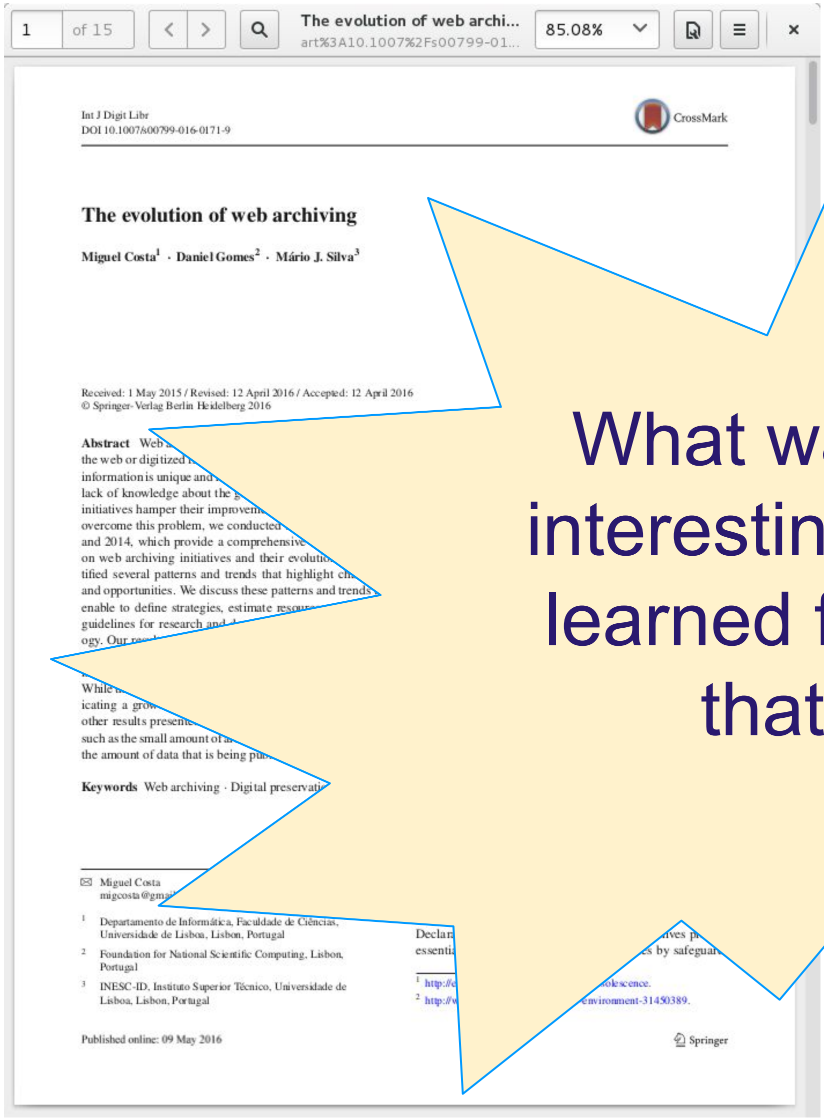
Web Archiving

The evolution of web archiving



The evolution of web archi...
art%3A10.1007%2Fs00799-01...

Int J Digit Libr
DOI 10.1007/s00799-016-0171-9

CrossMark

The evolution of web archiving

Miguel Costa[1] · Daniel Gomes[2] · Mário J. Silva[3]

Received: 1 May 2015 / Revised: 12 April 2016 / Accepted: 12 April 2016

Abstract Web
the web or digitized
information is unique and
lack of knowledge about the
initiatives hamper their improvem
overcome this problem, we conducted
and 2014, which provide a comprehensive
on web archiving initiatives and their evolutio
tified several patterns and trends that highlight ch
and opportunities. We discuss these patterns and trends
enable to define strategies, estimate resour
guidelines for research and d
ogy. Our re

While
icating a grow
other results presente
such as the small amount of
the amount of data that is being pu

Keywords Web archiving · Digital preservati

✉ Miguel Costa
migcosta@gma

[1] Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal

[2] Foundation for National Scientific Computing, Lisbon, Portugal

[3] INESC-ID, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal

Declar ives p
essenti es by safeguar

[1] http://e olescence.

[2] http://w environment-31450389.

