1.679 résultats
2011 1.173 résultats
2010 369 résultats
2009 59 résultats
jan. fév. mar. avr. mai juin jui. août sep. oct. nov. déc.
26 27
2 5 9 17 18 24 25 26
2 4 5 7 9
7 9 14 21 23 26
4 6 7
2 3 8 9
24 25 30
3 5
14
2008 67 résultats
2007 206 résultats
2006 13 résultats
FoxyProxy: archivesinternet.bnf.fr

Actualités

Le Monde.fr

Mise à jour à **19h23** - Paris

Le Monde.fr | le web avec YAHOO! Recherche sur Le Monde.fr

Recevez les newsletters gratuites
Abonnez-vous au journal
Le Monde : 16€/mois

Tempête : l'état de catastrophe naturelle pourrait être déclaré mardi

6

De son côté la Commission européenne s'est dite prête à actionner le Fonds de solidarité européen.

"La mobilisation est générale pour les assurances françaises"
La France à l'âge du feu

Les "désobéisseurs pédagogiques" interpellent Darcos

Cent-cinquante professeurs demandent au ministre le retrait du dispositif d'aide personnalisée aux élèves en difficultés.

Sri Lanka : vers la fin d'une guerre de trente ans ?

LES DÉPÊCHES

Toutes les dépêches

19:10 Jérôme Kerviel devrait être jugé en 2010, sans ses supérieurs Reuters

19:01 Le musée d'Auschwitz veut créer un fonds de 100 M EUR pour préserver le site AFP

18:45 Journalistes et avocat tués en Russie: il faut une "véritable enquête" (Unesco) AFP

18:44 France Télévisions candidat aux droits de la Coupe de la Ligue (Bilalian) AFP

EDITION ABONNÉS : 70 fils de dépêches thématisés

VOS RÉACTIONS

« C'est l'arroseur arrosé !!! TF1, qui avait si bien théorisé la fin de la pub (pour son plus grand profit) dans un livre blanc sur mesure pour le président, réclame à cors...

Alex sur La fin de la publicité sur les chaînes publiques ne profite pas à TF1 ni à M6 53

TEMPÊTE

A découvrir | Météo | Bourse

"Le Monde" fait peau neuve

Aujourd'hui, découvrez la nouvelle maquette du quotidien "Le Monde" en kiosque ou avec le Journal Electronique.

En savoir plus

Les blogs invités

- Abalo le Coréen
  LES "EXPERTS" : CROATIE
- Pour Gaëtan Gorce, s'opposer c'est négocier...
  PUZZLE SOCIALISTE
- "Je suis plutôt inquiet pour 2011 et 2012"...
  ENGRENAGES
- Nouvel an chinois
  L'ACTU EN PATATES
- United States of Tara - Moi, elle, lui et elle
  LE MONDE DES SÉRIES

BnF

Archives de l'internet

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://

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).

"Le Monde"

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.

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.

Carnets de voyage : le monde au bout des doigts

Les carnets de voyage, forme ancienne sur papier, peuvent être aussi divers que les voyages qu'ils racontent. L'internet prolonge les pratiques existantes et apporte de nouvelles

BnF

Archives de l'Internet

Retrouver les occurrences de ces mots :

"Le Monde"

Rechercher

Possibles :

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

Recherche par mot

13318089 résultats

pour : "le monde"

résultats 1 à 10 - 10 par page

1 2 3 4 5 6 7 8 9 10 ▶ ▶|

Bonjour tout le monde ! · eureka

http://albin.unblog.fr/2006/11/08/bonjour-tout-le-monde/

... Bonjour tout le monde ! · eureka eureka Bienvenue sur mon blog Accueil A propos Que tout le monde y trouve quelque chose navigation Bonjour tout le monde ! 8 novembre, 2006 Posté par albin dans : Non classé , trackback Bienvenue sur Unblog.fr . Ceci est votre premier article. Editez ... commentaire, identifiez vous et affichez les commentaires de votre blog, et vous pourrez l'effacer. Ce formulaire est protégé contre le ...

