Archiver le Web

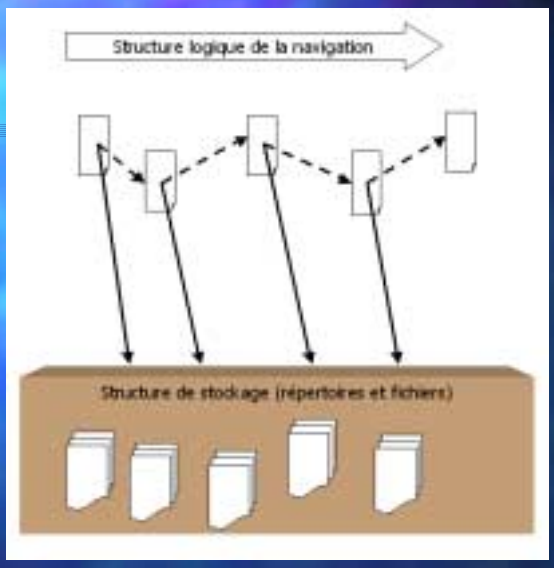
Julien Masanes

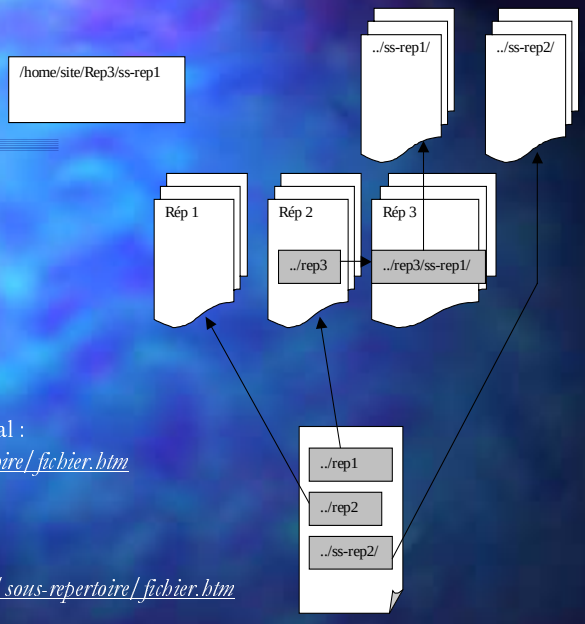
BnF

Présentation Groupe PIN

14 mai 2002

Le Web, un nouvel espace documentarie





Adresse sur le serveur local :
/ home/ repertoire/ sous-repertoire/ fichier.htm

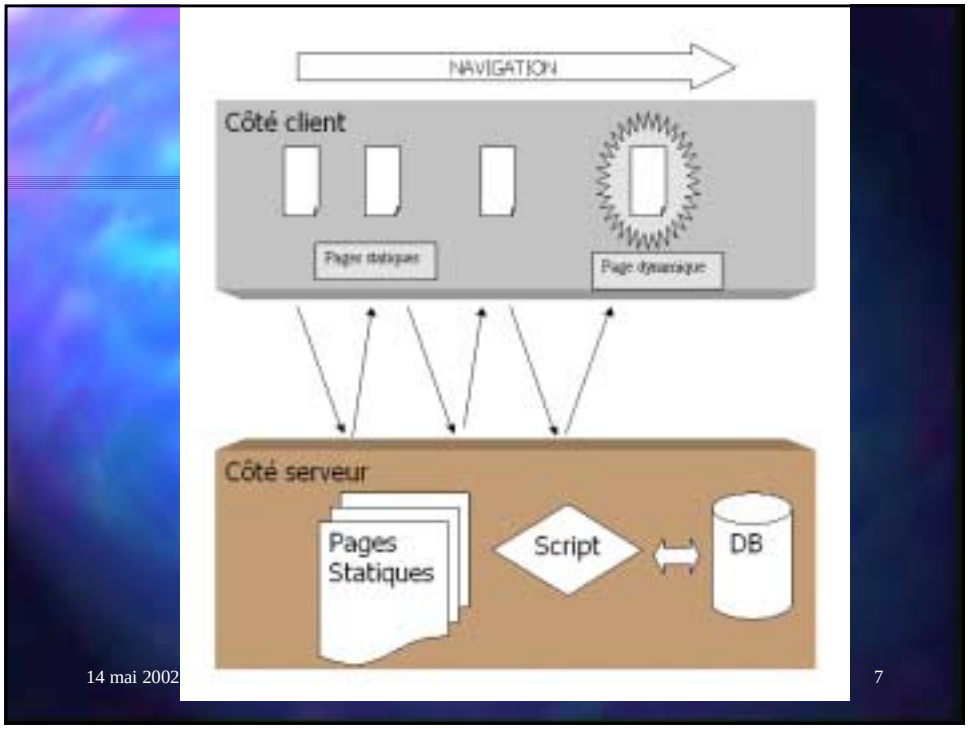
sur Internet cette adresse

http/ /:www.site.fr/ repertoire/ sous-repertoire/ fichier.htm

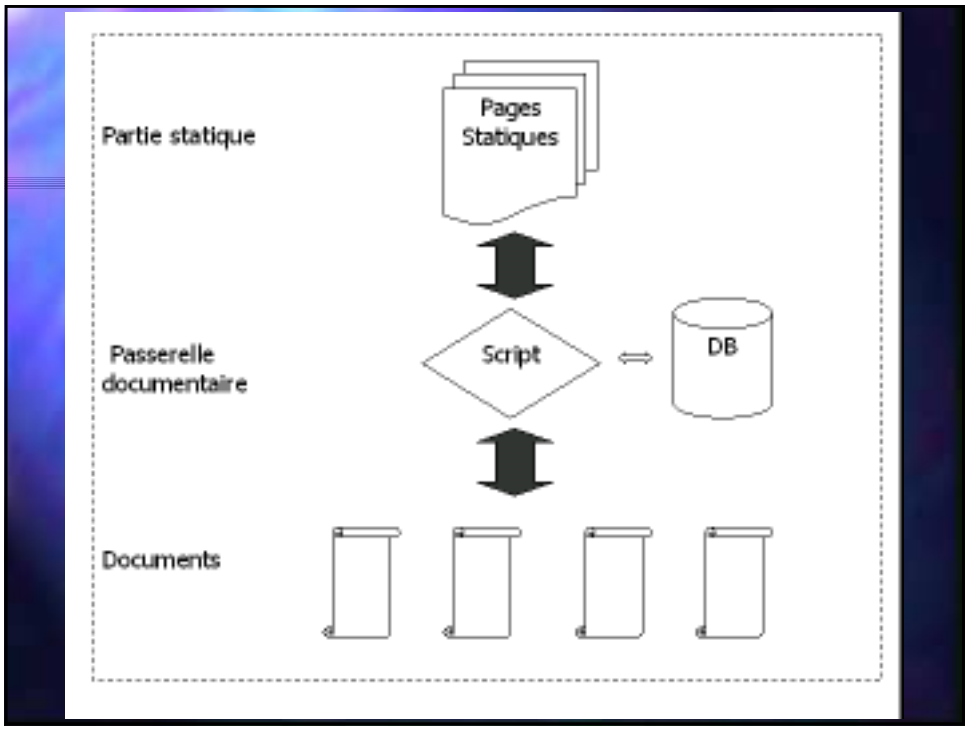
Archivage des sites statiques

- Navigation automatique par un robot
- Sauvegarde de toutes les pages et des éléments liés
- Liens absolus transformés en liens relatifs

Les sites dynamiques


NAVIGATION
Côté client
Page statique
Page dynamique
Côté serveur
Pages Statiques
Script
DB

14 mai 2002 7


Partie statique
Pages Statiques
Passerelle documentaire
Script
DB
Documents

Archiver les sites dynamiques

- Premier type :
  Revient au cas statique, chaque fois que les liens contiennent l'information (paramètres)
- Second type :
  Plus complexe, nécessite de faire un dépôt des fichiers et une migration de la partie base de donnée vers un format pérenne (XML).

Ce que Internet change pour le dépôt légal

- Publications non filtrées par des éditeurs (coût de publication très bas).

-> nécessité pour la bibliothèque d'assurer un rôle nouveau de sélection (mais pas dans une logique acquisition).

- Grand nombre de documents par site (parfois plusieurs dizaines de milliers) et nombre de sites très important

  (300 000 en France).

-> impossible de tout traiter manuellement, il faut utiliser des outils de traitement automatiques.

Ce qui a été fait par les autres pays

- Sélection manuelle de quelques sites (éditeurs classiques, institutions) traités individuellement. Australie, Canada.

-> couverture très insuffisante par rapport à l'ensemble du Web.

->Manque d'outil de repérage global de l'information sur le Web. Risque d'empirisme dans la sélection.