Published online: 09 May 2016

Springer

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

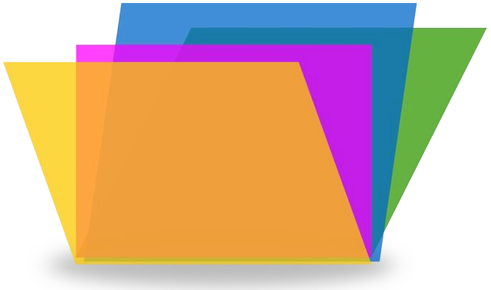
Agenda

1. Potential

source: Scott Maxwell

2. Approaches

3. Requirements



source: Scott Maxwell

Potential

The Potential of Web Archiving

1

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

1. Stirling, P., Chevallier, P. & Illien, G. (2012). **Web Archives for Researchers: Representations, Expectations and Potential Uses**. *D-Lib Magazine*, 18. DOI: 10.1045/march2012-stirling (6000 words)
2. SalahEldeen, H. M. & Nelson, M. L. (2012). **Losing My Revolution: How Many Resources Shared on Social Media Have Been Lost?** In P. Zaphiris, G. Buchanan, E. Rasmussen & F. Loizides (eds.), *Proceedings of the Second International Conference on Theory and Practice of Digital Libraries* (pp. 125–137), Berlin/Heidelberg: Springer. DOI: 10.1007/978-3-642-33290-6_14 (5000 words, feel free to skip the more technical parts)
3. Dougherty, M., Meyer, E. T., Madsen, C. M., Van den Heuvel, C., Thomas, A. & Wyatt, S. (2010). ***Researcher Engagement with Web Archives: State of the Art*** (Accepted Paper Series). Joint Information Systems Committee. URL: http://ssrn.com/paper=1714997 (**pages 9-17 only**, 6000 words)
4. Bhat, M. H. (2009). **Missing Web References — A Case Study of Five Scholarly Journals**. *LIBER Quarterly*, *19*(2), 131–139. DOI: 10.18352/lq.7957 (3000 words)
5. Costa, M. & Silva, M. J. (2010). **Understanding the information needs of web archive users**. *Proceedings of the 10th International Web Archiving Workshop*. URL: http://xldb.lasige.di.fc.ul.pt/xldb/publications/costa2010understandingneeds_document.pdf (6000 words)
6. Dougherty, M. & Meyer, E. T. (2014). **Community, tools, and practices in web archiving: The state-of-the-art in relation to social science and humanities research needs**. *Journal of the Association for Information Science and Technology*, 65, 2195–2209. DOI: 10.1002/asi.23099 (12000 words, split section "Findings" among group members)

The Potential of Web Archiving

2

or

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?

source: Paraschivu.Florin

Approaches

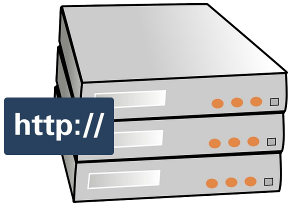
Overview

Stakeholders

Services

hanzo

Approaches



Software

INTERNET ARCHIVE

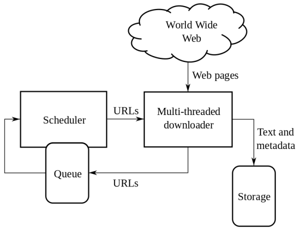
WayBackMachine

Selection Policies

.uk

Selection Criteria

Web Crawling

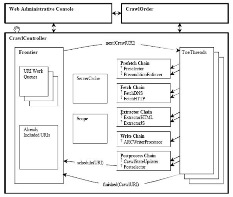


Standards

ISO

International Organization for Standardization

Heritrix



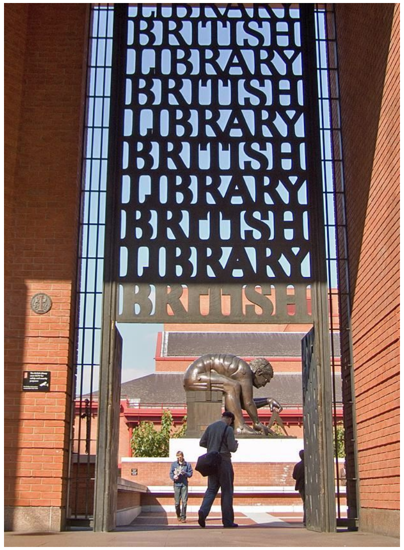
Stakeholders

source: Hugaholic

source: Ryan Wick

source: Holger.Ellgaard

source: Tilemahos Efthimiadis



source: Chris Beckett

IIPC
netpreserve.org

INTERNATIONAL
INTERNET
PRESERVATION
CONSORTIUM

and many more ...

