Web Archiving
The evolution of web archiving

What was the most interesting aspect you learned from reading that paper?
1. Potential

2. Approaches

3. Requirements

source: Scott Maxwell
Potential
The Potential of Web Archiving

1. Who are the stakeholders interested in/performing web archiving?
2. What are their interests? Why do they consider web archiving important?
3. What are their requirements?

source: Paraschivu, Florin
The Potential of Web Archiving


The Potential of Web Archiving

1. Who are the stakeholders interested in/performing web archiving?
2. What are their interests? Why do they consider web archiving important?
3. What are their requirements?
4. Why do you consider web archiving important?
Approaches
Overview

Stakeholders

Services

Approaches

Software

Selection Policies

Selection Criteria

Web Crawling

Standards

Heritrix

[Images of stakeholders, services, approaches, software, selection policies, selection criteria, web crawling, standards, Heritrix]
Stakeholders

source: Hugaholic

source: Tilemahos Efthimiadis

source: Ryan Wick

source: Holger.Ellgaard

source: Chris Beckett

and many more ...
Services

see also: https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives
Approaches

- Client-side archiving
- Transactional archiving
- Server-side archiving

source: future15pic
Software

collecting
- PANDAS
- Web Curator Tool
- NetArchiveSuite
- HTTrack
- Heritrix
- MemoryBot
- Nutch

indexing
- NutchWAX
- SOLR
- Elasticsearch

accessing
- Wayback
- Memento
- Kibana
- Zotero
- Diigo
- wget
- warctools

sources: WP Bakery Group (1), Chmee2 (2), Visitor7 (3)
Selection Policies

domain collections
● country
  ○ national domain (.uk, .pt, …)
  ○ hosting country
  ○ abroad, content focus
● other
  ○ .edu, .ac.uk, …
  ○ nhs.uk, nasa.gov, …
  ○ geocities.com, …

selective collections
● collections of individual websites
● theme or subject
e.g., digital artists, international development organisations, …
● event
e.g., olympics, elections, …

+ potentially most comprehensive
  - yet, often incomplete
+ relationships/links with sites
+ focused resource usage
+ higher selection quality
+ sites more likely complete
  - external links likely broken
  - selector bias
Selection Criteria

goal: operationalising high-level selection criteria
e.g., “British web sites”

aspects:

- seeds
- scope (domain, file type, file size, path depth, seed distance, URI scheme, prerequisites, …)
- inclusion/exclusion rules
- extraction (HTML, CSS, JavaScript, …)
- error handling (retries, logging, …)
- politeness (robots.txt, delay, bandwidth, …)
Web Crawling

- priority
- politeness
- load balancing

- request/response
- error handling
- link discovery

source: Dnet Svg

- seeds
- robots.txt
- selection criteria

- WARC
- indexing
- long-term preservation
Standards

URI: http://robots.txt

WARC, DOI, OAIS, Marc 21, Dublin Core, PREMIS, METS

sources: W3C, daPhyre
source: Dan Han (https://hhddkk.wordpress.com/2012/05/29/nutch-vs-heritrix/)
Requirements
Task

practice: to identify requirements for potential web archive users and get an overview on different use cases

1. read your use case, potentially do some research
2. identify requirements, constraints, challenges, etc. by considering the three phases of collecting, indexing, accessing
3. identify at least three open questions - both towards a web archiving consultant and your user

1. discuss your questions with the consultant

3. draft your solution, which should comprise the following information: type of crawl/archive, what to archive (as concise as possible), when/how often, who could do it, which technology, challenges/open questions
Crisis Maps

Amy is a social scientist working closely together with emergency relief organisations. She is interested in analysing the dissemination and evolution of crisis maps on the web.

She would like to use archived crisis maps for her research in collaboration with a computer scientist.

source: Editor abcdef
Since 1928 the Springfield History Club is documenting the history of the town. They are aware of the rich content that is available on the web related to their town - web sites of the council, from companies, clubs, and citizens. The club would like to create an archive of the most important web pages related to Springfield.
Gulp

Gulp is one of the largest producers of soft drinks with a vast online presence, including several national Gulp websites, Twitter, Instagram, and Facebook streams, the Gulp blog, and websites for other Gulp-owned brands. The first Gulp website was published in 1995. Gulp would like to establish a web archive to capture and preserve Gulp websites and social media.
First Web Bank Ltd.

First Web Bank offers bank accounts and online banking to more than 50 million customers worldwide. According to new regulations, all online banking transactions need to be archived for non-repudiation. This includes user actions, entered data, as well as the content of web pages and account statements shown to the user.
Fast Media

As a regional news and media agency, Fast Media would like to archive web pages about regional events and their own web pages. They would like to include rich media as well as social media and integrate this into their online presence. The archive will be valuable for their own news coverage and as a service for their customers.
Molvanîa is a small country with less than 7 million citizens. The new legal deposit legislation requires that the Molvanîa National Library archives all web pages from and about Molvanîa. The goal of the library is to build a comprehensive archive which should also include web pages which ceased to exist.
Web Archiving

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.