W3C Standard RDF Query Language
▷ SPARQL

W3C XML Path Language
▷ XPath/XQuery

W3C XML Query Language
▷ XPath/XQuery

Waiting Time Models
▷ Queueing Theory

Walk
▷ Semirings and Matrix Analysis of Networks

Web 2.0
▷ Topology of Online Social Networks

Web 2.0 Recommender Systems
▷ Recommender Systems, Semantic-Based

Web 3.0
▷ Twitris: A System for Collective Social Intelligence

Web Analysis
▷ Analysis and Mining of Tags, (Micro)Blogs, and Virtual Communities

Web Archives
Klaus Berberich
Max Planck Institute for Informatics,
Saarbrücken, Germany

Synonyms
Web archiving; Web preservation
Glossary

Crawler  A software that harvests content from the World Wide Web
URL  Uniform Resource Locator
Scope  The resources that a web archive seeks to preserve
Website  A collection of related URLs

Definition

Web archives are repositories of web contents collected in the past. They act against the ephemeral nature of the World Wide Web, where new contents are constantly added while others are removed and thus lost forever. Web archives counter this loss by preserving web contents as part of the cultural heritage for future generations. To this end, web archives select resources (e.g., specific websites) worth preserving, repeatedly acquire snapshots of these resources, store them together with metadata (e.g., a time stamp or keywords), and provide access to the archived web contents (e.g., via keyword search).

Institutions operating web archives include nonprofit organizations, universities, national libraries, and for-profit companies. Users of web archives include scholars (e.g., to study the evolution of the web), journalists (e.g., to retrieve documents no longer available on the live web), as well as companies (e.g., to comply with legal requirements regarding documentation).

Historical Background

The short-lived nature of contents on the World Wide Web was recognized soon after its arrival in the early 1990s. Brewster Kahle incorporated the Internet Archive (IA) (2013), whose mission he describes in Kahle (1997), in early 1996. Around the same time, several national libraries, including the Swedish Royal Library (Arvidson and Lettenström 1998), started preserving their national webs. The WebBase project at Stanford University (Stanford WebBase Project 2013) started in the late 1990s and is active to date.

In 2001, the International Web Archiving Workshop (IWAY) (2013) took place for the first time and became the prime forum of the web archiving community for the following decade. In 2003, the United Nations Educational, Scientific and Cultural Organization (UNESCO) (2003) acknowledged the need to preserve digital heritage, specifically mentioning web contents. In the same year, the International Internet Preservation Consortium (IIPC) (2013) was established by 12 institutions, including IA and several national libraries, as an international organization to foster and coordinate web preservation efforts. Key accomplishments of IIPC thus far include the now-standardized WARC file format (ISO 28500) and joint development of software such as the Heritrix crawler (2013). In 2004, the European Archive (now Internet Memory Foundation) (IMF) (Internet Memory Foundation 2013) was established as a counterpart to IA in Europe. More details on the history of web archives, which go beyond these key events, can be found in Toyoda and Kitsuregawa (2012).

Web Archives

Selection

One fundamental question that every institution operating a web archive has to answer is the question of its scope. It thus has to select those resources that it aims to preserve. At least five types of scope can be distinguished.

Broad scope refers to a nonselective policy that attempts to include all available resources in the web archive. While this seemed feasible in the early days of the web and was therefore initially adopted by IA, today’s size and growth of the web make it impractical.

Site-centric scopes define one or multiple websites (e.g., of a company or the government of a country) whose contents are to be preserved.

Topic-centric scopes aim to preserve all resources that are relevant to a specific topic, often defined based on a real-world event (e.g., an election, sports event, or natural disaster). As a concrete example of a topic-specific web archive collection, consider the US Elections Web
Archive provided by the Library of Congress (LOC) (2013) that preserves content related to candidates, parties, polls, etc. for all presidential elections since 2000.

**Domain-centric scope** refers to a selection of resources based on the Domain Name System (DNS). For instance, a web archive could aim to include all resources belonging to a specific national top-level domain (e.g., .fr, .de, or .ch).

**User-driven scopes** involve users of web archives to determine which resources are to be preserved. Implementations of such a selection policy can be found in on-demand archiving services like Archive-It (2013) or Archive The Net (2013) operated by IA and IMF, respectively.

Orthogonal to their scope, web archives have to select content types (or MIME types) that they aim to (or are able to) preserve. While this typically includes established content types (e.g., text/html, application/pdf, or image/png), others are excluded because of their limited acceptance or because their acquisition is technically difficult (e.g., for streaming-media content types).

