{BnF

Meeting researchers' needs in mining web archives: the experience of the National Library of France

Sara Aubry, IT Department
Peter Stirling, Legal Deposit Department
Bibliothèque nationale de France

LIBER Annual Conference
Lille, 5th July 2018

{BnF

1. Web archives at the BnF
   - Access services
   - Research projects : use cases
2. Legal context and framework
3. Organisational questions
4. Technical aspects
5. Lessons learned

{BnF

Web archives at the BnF

- Captures of websites at a moment in time
  - Pages, images, style sheets…
  - Use of the open source crawler Heritrix and the WARC format (ISO 28500)
- Governed by legal deposit legislation
  - Definition of the “French” internet
- Constitution of the collections
  - Annual domain crawls
  - Focused crawls of sites selected by librarians and partners
  - Historical collections from Internet Archive (1996-2005)

Access services

- An application called *Archives de l'internet*
  - Based on OpenWayback (open source)
  - Controlled environment with dedicated but regular web browser and plugins
  - URL search, page display and browsing
  - Full-text search for a small number of collections
  - "Guided tours" to present collections

- Well suited to use of individual captures as documents rather than to large-scale analysis

{BnF

Research projects: use cases

- Cartography of websites on the First World War and analysis of a discussion forum (Télécom ParisTech, as part of the cluster of excellence The Past in the Present)
  - https://hal.archives-ouvertes.fr/hal-01425600
- Web90, project studying “Heritage, memories and history of the web in the 1990s”(ISCC)
  - http://web90.hypotheses.org
- Néonaute: a search engine to study the use of neologisms (Paris 13/Université de Strasbourg)
  - Focus studies on identification of neologisms, use of recommended terms and the feminisation of words for jobs, titles, grades and roles

{BnF

Researchers: who's who?

- Academic research teams
  - single field of study or cross-disciplinary
  - from humanities to automated language processing
- Research engineers / data scientists
  - with technical skills
  - hired for the period during which the project is carried out
- Project sponsors and funders

{BnF

Legal context and framework

- Access is controlled under legal deposit, intellectual property and data protection legislation
- Collections are accessible onsite in BnF research library reading rooms and in a regional library network
- Admission is granted to anyone giving a proof of academic, professional or personal research activities
- Users can search/view/cite but not download documents

{BnF

Legal context and framework

- Aim to allow analysis of web archive collections while respecting the relevant legislation
- Use of research agreements
  - List the data and metadata to which researchers have access
  - Conditions of use, both of data onsite and exported metadata and results
  - Define organisational aspects and responsibilities of all parties
- Requires signature by the BnF and partner institutions in the project

{BnF

Organisational questions

- Physical reception of researchers
  - No dedicated reading room for the project
  - Use of internal meeting rooms for discussions and training rooms for hand-on workshops
  - Accommodation of research engineers in offices close to web archive team (legal deposit or IT)
- Work organisation
  - Opportunity to develop new tools and services in parallel: metadata generation, full text search and corpora identification

{BnF

Organisational questions

- Use of agile methodology and specifically Scrum project management (also used for IT projects at BnF)
  - Shared monthly sprints with daily or weekly checkpoints
  - Initial planning and review at the end
- Support and accompaniment
  - Meetings and exchanges with BnF staff
    - Content curators: collections scope and content
    - Crawl operators: how the collections are built
    - Metadata and format specialists: how the data is described and stored
    - Technical support: how the data can be accessed and parsed

{BnF

Technical aspects

- Exchanges on technical aspects are adapted to researchers' skills and objectives
- Digital equipment
  - Development environment
    - Virtual machine running Linux, same as BnF developers
    - Extended permissions to install tools and libraries
  - Runtime environment
    - Physical machine with adapted memory and storage (SSD)
    - Only single machine environment for the moment
  - Ensure collection security, still offer read access to data

{BnF

Technical aspects

- Coding
  - Started with metadata but no tools
    - Identify, test and use middleware applications
  - Python seems to be THE language for TDM and machine learning
    - BnF developers are Java experts, need for transfer of skills
- Narrowing and scaling
  - 31 billion URLs, 1 PB of data – too big to be mined all at once
  - First identify which parts can answer the research question
    - Based on collection building procedures and processes
    - Crawl on demand service
  - Define representative subsets (1%, 10%) to smooth out the processes

{BnF

Lessons learned

- Different projects have different needs
  - But individual solutions take more time and resources
- Cooperation promotes the exchange of technical expertise
  - Possible re-use of applications, tools or pieces of code to improve BnF general access tools
  - Technical “toolkit” of software and idea of infrastructure needed
- Need to simplify organisation to answer researchers’ needs
  - Service rather than co-development
  - Use of data and metadata: licence rather than agreement?
- An institution-wide solution to offer a coherent service
  - “Corpus”: four-year BnF project to provide digital corpora to researchers

{BnF

Questions?

twitter.com/DLWebBnF

depot.legal.web@bnf.fr

www.bnf.fr/fr/professionnels/innov_num_dl_internet.html