Archive du 29/11/2006 à 01:08 | autres dates — html - 19 Ko

Commentaires sur Bonjour tout le monde !

http://accesmodeintime.unblog.fr/2006/09/08/bonjour-tout-le-monde/feed/

... Commentaires sur Bonjour tout le monde ! http://accesmodeintime.unblog.fr/2006/09/08/bonjour-tout-le-monde/ Bienvenue sur le blog harmonie intime entre adultes consentants dont la devise est " un(e) pour tous & tous pour un(e) " interdit ... 01:12:02 +0000 http://wordpress.org/?v=MU par : Admin http://accesmodeintime.unblog.fr/2006/09/08/bonjour-tout-le-monde/#comment-1 Fri, 08 Sep 2006 12:09:43 +0000 http://accesmodeintime.unblog.fr/2006/09/08/bonjour-tout-le ...

Archive du 29/11/2006 à 02:11 | autres dates — xml - 1 Ko

Broc collection - Le monde du reduit et de la collection

http://boutique.broc-collections.fr/epages/ce_fr.sf/secrdZCvGMcgOc/?ViewAction=ViewContactForm&ObjectPath=/Shops/155816

... Broc collection - Le monde du reduit et de la collection Broc collection Le monde du reduit et de la collection Page d'accueil Coordonnées Nous contacter Conditions générales de vente Informations client Politique de ... Adresse e-mail * Sujet * Message * * champs requis Panier 12 Produits 176,30 € Colissimo 6,50 € Montant total 182,80 € Afficher le

Archive du 18/10/2007 à 15:20 | autres dates — html - 11 Ko

Forum le monde de la magie blanche

http://96334.aceboard.fr/recherche.php?login=96334

... Forum le monde de la magie blanche FORUM , Forum Discussion , Forum Gratuit , Nom de domaine , Nom de domaine gratuit , Redirection gratuite , Administrateurs : lune ... en ligne : 1 inconnu visite le forum Inscription | Profil | Messages Privés | Recherche | Online | Aide | Créer un blog gratuit Recherche sur

BnF

Archives de l'internet

Parcours guidés

Mis à jour le 15/06/2012

Tous les parcours

Les parcours guidés permettent de découvrir le contenu des archives et se familiariser avec les outils de recherche et de consultation.

Presse et actualité

Créé le 21/11/2013

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.

Carnets de voyage : le monde au bout des doigts

Créé le 24/06/2014

Les carnets de voyage, forme ancienne sur papier, peuvent être aussi divers que les voyages qu'ils racontent. L'internet prolonge les pratiques existantes et apporte de nouvelles possibilités : visibilité accrue, immédiateté et interaction, variété de médias. Tout en insistant sur les carnets novateurs, ce parcours guidé propose un échantillon qui se veut représentatif et montre comment les voyageurs se sont approprié le web pour renouveler le genre.

Parcours guidé

Carnets de voyage : le monde au bout des doigts

Avec qui voyager ?

Avant de raconter son voyage, il faut l'organiser : décider d'une destination, d'un mode de transport, préparer ou improviser son itinéraire et son mode de vie une fois sur place, mais aussi proposer à une autre « personne » de participer au voyage... ou pas.

Parmi les sites ou blogs sélectionnés sur le web, il faut distinguer les blogueurs qui partent à plusieurs : en famille, en couple comme des jeunes mariés ou avec une classe d'élèves. Certains blogueurs partent seuls accompagnés d'un objet insolite ou de leur animal « favori ». Et puis, il y a les solitaires !

L'aventure Caraïbes en catamaran

Blog d'une famille (les parents et leurs deux filles de 10 et 12 ans) qui voyage en catamaran *le Delphis* sur les eaux des Caraïbes, durant 1 an. Le blog montre leur vie de famille avec de nombreuses photos et commentaires.

http://delphis2.uniterre.com

▶ Archive du 13 juin 2013 à 12:33 | autres dates

Voyage au Burundi

Joelle, l'ainée de 4 enfants raconte dans son blog, le voyage de ses parents, frères et soeurs au Burundi, pendant 3 semaines en 2007, Ce blog a permis de partager la découverte de ce pays méconnu des touristes.

http://burundi-decouverte.skyrock.com

▶ Archive du 13 juin 2013 à 12:33 | autres dates

Thèmes de ce parcours


Consultation
des Archives
de l'Internet
WWW


Archives de l'Internet
Archives de l'Internet

Danke für die Aufmerksamkeit

Fragen?