Conducting an event-based web archiving project: the example of the French national elections crawl

➔ A growing number of political websites emerge and disappear on the World Wide Web. Political parties and candidates as much as journalists and citizens express themselves on the better ways to match the problems of the nation. These debates played a major role in the 2007 French national elections: presidential elections and parliamentary elections.

➔ However, there is a strong risk to loose the memory of these manifestations of democratic vitality. This is why the French national Library (BnF) decided to archive the 2007 electoral websites. Relevant websites were identified by a team of librarians and harvested by the BnF Digital Library department and IT department staff.

➔ Project features :

- Duration : ten months, from October 2006 to July 2007
- Staff involved: more than 40 librarians, at partial time (form 3 hours per week to 4 days per week)
- For BnF, this project was a large-scale test-bed to build its internal crawling infrastructure and enforce awareness within and outside the library.

Websites selection criteria

Selected websites sample shall:

- deal with 2007 presidential and parliamentary elections;
- be significant and representative in this respect;
- be regularly updated;
- ensure a fair balance between the different political trends and ideologies;
- be located in France (authors, publishers) = legal deposit;
- be French speaking (+ regional languages);
- ensure representativity from different geographical areas of the country;

Policy guidelines:

BnF defined a typology aimed at controlling the overall constitution of the sample as the collection was being harvested. This typology relies on an analysis of the public space of the political web rather than on a differentiation on websites by editing techniques or file types.

- ◆ 0 – Government publications, rules of the campaign
- ◆ 1 – Candidates and their organisations
  - 1.1 Campaigning candidates: official or personal, temporary or permanent websites
  - 1.2 Political organisations websites
  - 1.3 Other supporting organisations
- ◆ 2 – Opinions and points of view upon the campaign
  - 2.1 Observatories, polls, research institutions and analyses
  - 2.2 Traditional medias and press
  - 2.3 Associations, trade unions and other non-campaigning organisations : pre-existing organisations which have an opinion about the campaign
  - 2.4 Individuals and communities expressing themselves on the Internet (opinion blogs, comics, caricatures…) – and only there.

Cost of project

Human resources

Skills and activities	Implied agents	Hours spent	Global costs
Project definition (March-September 2006) Defining the project goals and organisation, setting selection criteria, developing the curator tool	4 agents	774 hours (110 men / days)	34 150 €[1]
Seeds selection This work is done by collection librarians from Law, economics, politics department and from Philosophy, history, social sciences department. Two major tasks are required: - Websites URL identification - Choice of the crawl settings : depth and frequency of the harvest Librarians should also fill some description fields (type of site, name of the candidate (if any)…). The goal was to ensure the consistency of the collection sample, not to catalogue the websites.	18 agents	2 160 hours (310 m/d)	57 730 €[2]
Websites archiving The Web Legal Deposit team is in charge of the technical execution: 4 librarians of the digital library department, 2 engineers of the IT department. - technical validation of the websites harvest proposals - scheduling - harvesting and monitoring - quality assurance - indexing - data storing on access and preservation servers	6 agents	3 450 hours (492 m/d)	152 215 €
Total for BnF	**24 agents**[3]	**6 384 hours (912 m/d)**	**244 095 €**
Librarians from local libraries From April to July 2007, 8 libraries from different French regions were associated to the project. Their agents worked the same way as BnF librarians, identifying and selecting websites to be archived by BnF.	15 agents	800 hours (115 m/d)	21 382 €
Total of costs for human resources	**39 agents**	**7 184 hours (1027 m/d)**	**265 477 €**

Technical resources

Current resources and upcoming needs	Need	Designation	Global costs
Harvesting The current harvesting infrastructure was sufficient to collect the wanted material. However, it will be necessary to increase our capacity to match the needs of broader crawls in future years.	Hardware	6 computers Processors: 3 GHz Hard drive: 518 Gb RAM: 2 Gb	5 320 €
	Software	Heritrix	0 € *(open source)*

[1] Figures based on the average salary of a curator / engineer (67 945 € per year, 220 working days in a year).

[2] Figures based on the average salary of a cataloger (people in charge of selection and identification ranked from the storekeeper to the curator), 41 160 € per year, 220 working days in a year

[3] The 4 agents quoted in the project definition are also included in « Seeds selection » and « Websites archiving » tasks.

