

Archiving the Web : experiments at the BnF

Overview

- The context
- An integrative approach
- Improve the present generation of harvester
- Archiving the surface Web
- Archiving the deep Web

The context

- legal deposit in France
 - books : 1537
 - serials : 1921
 - music, video, multimedia and software : 1992
- French Law modification : June 2001 proposition of a new article to be added to the legal deposit law in. It has not been adopted by the parliament yet.
- Minister of culture has commissioned BnF and INA to experiment the best policy and tools to archive the Web

What's new (with regard to legal deposit) ?

- We have to deal with unfiltered content

We have to assume selection but in a different logic than for acquisition

- Content is dynamic and you have to monitor this change by yourself
- Online gathering of content is, to some extend, possible

An integrative approach

Philosophy : do as much as we can in an automatic way, human intervention takes the relay when necessary.

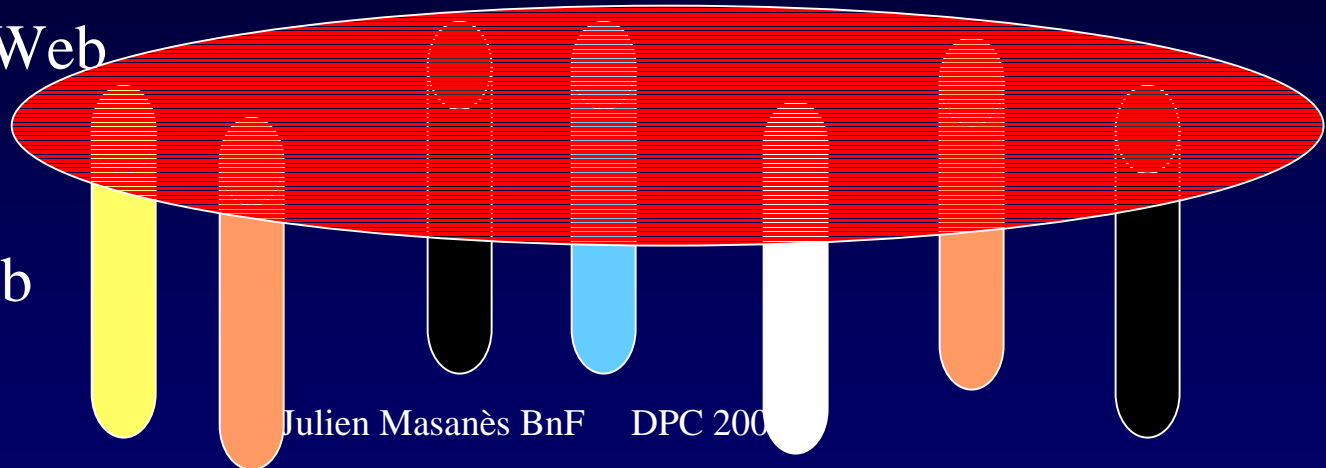
Automatic and manual tasks

Surface Web : Where robot can go and collect content

Deep Web : where they can't go. Manual deposit is necessary

Surface Web

Deep web



Improve the present generation of harvester

exploit more information :

- Linking information
extract the link matrix and make some notoriety estimation based on the number of in-going links to a page or a site.
- Content evaluation particularly lexical information used by search engines (rare words add value)

