Archival Tools to Match the Web: Open, International, Comprehensive

Gordon Mohr

Internet Archive, 4 Funston Ave, San Francisco, CA, 94129, USA
gojomo@archive.org

Abstract. Together with a number of national libraries, the Internet Archive committed itself in 2003 to international collaboration to create open source tools and standardized formats for web archiving. This project was motivated by our experience as home to over 100 billion archived web resources dating back to 1996, and as a partner to memory institutions building thematic web archives. Resulting tools include the *Heritrix* archival web crawler/harvester, the *Wayback* archive browsing service, and the *NutchWAX* archive full-text index and query utilities. A standard ingest/archival format for web resources called *WARC* has also been developed. Software with full source code is free to download and reuse, and organizations worldwide have adopted and contributed to these tools. Working with large collections remains a challenge, and the web itself is constantly growing and changing, so we continue to seek international cooperation to expand and improve this web archive tool set.

Keywords: World wide web, internet, harvesting, crawling, archives, indexing, search, HTTP, open source, collaboration.

1 Introduction and Background

Starting in 2003, the Internet Archive began developing open source tools for web archiving, with the support and assistance of many national libraries.

The Internet Archive is a non-profit Internet library based in San Francisco, California, USA. We are known for our 'Wayback Machine', offering public access to a web site archive of over 100 billion captured URLs dating back to 1996. We also have a leading role in the Open Content Alliance mass book digitization effort, and host popular free audio and video content collections, including thousands of live music shows and educational presentations.

The bulk of our web collection has come from raw content donations by a commercial partner, Alexa Internet. However, in 2003, the Internet Archive, together with the national libraries which would go on to form the International Internet Preservation Consortium (IIPC), determined there was a need for new tools and standards, built in an open collaborative model, for web archiving. Prior tools lacked the flexibility, archival focus, and unencumbered licensing possible with an open source approach.

D.H.-L. Goh et al. (Eds.): ICADL 2007, LNCS 4822, pp. 7–8, 2007.

2 Tools

Development of these tools began with a crawler, *Heritrix*, for harvesting web content. They have grown to also include a standard format, *WARC*, for storage and interchange of web content; a browsing service, *Wayback*, for viewing archived content; and search utilities, *NutchWAX*, for full-text indexing and querying of archived content using only free software. All are now available for free download and use, with full source code. Software is primarily implemented in cross-platform Java, and available for embedding and customization for other projects.

***Heritrix* archival crawler.** Heritrix was designed for faithful and complete content archiving, with a high level of configurability and customization. At the Internet Archive and elsewhere, Heritrix has been used for crawls of various frequencies – daily, weekly, monthly, quarterly, yearly, or one-time – and sizes – from a few thousand captured URLs to billions. Heritrix may be remote-controlled by other software or incrementally extended with new code modules.

***WARC* archival file format.** For over a decade, the Internet Archive stored captured web content in its own simple concatenated-responses format, called ARC. To better handle collection-time metadata, duplicate-reduction, format evolution, and other related storage needs, the Archive and other libraries designed a successor format, called WARC, now under consideration as an international ISO standard.

***Wayback* archive browsing service.** Wayback software allows URL-based lookup and browsing of archived web content, in a browser, as if viewing the original website near a desired time. Wayback has enabled access to multi-billion-URL collections.

***NutchWAX* full-text index and search utilities.** NutchWAX adds Google-style search to web archives, based on other open source projects. These include Lucene, a raw full-text search engine; Nutch, an adaptation of Lucene for web content; and Hadoop, a system for parallelizing large processing jobs. NutchWAX is being used in distributed configurations at the Internet Archive to provide search in collections containing over a billion URL captures.

3 Future Directions

The Internet Archive actively maintains each of these software projects, and improvements are often sponsored or contributed by our library partners and other software users. However, web archiving software continues to face serious challenges as the web grows in size, in diversity of content types and technologies, and in the level of adversarial content manipulation (web spam).

The collaborative, open source, international approach has worked to efficiently built a shared base of web archive capabilities, and we seek new collaborations to continue this progress. We hope the Asian library community will find these tools useful, report feedback from their experiences, and join us to build new functionality.