- Collecte automatique régulière (snapshot). Etats-Unis (LC/Internet Archive depuis 1996), Suède (KB) depuis 1996).

->une partie du Web échappe aux robots ('web invisible')

Orientation des expérimentations de la BnF

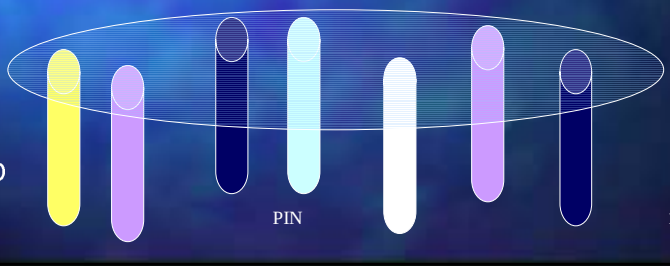
Articuler les deux approches pour améliorer la couverture et le suivi des contenus

- Se servir d'un robot pour archiver les contenus dispersés sur le Web et traiter les gros volumes automatiquement.
- Extraire la cartographie globale du Web pour permettre une sélection et un suivi moins empirique des sites.
- Sur un nombre limité de sites assurer un archivage individualisé pour permettre notamment un dépôt des contenus qui ne peuvent être capturés en ligne.

TTraitement automatisé du Web de surface, collecte en ligne
Pour les sites inaccessibles aux robots, le relais sera pris pour un traitement individuel avec dépôt si besoin

Web

Deep web



1- Adaptation d'un robot collecteur

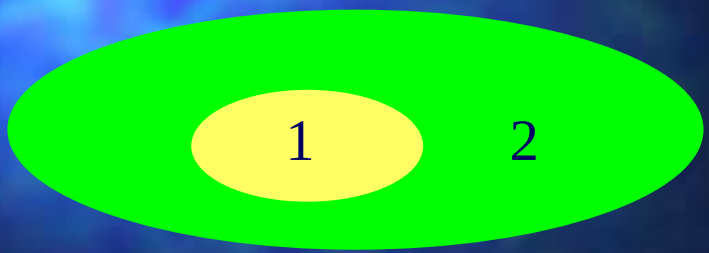
- Conçu comme un outil central pour le DLI
- Devra donner une cartographie globale et des informations pour s'orienter dans la masse documentaire (localisation, volume, notoriété, type de contenus).
- Permettra de traiter de gros volume à archiver et de conserver ainsi une réelle archive du Web (non exhaustive mais assez large pour être représentative).

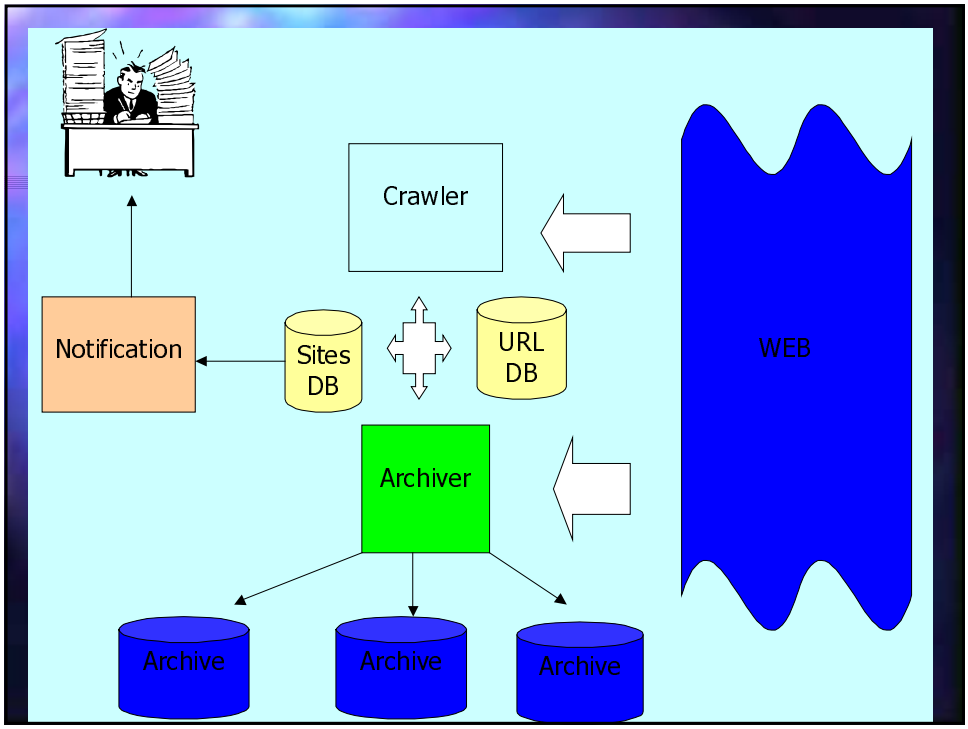
Informations

- Nombre et volumes des sites
- 'Importance'
  Indice de notoriété
  Indice de mots rares
- Détection des sites difficiles à collecter