Improve the present generation of harvester

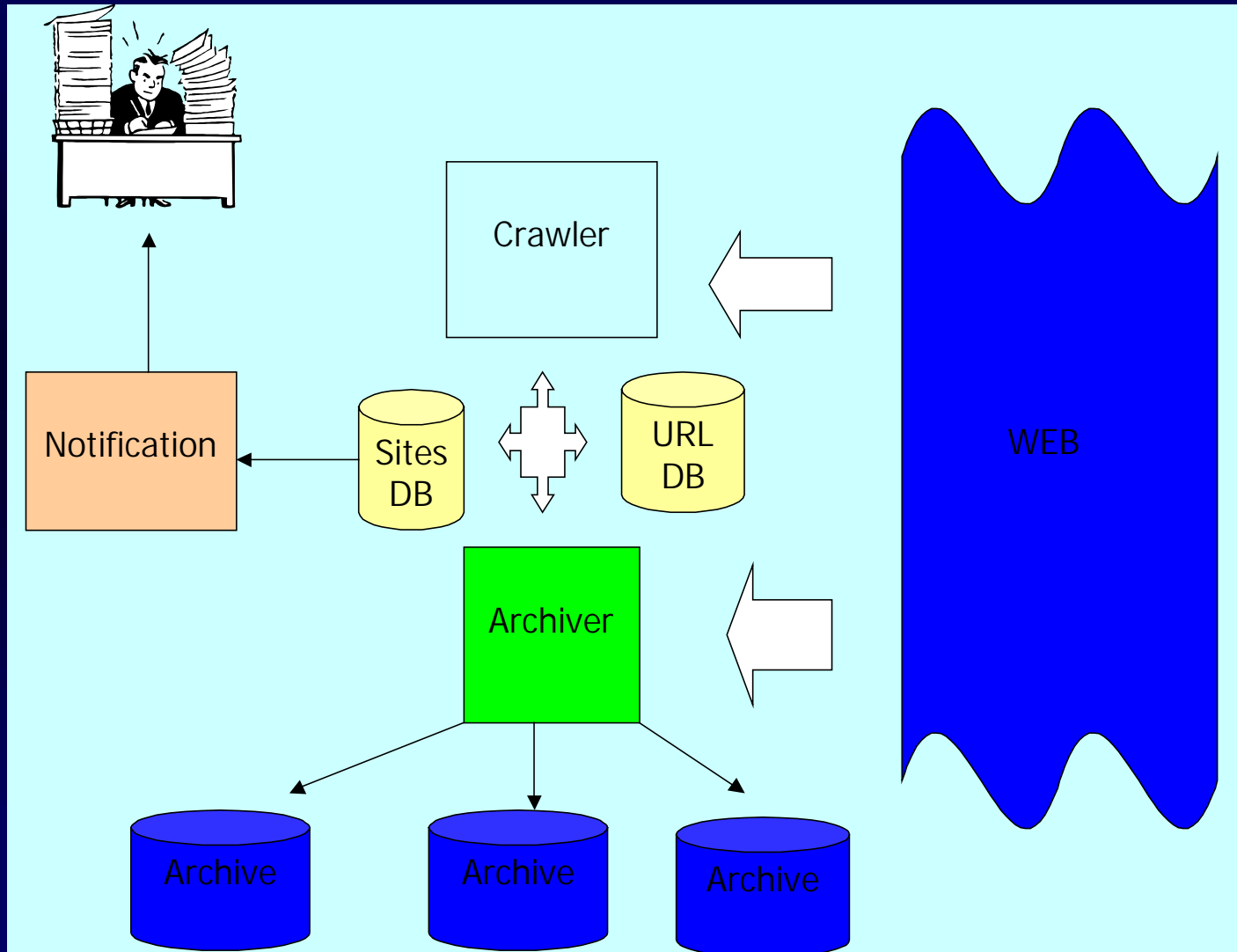
- While crawling, search for technical indice of deep web → in these cases, contact the webmarter for a deposit

Improve the present generation of harvester

With this information it will be possible to :

