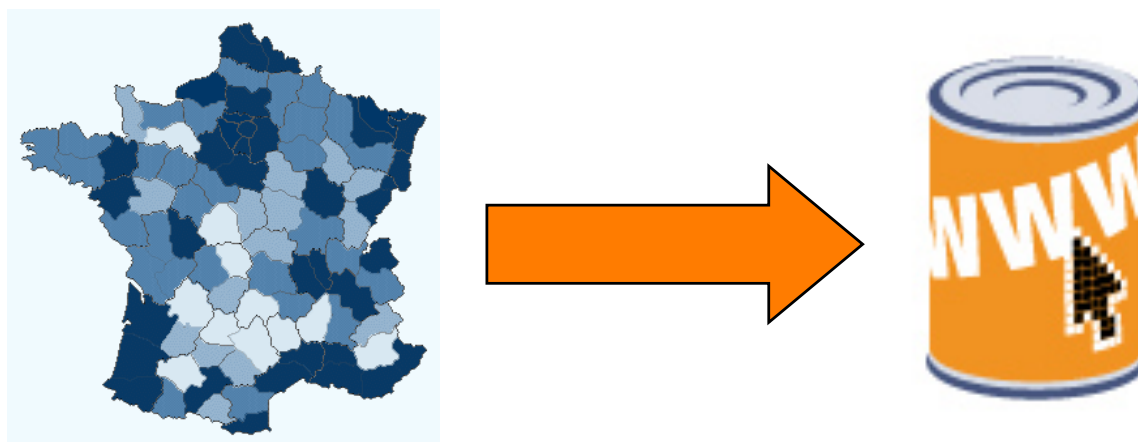


Legal deposit of the French Web

Harvesting strategies for a national domain



International Web Archiving Workshop
Aarhus, Denmark, 18th & 19th September 2008

France Lasfargues, Clément Oury, Bert Wendland

Digital Legal Deposit, French National Library

Legal Deposit of the French Web - IAWW 2008



The French Legal Deposit and the Web

- A five-centuries-old tradition

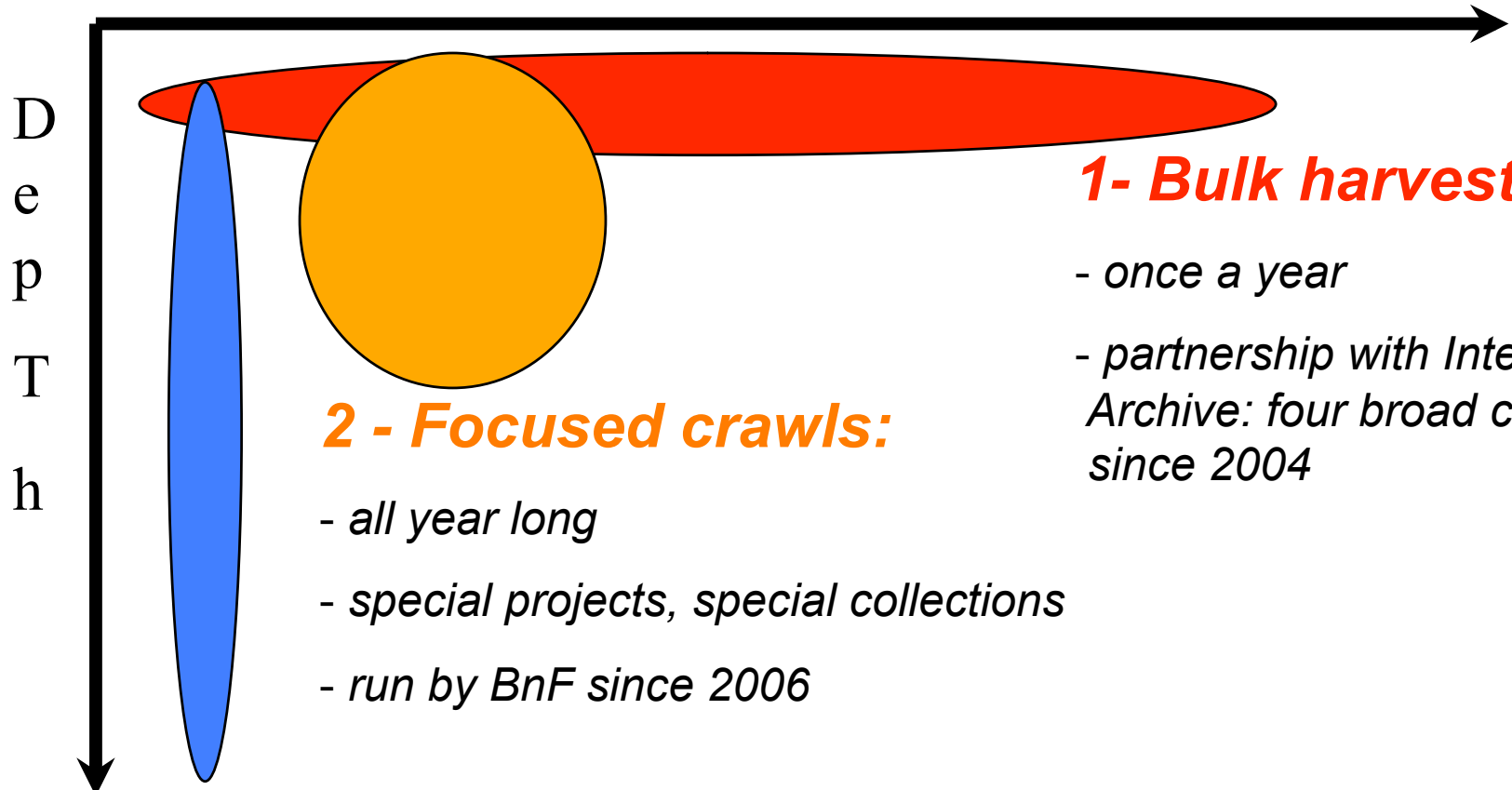
- ... mainly based on three criteria...
 - publication
 - media
 - territory

