


Existing Archival Practice:

Web Archives as **Infrastructure**

Ed Summers <edsu@stanford.edu>



Web Archives and Archival Studies

A bit about me and my research interests:

- **Appraisal**: how are decisions made about what to archive from the web.
- Appraisal is driven by beliefs about **use, misuse, disuse** (Ahmed, 2019)
- Appraisal is **sociotechnical**: delimited by tools, which embody needs.

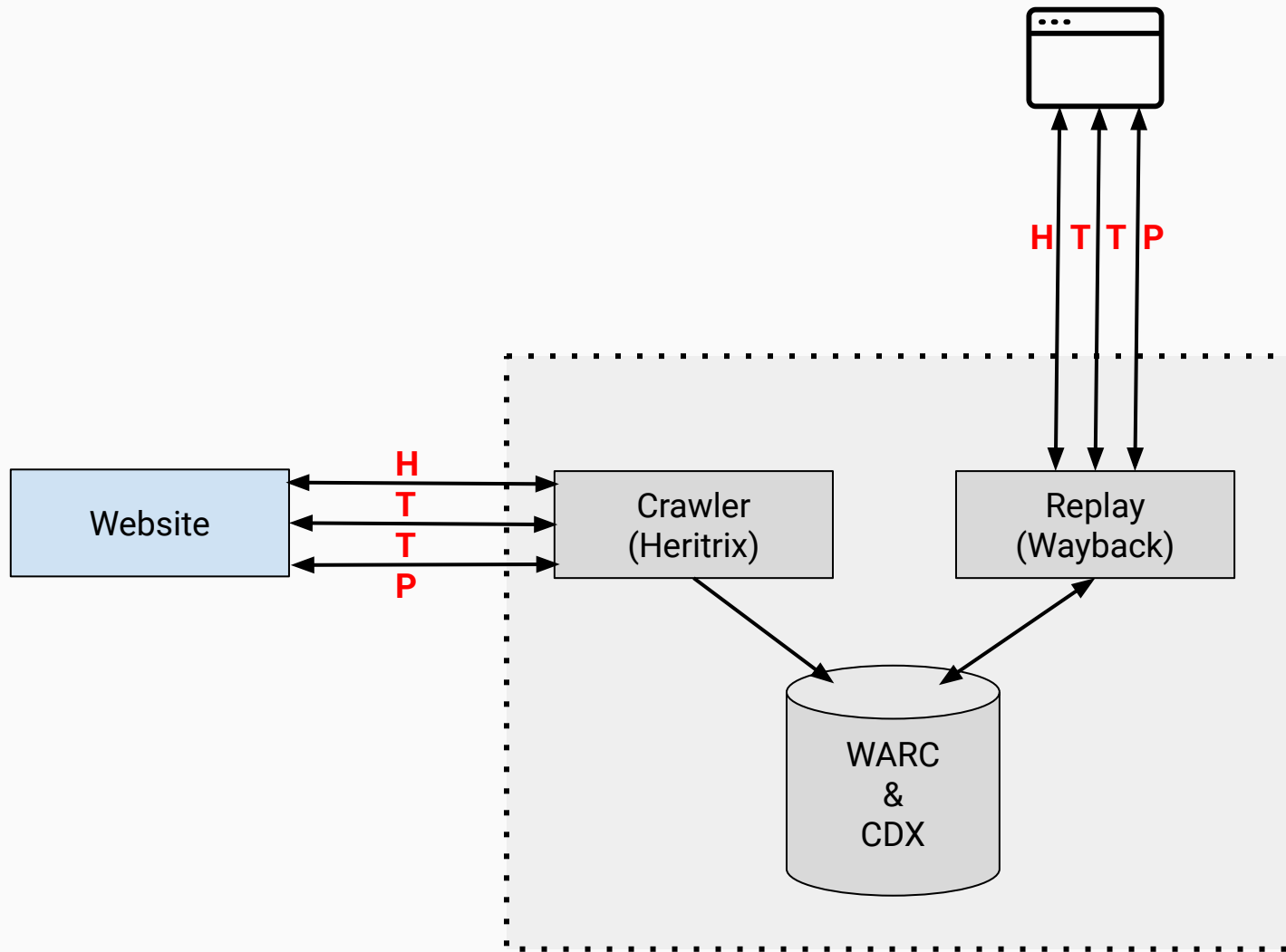
Putting these ideas into **practice** at Stanford University Library, helping to maintain the Stanford Web Archiving Program, while also contributing to the Webrecorder project (whose tools we use).