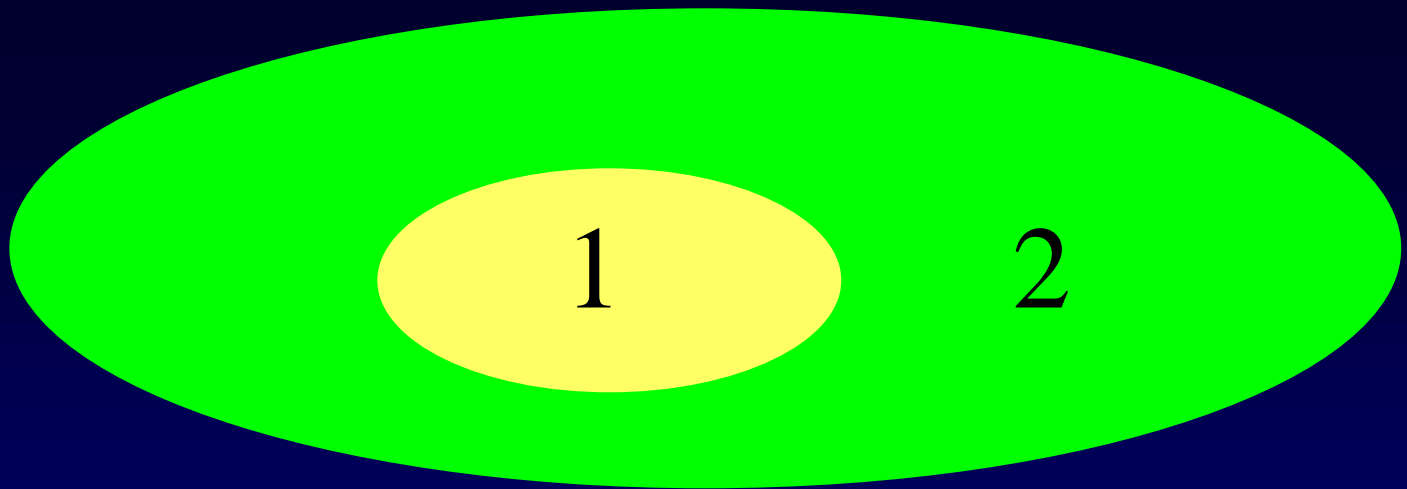
- Define an archiving policy for content that can be treated in an automatic way
- Enhance selection relevance for site that can't be archived online and need deposit

Simplified schema of the robot



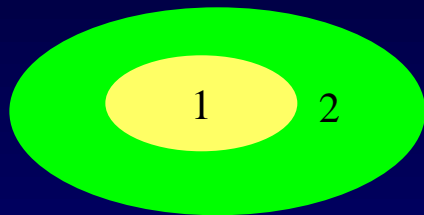
Archiving the surface Web

- On the basis of a valuated topography made by the robot : delimitation of subsets of the surface web



The crawl and the archiving process can be disconnected

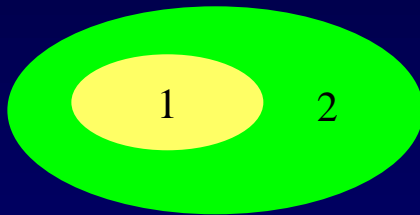
Several policy are then possible regarding crawling and, most of all, archiving.



Crawl policy

You can :

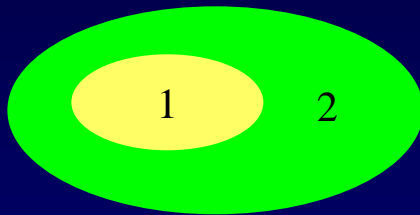
- crawl permanently 1 and 2
- crawl both but with priority to 1
- crawl 2 just once a year and focus on 1




Archiving policy

You can :

- archive 1 and 2 the same way
- archive both but with refreshing 1 permanently
- archive only 1

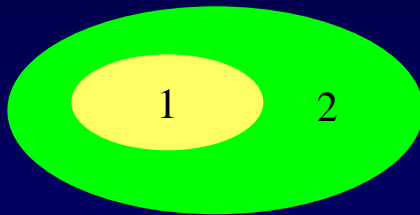


Comparison with the snapshot approach



crawl both but with priority to 1
archive both and refresh 1 permanently

- Archiving capacity needed : more than 1 snapshot but maybe less than 2 :
(depending on the size of 1 and its dynamic)



- Full use of the crawling facility

- Better follow up of subset 1

Advancement

- Feasability study with two robot providers
- Evaluation of notoriety parameter
more than 500 sites evaluated by 8 librarians → correlation with the notoriety factor of these sites (almost finished)