	Bandwidth	40 Mb/sec Since July 2007: 100 Mb/sec	13 350 €
Indexing The harvesting capacities were superior to the indexing capacities, hence a recurrent problem of saturation. It is necessary to increase the computing power at the indexing step. These needs should also grow dramatically with full-text indexing.	Hardware	“Wayback Machine” indexing: 1 computer Processor: 3 GHz Hard drive: 518 Gb RAM: 2 Gb	940 €
		Since May, 2007: 1 computer for full text indexing. Processor: 4x2,6 GHz Hard drive: 1,2 Tb RAM: 4 Gb	6 375 €
	Software	Arc2Dat / Dat2CDX	0 € (delivered by par IA on Petabox n°1)
		NutchWAX	0 € (*open source*)
Back-up and Long term storage Data was stored on the BnF long-term storage repository: SPAR (“Preserving and archiving distributed system”). At this time, this system only insures the bit stream preservation.			Costs included in the global building of BnF digital repository
Access Data was stored for fast access on the Petaboxes, large-scale data repositories delivered to BnF by Internet Archive together with the data of the *.fr* broad crawls of 2004, 2005 and 2006. Petaboxes are designed by Capricorn Technologies.	Hardware	Empty space on Petabox n°2	5 920 €
	Software	Wayback Machine	0 €
Total of costs for technical resources			**31 905 €**

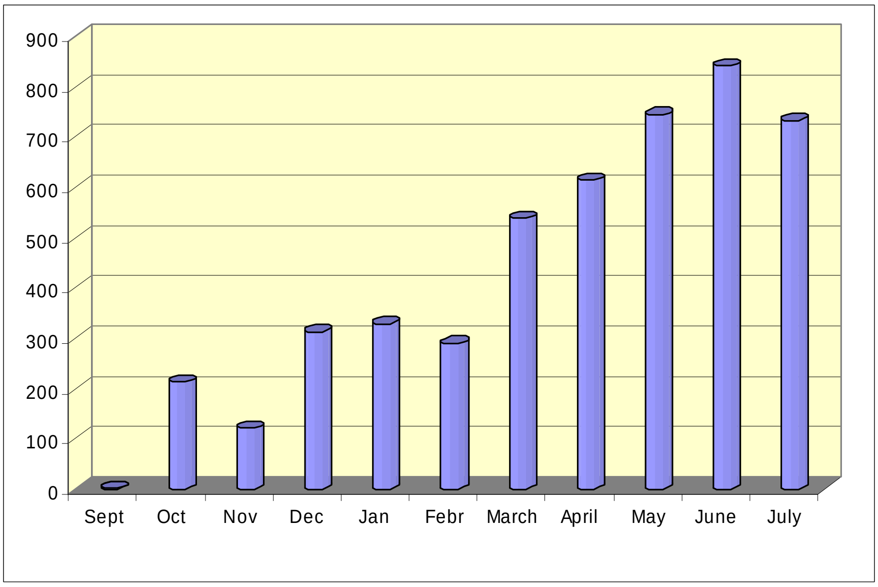
Projects outcomes and benefits

Key figures	
Chosen websites Number of websites (seeds) proposed by librarians to be totally or partially harvested	**5 813**
Harvested hosts Number of different hosts on which at least one URL was harvested	**66 223**
Harvested URL	**63 519 520**
Compressed size of harvested data (in Tb) This total is similar to the size of the annual broad crawls of the *.fr* made by Internet Archive on behalf of BnF	**3,38**

The building of a Web archiving infrastructure

As the project was a large scale test-bed for the building of BnF web archiving infrastructure, it was necessary to increase progressively the amount of data collected – this condition was allowed by the long duration of the project.
The web legal deposit process was constituted step by step: the first function managed by the Library was web files harvesting, then Wayback Machine indexing, bit-stream preservation… Full-text indexing and access in reading rooms are scoped in future month.
This infrastructure should be used:

- in 2008: persistent focused crawls on websites selected by librarians in their thematic fields, other event-based focused crawls;
- in 2009: broad crawl of the *.fr* domain.



Data harvested per month in Gb (uncompressed), 2006-2007

Other results

➔ Communications: the web campaign strategies of the different candidates were largely discussed by newspapers and audiovisual press. The library organised a press conference about the project on October 2006.

➔ Strong networks were built thanks to this project. Within the library, almost twenty people were associated to the selection task; most of them will pursue this job for the persistent focused crawls in their thematic fields. Moreover, this project experienced a new way of collaboration between the national library and local libraries, and with scholarly communities as the political sciences institute ("Sciences-Po"). Partnerships were also established with institutions not traditionally associated with BnF, such as the Internet Rights Forum (Forum des droits sur l'internet: http://www.foruminternet.org/), an association grouping delegates from public administration, lawyers, technical experts and major actors of the web (well-known bloggers…).