Web Archives are **Archives**

The fundamental concepts of archival theory are rooted in concepts embedded in Roman **law**, which have lingered for centuries and are so ingrained in our Western culture that we keep perpetuating them even when we cannot remember the reason for doing so. (Duranti, 1994)

A **document** is evidence in support of a fact. (Briet, 1951)

The existence, preservation and availability of archives, documents, records in our society are very much determined by the distribution of wealth and power. That is, the **most powerful**, the richest elements in society have the greatest capacity to find documents, preserve them, and decide what is or is not available to the public. This means government, business and the military are dominant. (Zinn, 1970)



Use Cases for Decentralized Web Archives

04 December 2022

▼ More details about this document

This version:
<https://specs.webrecorder.net/use-cases/0.1.0/>

Latest published version:
<https://specs.webrecorder.net/use-cases/latest/>

Editors:
[Ilya Kreymer \(Webrecorder\)](#)
[Ed Summers \(Stanford University\)](#)

Authors:
[Lorena Ramirez-Lopez \(Webrecorder\)](#)
[Cade Diehm \(New Design Congress\)](#)
[Benjamin Royer \(New Design Congress\)](#)
[Nicholas Taylor \(Los Alamos National Laboratory\)](#)
[Sua Yoo](#)

Additional Documents
[Specification](#)

Repository
[Github](#)
[Issues](#)
[Commits](#)

Copyright © 2022 Webrecorder CC-BY

Abstract

This document sets out use cases and requirements for a new packaging format for distributed web archive collections.

This document was published by the [Webrecorder Project](#) as part of a [grant from the FileCoin Foundation](#). It is being actively developed and feedback on new use cases is encouraged using [GitHub Issues](#).

TABLE OF CONTENTS

- Abstract**
- 1. Introduction**
 - 1.1 Motivation
 - 1.2 Existing Work
 - 1.3 Concepts of Distributed Web Archives
- 2. Use Cases**
 - 2.1 Researcher Saves an Article
 - 2.2 Social media user preserves and deletes account
 - 2.3 Curator preserves application in an institutional repository
 - 2.4 Preserving organizational social media accounts in institutional repository
 - 2.5 Resources for investigative journalism
 - 2.6 Curate art works for transfer to museum
 - 2.7 Aggregate access to routine automated crawls
 - 2.8 Author copies of work published elsewhere on the web
 - 2.9 Locally hosted copies of at risk video to guard against link rot
 - 2.10 Extract collection from larger archive for research
 - 2.11 Requesting materials during electronic discovery
 - 2.12 Manuscript publication
 - 2.13 Reconstructing local archive of lost website
 - 2.14 Extracting web pages from web archive for local full text search
 - 2.15 Court staff archive open web resources cited in opinions
 - 2.16 Offline first curriculum materials
 - 2.17 Evaluate trust in the creator of a web archive
- 3. Threats**
 - 3.1 Archived web content as evidentiary weapon
 - 3.2 Web archive for delivering trojan attack
 - 3.3 Surveillance and disinformation campaigns

Trust

- Centralized web archives have trust characteristics associated with them.
- Centralized web archives are single points of failure.
- Decentralized archives present challenges for knowing what and whose archived content to trust.
- Are there hybrid solutions?
- Trust by Who?

Transparency

- Centralized archives often hide (intentionally or not) their inner mechanisms for collecting and presenting archived content
- Decentralized technologies (blockchain and cryptocurrency protocols) can obfuscate power dynamics that are actually at play.
- Transparent for Who?

Portability

- Centralized archives bind code and data together in processes of crawling, indexing and replay which make it difficult to move archived content from one place to another.
- For web archives to be used as evidence they need to be storable and viewable in different (sometimes secure) contexts.
- Portable for Who?

Ownership

- Centralized web archives aggregate and control data **by design**.
- Archives that are used as evidence need to be controllable by individuals and organizations.
- The creators and providers of web archives need to be verifiable.
- Owned by Who?

Thanks!

edsu@stanford.edu