

A case study for harvesting e-diaries: report from BnF



France Lasfargues

Digital Legal Deposit, National Library of France



Agenda

- How the project came to us
- Project organization and roles
- Defining scope and selection criteria for the e-diaries project
- Project workflow and harvesting
- Feedback and lessons

How the project came to us

- BnF crawl strategies: mixed model combining broad and focused crawls

- E-diary: a focused crawl made with an external partner: l'association pour l'autobiographie (APA)
 - A French non-profit foundation working on the preservation of diaries
 - A meeting place for researchers and writers
 - Diaries are going online, they became even more ephemeral and fragile
 - Their preservation is at risk!

- Key dates of the partnership:
 - May 2007: first meeting APA/BnF
 - July 2007: definition of role distribution and selection criteria
 - August 2007: first focused crawl (100 websites)
 - September 2008: project became on-going

Roles distribution

- ➔ Share responsibility between different actors (members of the foundation, librarians from BnF literature department and web team)
 - APA and literature curators:
 - define criteria
 - make a list of selected websites
 - Web team:
 - organize the production
 - check URL and validate technical feasibility
 - plan and monitor the crawl
 - After the crawl, literature curators:
 - control the quality of crawled data
 - promote the project

Selection criteria

→ Content based criteria:

Blogs which have been selected are not only diaries but must express an individual point of view

- Must claim his singular and subjective expression
- Quality of writing expression
- Blog must be regularly updated

→ Examples :

- <http://www.cachemireetsoie.fr>
- <http://julie70.blogspot.com>
- <http://www.desordre.net>

Workflow and harvesting

→ Scheduling:

- twice a year
- jobs are distributed on different servers according to depth size and sites with technical obstacles

→ Special parameters:

- include all the blog archives
- special scripts for videos from platforms such as dailymotion, youtube
- special scope to have comments and images which are stored on other hosts

→ 2008 figures:

2 crawls

30 days

459 websites

5 358 108 URL

172 Gb

{ BnF

Browsing e-diaries in the Archive

→ Different kind of problems

- collect js, video, flash
- display the data collected

Feedback

- Good feedback from the different actors of the project and the real desire to continue this collection
- The e-diary collection is available in BnF reading rooms:
 - researchers, writers come to browse the collection and actually become the first public to use the web archive
- BnF/APA project to highlight this collection and create *guided paths*
- Contact with producers: a way to exchange around our practice
- APA wrote about this experiment both
 - in their journal
 - on their websiteand speak about it in a blogger meeting

Lessons

- Specific projects bring useful answers for our harvesting practices (templates...)
- A virtuous circle for everybody
- How do we change a « project » into a living collection?