Web Archiving Workshop

Leïla Medjkoune, Internet Memory
Fiat/ifta, London, September 2012
What is a Web archive?
• A copy of a website
• Recorded by a crawler
• At a specific date and time
• Look and feel like a real website

Whom?
Any institution with the aim:
• to collect and preserve web material
• for historical, cultural, heritage or legal (compliance) purpose

Why?
• Pervasive
• Dynamic
• Valuable
• Web content

variety of format

Ephemeral
2010 European survey (LIWA)

- 74 heritage institutions answered (26 Broadcasting Archives)
- 55% are engaged in Web Archiving (30% fully operational)
- 71% conduct selective / thematic crawls
- 41% provide online access without restrictions
Context

Explosion of Social media:
• Facebook (n°1): 250 millions of pictures/day
• YouTube (n°2): 830 millions of videos/day
• Twitter (n°15): 175 millions of Tweets/day
• LinkedIn (n°29): 135 millions profiles

Multiplication of production means: increase of user generated content
Technical challenges

Capture
• Dynamically generated content, deep web, etc.
• Non HTTP protocols (e.g.: RTMP)
• Social media platforms, ...

Access
• Replicate live functionalities and look & feel
• Provide access to very large files

➡ Fast evolving technologies
➡ Ephemeral content
Solutions?

- Execution based crawling (vs parsing)
- API crawling
- Application aware crawling
- Bespoke fetchers

➡️ Orchestration of tools
Solutions?

Access tool:
• Player replacement: reproduce players functionalities
• Adapt access solution to type of content/platforms (generic solutions)

Storage infrastructure / format:
• Enable access to large files
• Fast access to large amount of content to facilitate search & retrieval
Web Archivists?

- Find new selection methods and tools
- Find new crawling methods and tools
- Adapt crawl frequencies
- Create access solutions
- Tackle new legal issues
Web Archivist tools

• Open source tools:
  – Netarchivesuite,
  – Web Curator tool, etc.

• Outsourcing services:
  – Archive-it,
  – ArchivetheNet (AtN), etc.
ArchivetheNet

• Shared platform designed as a Software as a Service (SaaS)

• Management of Web Archiving workflow, from collection management to crawl, quality assurance, reporting and also hosting and access

• Friendly user interface

• Dedicated QA and crawl engineering teams
ARCOMEM: ARchive COmunities MEMories

EU funded project, under the FP7 program, research groups and two audiovisual archives: SWR & Deutsche Welle:

• innovative models and tools for social Web driven content appraisal and selection, and intelligent content acquisition
• novel methods for social Web analysis, Web crawling and mining, event and topic detection and consolidation, and multimedia content mining
• reusable components for archive enrichment and contextualization
• two complementary example applications, the first for media-related Web archives and the second for political archives
ARCOMEM Archivist tool?

Set and follow web archive campaigns

• V1: A crawler cockpit and a search and retrieval application

Intelligent content acquisition:

• Seeds URLs
• Keywords
• Social web sites APIs
• Social Media Categories (SMC)
Search and retrieval interface:

- Advance search functionalities
- Filtering via faceting
- Sorting by content type, Social media platform, text/image contextual information (event, entity,...), etc.
Crawler Cockpit Interface

- Create/select a campaign
- Describe campaign (title, description, comments, etc.)
- Define scope: select criteria such as language, keyword, url, organisation, etc.
- Select social media categories and APIs to explore
- Set precedence rules for some content type or source (images, videos, tweets, news, etc.)
Crawler cockpit interface

http://ia200134.eu.archive.org:5000/cockpit/
ARCOMEM Archivist Tool V2

• Refinement mode: Refine crawl parameters to improve crawls
• Improve access application (SARA): Preview function so that the users can review the results of the campaign set up
Research

ARCOMEM: Collect-All ARchives to COmmunity MEMories

Longitudinal Analytics of Web Archive data

SCAlable Preservation Environments

LiWA
Living Web Archives
Thank you
Any questions?

Leïla Medjkoune, Internet Memory