WARCreate

Create Wayback-Consumable WARC Files from Any Webpage

Mat Kelly and Michele C. Weigle

mkelly@cs.odu.edu, mweigle@cs.odu.edu

Contribution: Enable WARC generation from any webpage

Preserving content behind authentication on the web is difficult. Getting this content into the WARC format[2] is laborious and sometimes not possible. We developed a Google Chrome extension to allow a user to preserve the content of any webpage into the WARC format.

Problems Addressed By Tool

No Context

- Crawler does not have account on social media sites[1]

Not WYSIWYG

- A remote crawler may not see the same content as a user (e.g. Craigslist)
- Many webpages are rendered differently for different user agents (e.g. social media)

Unsuitable for Personal Archiving

- Ajax driven (e.g. social) sites could cause content to vary vs. original perspective

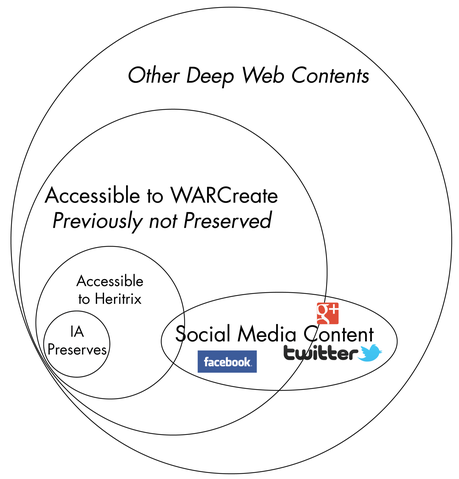
No Authentication

- Support for authentication by crawlers is limited[1]

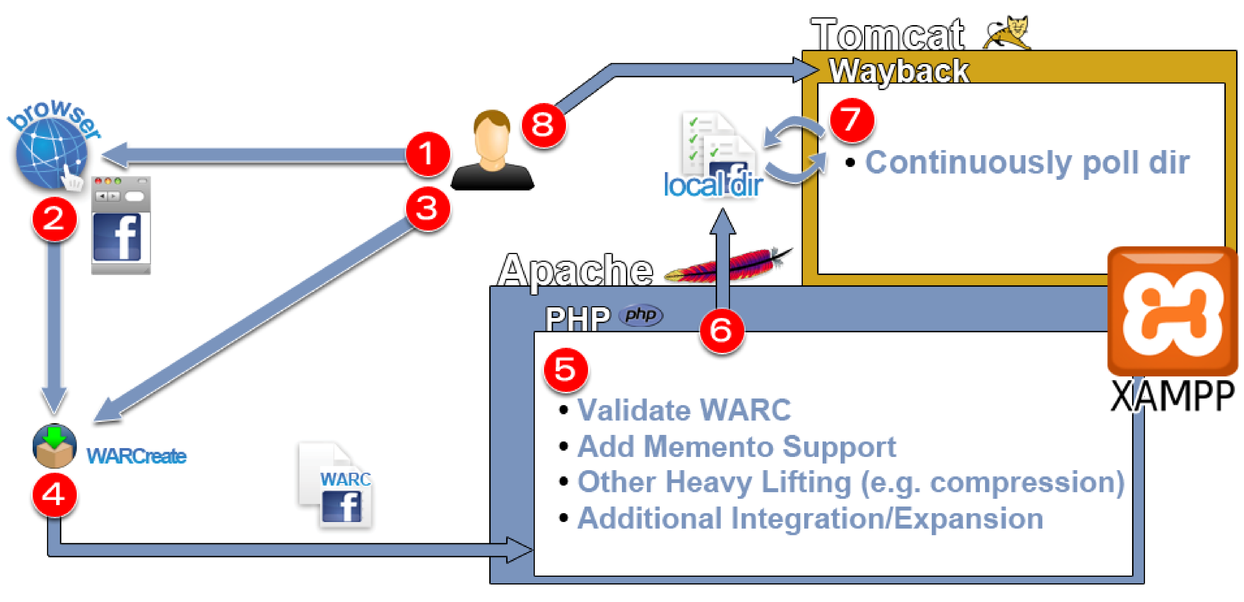
Impact

- Provides access to webpages that are inaccessible to crawler
- Allows user manipulation of webpages prior to archiving to ensure desired state
- Provides complete personal web archiving solution through the use of complementary open source technologies
- ***Content that was previously lost can now be preserved***

Preserve More!



How It Works and Fits Into a Personal Web Archiving Suite



1. User visits webpage via web browser
2. WARCreate captures HTTP Headers
3. User presses “Create WARC” button in extension
4. WARCreate packages WARC and sends it to (XAMPP-based) “Server”
5. Server script ensures WARC is fit for consumption and potentially extends features
6. Script sends OK’d WARC back to be saved to local directory
7. Local Wayback (via XAMPP) polls this directory
8. User views archived content in local Wayback

References

[1] M. K. Bergman The deep web: Surfacing the hidden value. *Journal of Electronic Publishing*, 2000.

[2] ISO. Information and documentation - WARC file format, 2009

[3] C. C. Marshall Rethinking Personal Digital Archiving, Part 1: Four Challenges from the Field. *D-Lib Magazine*, 14(3/4):2, 2008.

[4] C. C. Marshall, S. A. Bly, and F. Brun-Cottan The Long Term Fate of Our Digital Belongings: Toward a Service Model for Personal Archive In *Proceedings of IS& T Archiving*, pages 25-30, 2006.

[5] C. C. Marshall, F. McCown, and M. L. Nelson Evaluating Personal Archiving Strategies For Internet-Based Information In *Proceedings of IS& T Archiving*, pages 151-156, 2007.

Future Work

- Implement a completely decentralized personal web archiving process
- Port extension to other browsers
- Overcome bleed over (e.g. two users sharing facebook.com, different content)
- Address security concerns