Advancement (2)

- Make a crawl with deep web technical indices detection and explore the possibility of a systematic deep web detection (on the way)
- Harvest the entire .fr domain (june) to make evaluation

Archiving the deep Web

Limits of the deep web

First phase with a panel of 16 sites, selected by the audiovisual dpt (music, videos, multimedia)

- Get from the webmaster information concerning the site
- Try to collect their content with small robots (wget, teleport pro)

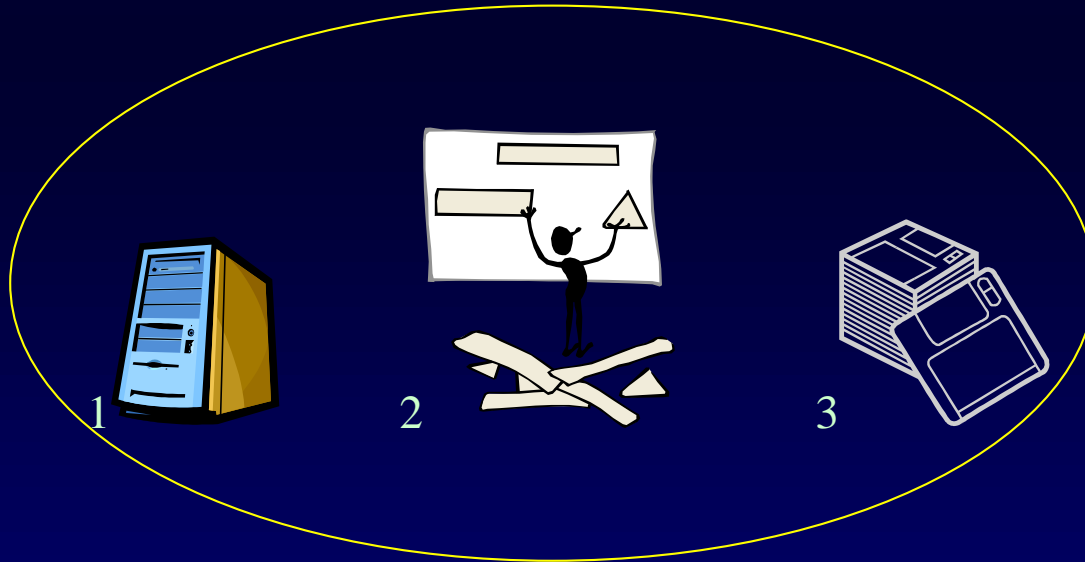
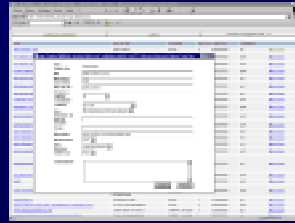
Results :

- 1 was actually loaded
- Few gave some pages and images but the real content was missing
- Many site were absolutely impossible to load
We found obstacles we didn't know (flash site, tricky redirections)
- Some site were unexpectedly enormous (ex vitaminic 600Go)
- Most of the site that weren't yet dynamic were about to migrate to dynamic.
- Few were already using cache streaming techno (Akamai)

Test of deposit procedures

Second phase with a larger panel of sites (100)

- Selected by the every dpt (Science, Litterature, Map...)
- Contact the webmaster and get information about the server architecture and formats used.
- On this basis and in concertation with the services specify one or several deposit (ftp, Cdrom, DVDRom DLT, Mail...)