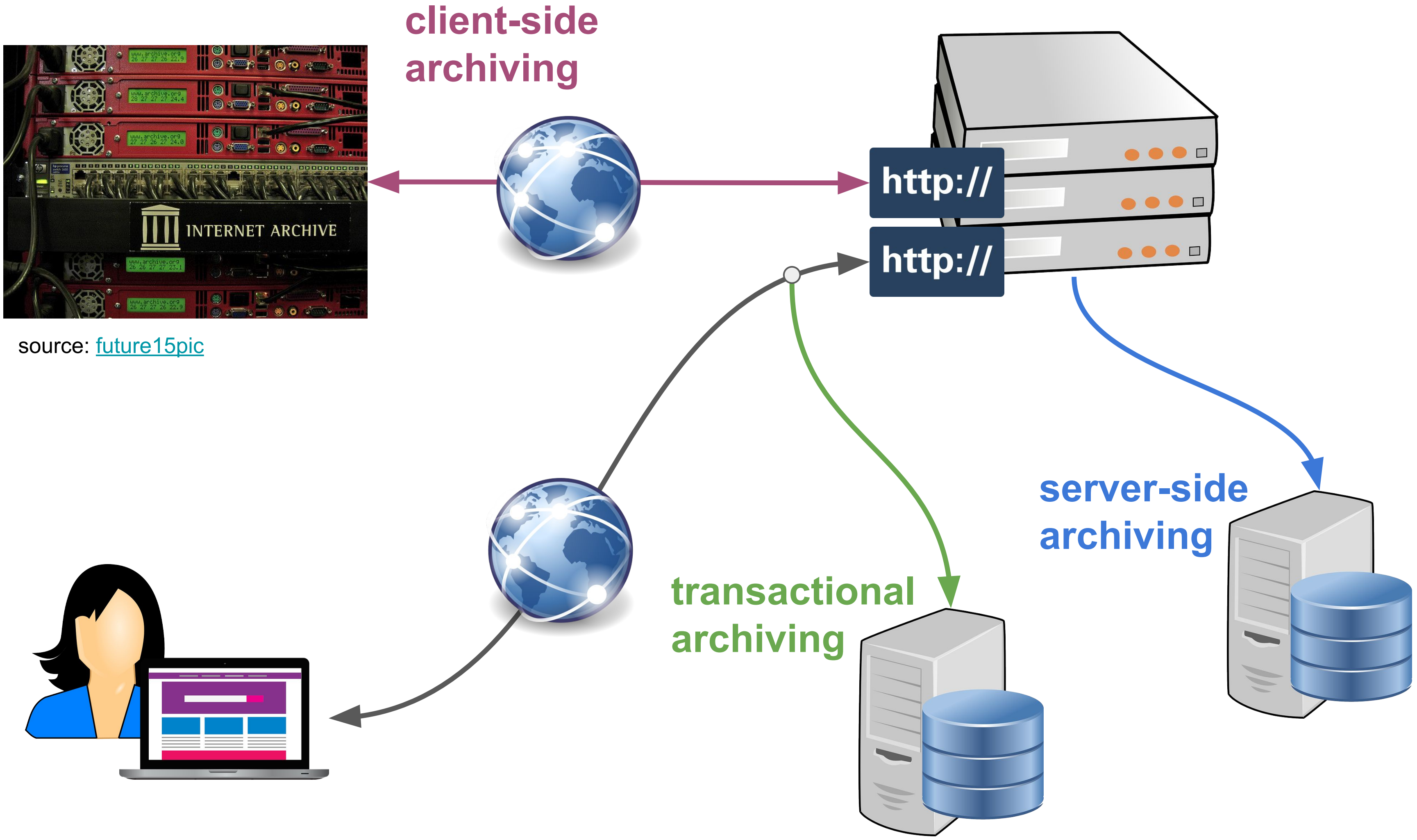
Services

Archivethe.net
rec

PageFreezer

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

Approaches



client-side archiving

http://

http://

transactional archiving

server-side archiving

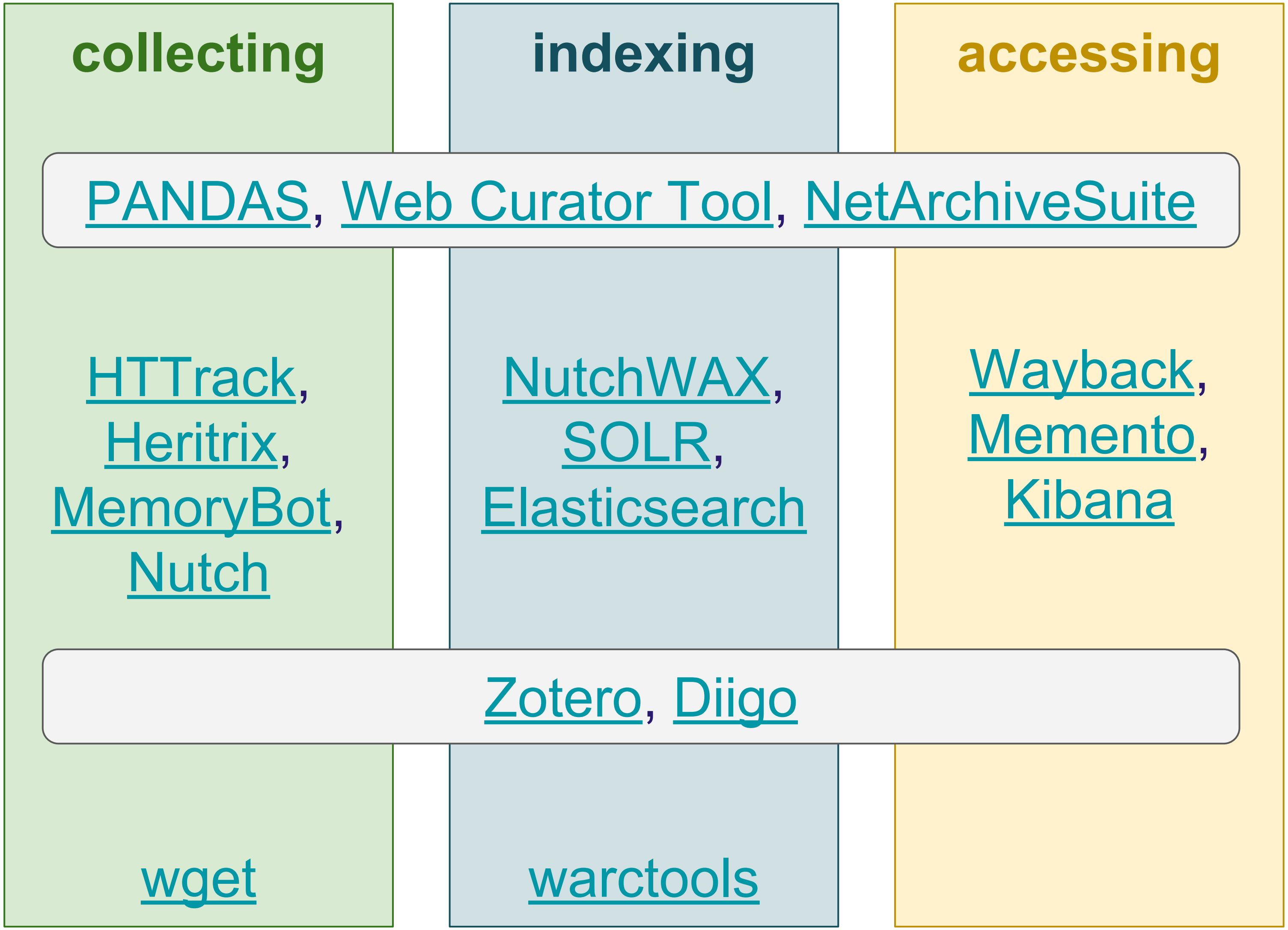
source: future15pic

Software

1

2

3



collecting

indexing

accessing

PANDAS, Web Curator Tool, NetArchiveSuite

HTTrack, Heritrix, MemoryBot, Nutch

NutchWAX, SOLR, Elasticsearch

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, …

+ potentially most comprehensive
- yet, often incomplete
+ relationships/links with sites

selective collections

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

+ 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”