- ... that are not easily applicable to Web harvesting

- The traditional goal of having an all-comprehensive collection is replaced by the building of a representative collection

BnF's Harvesting "Mixed Model"

W i d t h



1- Bulk harvesting:

- once a year
- partnership with Internet Archive: four broad crawls since 2004

2 - Focused crawls:

- all year long
- special projects, special collections
- run by BnF since 2006

3- E-deposits: - experimental

- expensive

{ BnF

The 2007 broad crawl



Focusing on the “French” Web

- The seed list of the crawl is made of...
 - the all-comprehensive list of domain names hosted on French TLD, thank to an agreement with AFNIC (French registry)
 - the host list coming from the previous .fr broad crawl (2006)
 - .fr host list coming from Alexa crawls extraction
- Crawl scope:
 - .fr and .re TLD
 - redirections from these TLD
 - restricted but economic

Shaping the crawl

→ Defining the crawl settings:
a decisive step to ensure a
representative crawl



→ Host vs. Domain

- host approach: more space to big websites (institutional sites or hosting platforms)
- domain approach (preferred): give more chance to small sites

→ Maximum budget for each domain set to 10 000 URL

→ Taking care of specific media



IA/BnF relationship during the crawl

- Organization of the the crawl: scheduling, monitoring, knowledge transfer
- Sharing experiences and roles in combining technical expertise of IA engineers and policy content expertise of BnF librarians
- Analysis of the “frontier report”, once a week, to:
 - assess the domains reaching the maximum budget of 10 000 URL
 - pay attention to mirror websites and domain names using domain farming
- The idea was not to ensure a 100% quality crawl but to ensure that resources (time, size) were properly spent.

Key figures

→ More data...

→ ...with a different distribution:

- growing number of domains
- less URL per domain but more URL per host

Number of	2006 (host)	2007 (domain)
URL	271 697 456	337 322 200
Hosts	2 928 364	1 589 458
Domains	382 540	1 062 317
(on which .fr domains)	131 136	791 940
URL per domain	710	318
URL per host	93	212
Unique ARC files	73 073	91 745
Compressed size of unique data (Tb)	7,2	8,8

Crawl outcomes in light of initial goals - 1

TLD and French *territory* ?

77% of crawled documents belong to the .fr

But we also had other TLD coming from:

- redirections
- external files (images, pdf, videos...)

(settings allowed to collect these contents)

TLD	Number of URLs	%
fr	259 869 452	77,12
com	59 843 624	17,76
net	4 951 932	1,47
org	3 171 196	0,94
de	2 808 359	0,83
eu	993 546	0,29
info	900 544	0,27
be	660 834	0,2
ch	461 021	0,14
uk	434 315	0,13
re	381 746	0,11
other TLD	2 471 064	0,73



Crawl outcomes in light of initial goals - 2

Crawling depth : *publication*

More small and medium websites,
fewer big websites

- those results fit with the general idea of our legal deposit as national “mirror” : we have sampled everything rather than selected the best or most popular
- small, medium and big got the same chance to be collected
- what about big websites?

Number of URLs	Number of domains
<10	498777
10-100	146356
100-1000	103370
1000-10000	43101
>10000	334

Crawl outcomes in light of initial goals - 3

- Document types and *media* approach
- MIME type analysis:
 - More and more different MIME types : 554 in 2004 crawl vs.1604 in 2007 crawl
 - Decreasing part of GIF in favor of JPEG and PNG
 - RSS
 - Videos (120 000)

MIME Type	Number of URL	%
text/html	229 257 942	67,96
image/jpeg	64 222 287	19,04
image/gif	25 376 262	7,52
image/png	3 955 885	1,17
application/pdf	3 955 463	1,17
text/plain	2 256 759	0,67
application/x shockwave-flash	1 594 342	0,47
text/css	1 432 809	0,42
application/x javascript	1 415 230	0,42
application/xml	1 083 991	0,32
other	2 771 213	0,82

Conclusion: domain settings

- Defining settings for a broad crawl is like writing a collection development policy, but at a *much* larger scale and with different tools
- The settings we chose did answer our goals :
 - Territory: French domain in .fr - but not only
 - Publication: settings proved to be in favor of the small and medium websites. Wider scope, less depth. Like a Robin Hood policy: take the budget from the big websites to give to the small ones!
 - Media: crawl settings allow for specific document type approaches, e.g videos, blogs, etc.

Questions for the future

- Strategies for big websites
 - Harvest the big within focused crawls (mixed model)?
 - Special crawl projects just for big platforms (YouTube...)?
 - More budget for the broad crawls ?

- Clarify and validate goals by involving other BnF librarians in a discussion on what should be in broad crawls

- Develop and support new tools for crawling *smarter* (de-duplication, queue priority management..)

- Explore other TLD domain name identification and localization strategies