Sélection automatique

- Délimitation d'un sous-ensemble du Web





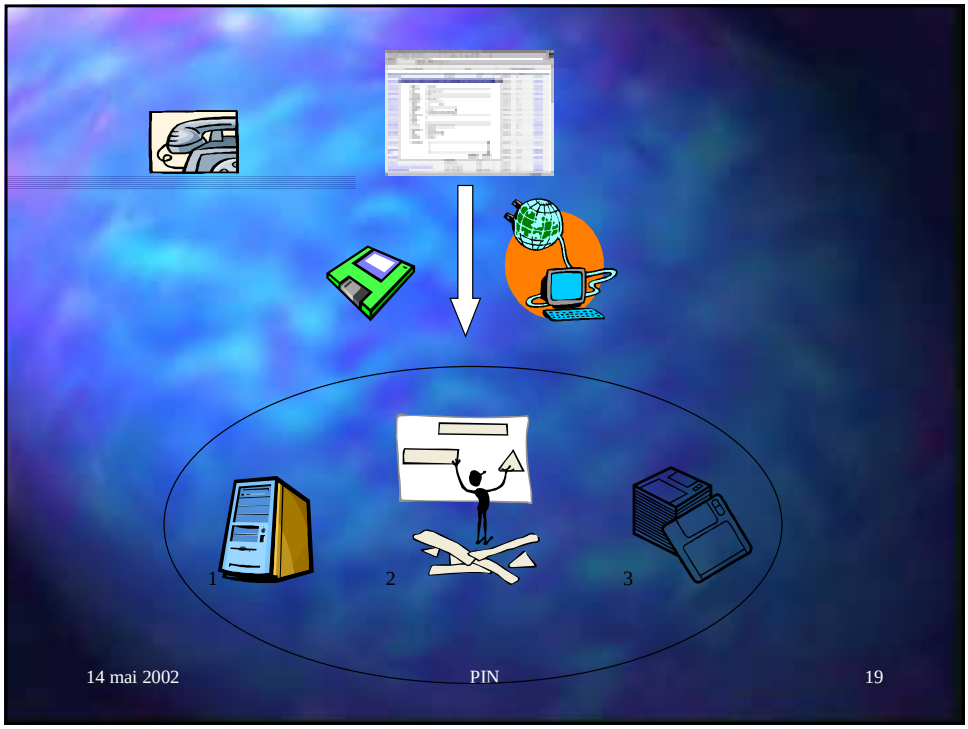
2- dépôt de sites

- Test des limites de la collecte en ligne sur un échantillon sélectionné de sites à fort contenu multimédia.

-> meilleure connaissance des limites techniques

-> collecte en ligne inutilisable pour ce type de sites

- Définition et test des procédures de dépôt avec un échantillon de sites plus large



Avancement

- Plus d'une centaine d'éditeurs de sites contactés depuis l'automne 2001
- La moitié ont accepté
- 25 ont actuellement déposé

Typologie

- Sites Web 'classiques' dont certains sites personnels
- Revues en ligne (EDP Science, Ciel et espaces)
- Sites de vente de e-book (00h00.com)
- Portails (Revues.org, Servicepublic.fr)
- Réservoirs documentaires (Vitaminic, Medpict, arXiv)
- Flux d'informations, chaînes (Le Monde, AFP)
- Lettres d'information (FTPresse)

Collecte thématique des sites électoraux (2002)

- 1200 URL sélectionnées (sites, parties de site, documents isolés) par une quinzaine de bibliothécaires
- Fréquences de collecte adaptées
- Typologie des sites
- Possibilité de suggestions extérieures

- www.bnf.fr/pages/infopro/dli_ECDL2001.htm
- Liste de discussion sur le sujet : web-archive@cru.fr

julien.masanes@bnf.fr