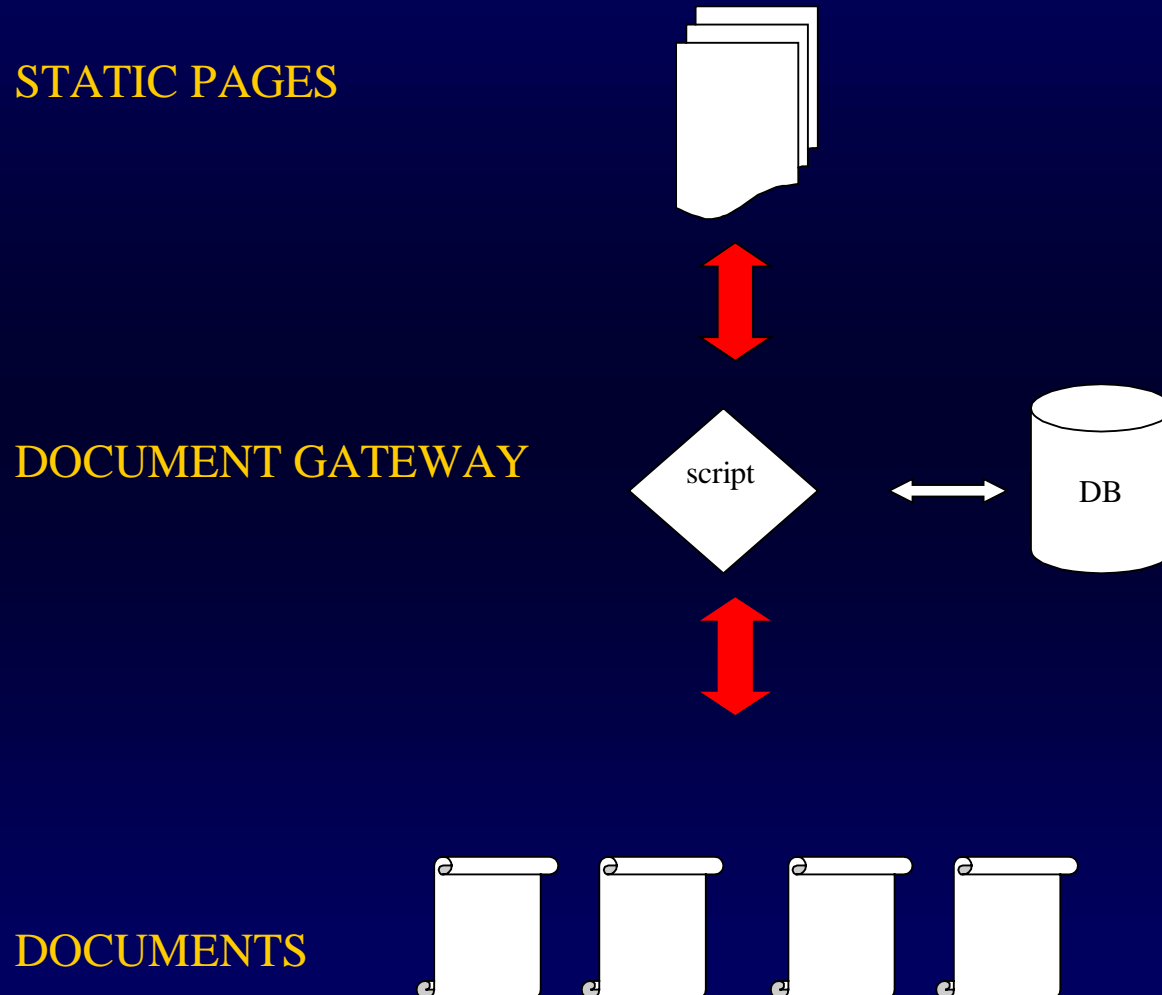


- At the moment 20 sites have made a deposit
- Help us to elaborate criteria and typology
- For some site, a migration procedure is tested

Typology

- Classical Web sites among them few personal ones
- Online serials (EDP Science, Ciel et espaces)
- E-book selling sites (Cytale, 00h00.com)
- Portals (Revues.org, Servicepublic.fr)
- Big content repository (Vitaminic, Medpict, arXiv)
- Content providers (Le Monde, AFP)
- Newsletters (FTPresse)

Migration of dynamic sites



STATIC PAGES

DOCUMENT GATEWAY

DOCUMENTS