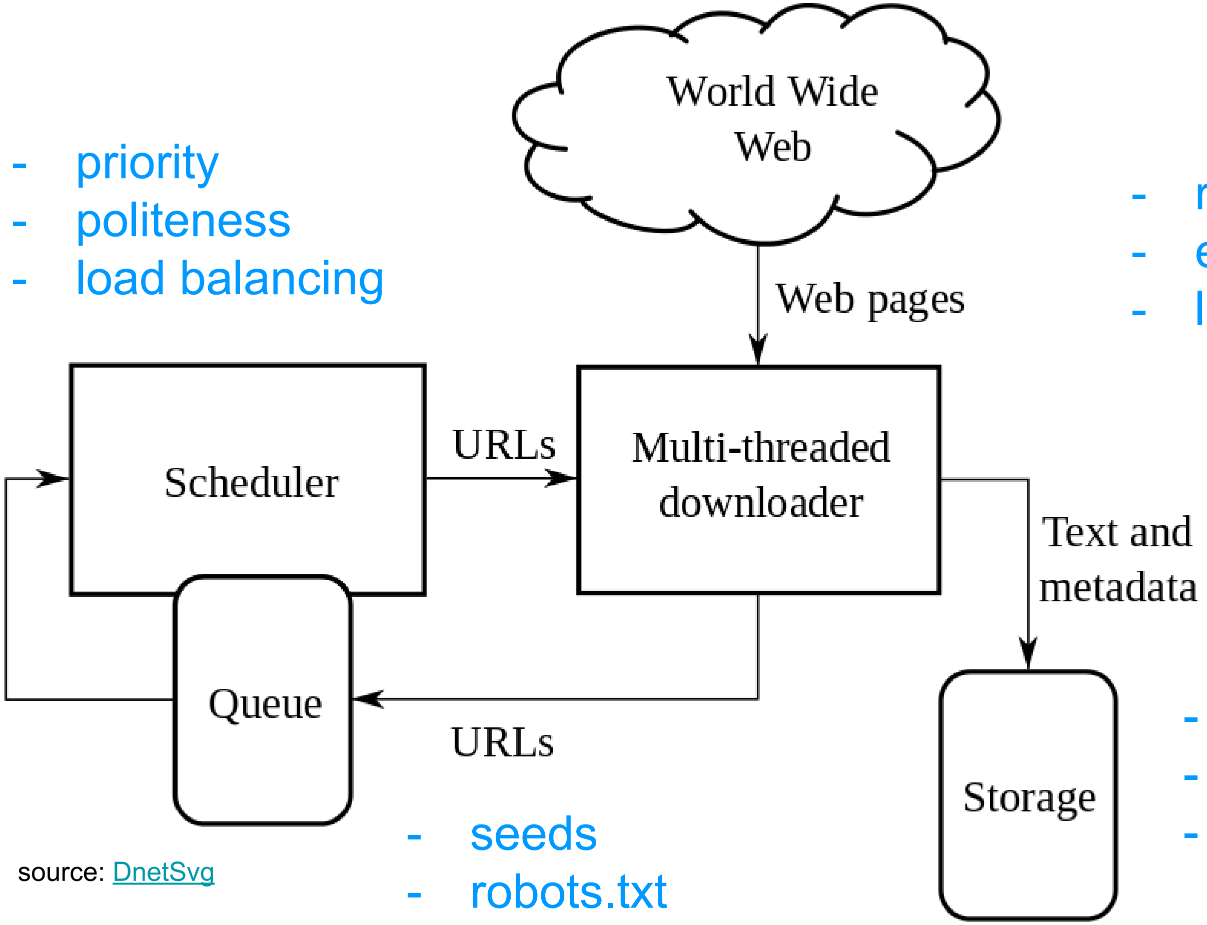
aspects:

- 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: DnetSvg

- seeds
- robots.txt
- selection criteria

- WARC
- indexing
- long-term preservation

Standards

W3C® HTML CSS JS

content

URI http:// `robots.txt`

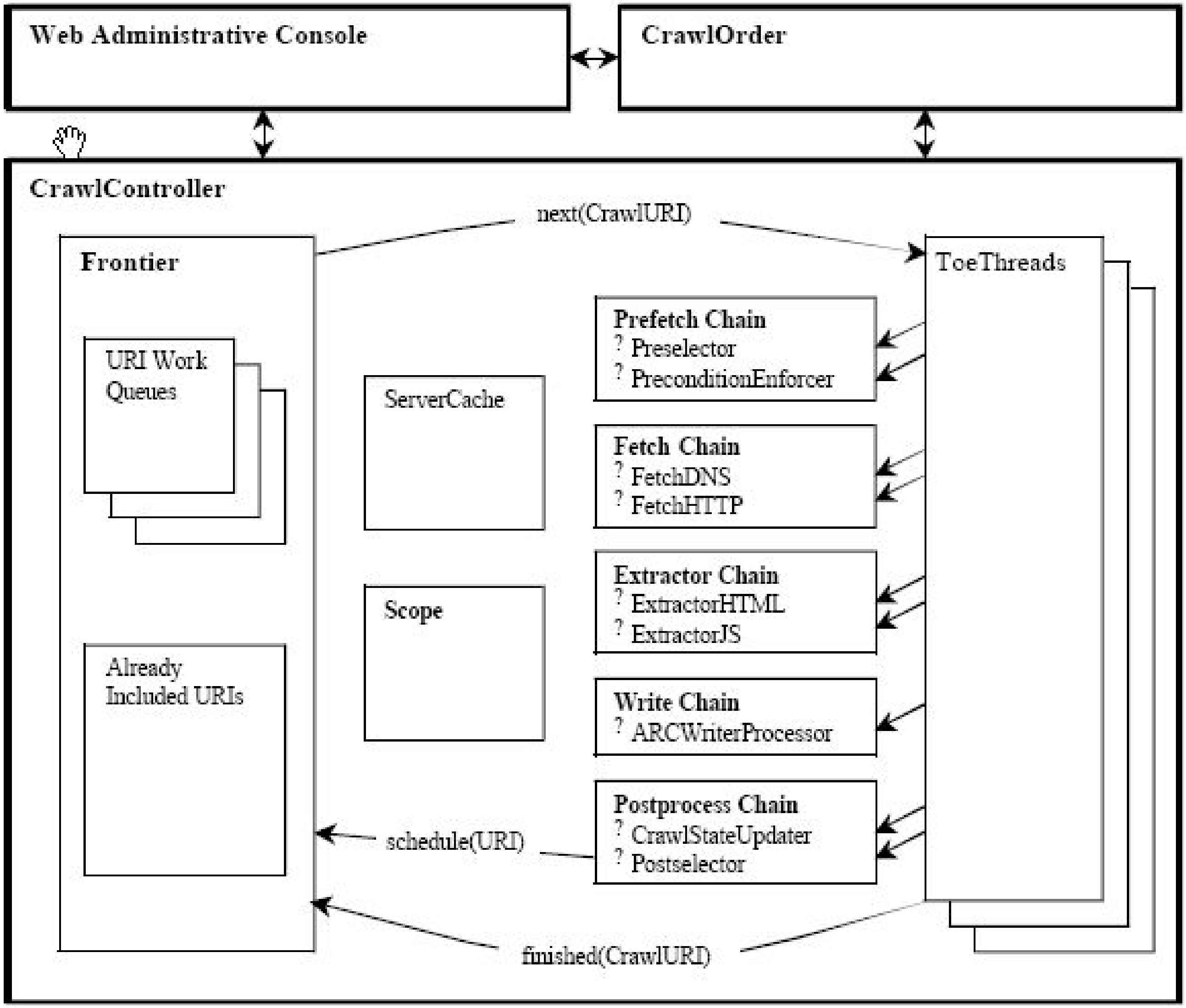
access

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

archive

sources: W3C, daPhyre

Heritrix



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

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

2

discuss your questions with the consultant

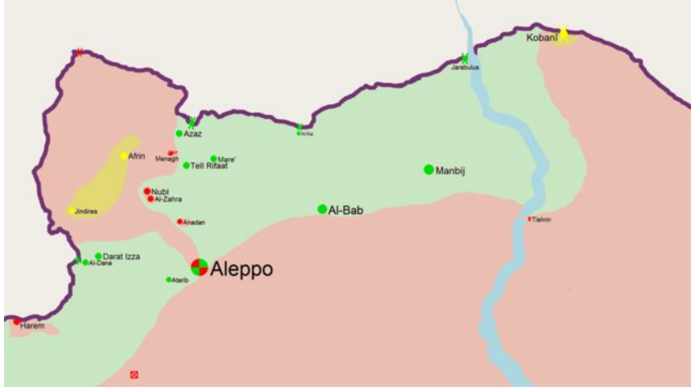
3

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

Springfield History Club

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 world wide. 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 National Library

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