Web archives also differ in how they interact with the original content providers (e.g., the company operating a website) in the selection process. When pursuing an opt-out approach, as done by IA, content is removed from the web archive upon explicit request by the original content provider. LOC, in contrast, for legal reasons, pursues an opt-in approach, archiving web contents only with the explicit consent of the original content provider.

**Acquisition**

There are different modes of content acquisition for web archives. Masanès (2006) distinguishes three of them based on where contents are captured.

**Server-side archiving** directly copies all files hosted by a web server on a regular basis. This mode assumes privileged access to the web server and its file system. It is only suitable for static web contents, corresponding one by one to files in the file system of the web server, and therefore of decreasing relevance in practice.

**Transactions archiving** steps between clients and web servers by listening and (selectively) recording their communication via the Hypertext Transfer Protocol (HTTP). While this mode can also acquire dynamic web contents, it only captures web contents that have actually been accessed by users. To implement it, web servers or other components such as proxies need to be instrumented to eavesdrop on HTTP communication. Legal accountability (i.e., being able to prove what contents were shown to a user) is often a motivation for using this mode in practice.

**Client-side archiving**, as the mode most common in practice, requires no privileged access to web servers but acquires web contents by means of interacting with web servers via HTTP much like a human user with a browser. Crawlers (Olston and Najork 2010), as the software used in client-side archiving, have their origin in web search engines. Initialized with a set of seed URLs (e.g., the index pages of all websites to be archived) as starting points, a crawler fetches their contents, stores them, extracts links to other resources, and adds them to a priority queue of resources that are yet to be fetched (the so-called frontier). This process is then repeated until all resources have been fetched or other stopping conditions are met. While conceptually simple, today’s crawlers are sophisticated distributed software systems that carefully allocate work across multiple machines and balance the load that they impose on web servers. Some key differences exist between crawlers employed by web search engines and those used by web archives. Web search engines typically only fetch resources that they plan to index and make searchable, whereas web archives have to acquire all resources that are needed for an authentic reconstruction of web contents later on (e.g., including style sheets and images). Moreover, as a challenge not faced by web search engines, web archives aim for temporal coherence (Denev et al. 2011), so that they can accurately reconstruct web contents as of specific times in the past. Crawlers in web archives thus have to fetch resources that belong together (e.g., a web page and its embedded images) in a timely manner to ensure temporal coherence of collected web contents.
At the same time, they have to respect politeness requirements of web servers (e.g., by issuing at most one request per second) specified via the robots exclusion protocol to avoid overloads. Heritrix (2013) is an open-source crawler worth mentioning that is specifically suited for web archiving and has been developed by IA and IIPC among others.

Crawlers can fetch web contents only if they are reachable via a sequence of links from one of the seed URLs. However, a non-negligible portion of today’s web contents is hidden behind forms and only generated on demand. To also harvest this so-called deep web, crawlers need to be extended (Madhavan et al. 2008; Masanès 2006) to automatically fill out forms with plausible input values in order to surface hidden web contents.

Once web contents have been fetched, web archives usually enrich them with metadata. In rare cases, this includes manually assigned metadata such as keywords or a summary describing the contents. More often, metadata is generated automatically and comprises information such as the time stamp of acquisition, the requesting/responding IP address, the content length, or the MIME type of the acquired contents.

Storage
Web archives have to ensure durability and fidelity when storing acquired web contents.

Durability is accomplished through the use of reliable storage technology. Nowadays, redundant storage on clusters of commodity computers, either in-house or in the cloud, is common practice. Before that, it was usual to store acquired web contents on disk arrays or even tape. As this shows, acquired web contents are not stored once and forever, but are likely to be migrated, as storage technology continues to evolve.

Fidelity is fostered by storing acquired web contents that are related (e.g., a web page and its embedded images) with its added metadata, about the resources or the acquisition process, in one place. The Web ARCHive format (WARC), developed by IIPC and an ISO standard since 2009, supports storing acquired contents of multiple web resources including metadata in one file. Changes in technology pose a serious threat to fidelity. It may thus happen that archived web contents become useless, since the technology (e.g., a browser plug-in or other software) required to view them is no longer available or does not run on current hardware. As countermeasures, preservation strategies have been devised that also archive required technology or convert all web contents into commonly used formats.

Access
To be more than vaults of the web’s past, web archives have to provide users with convenient access to their contents. Typically, for copyright reasons and in order not to compete with original content providers, there is a lag of several months until collected web contents are made accessible. For legal reasons, access to some web archives is still restricted to specific user groups. The French National Library (BNF) (2013), for instance, which has been operating a web archive since the early 2000s, provides access only to on-site visitors of the library. IA (Internet Archive 2013) and IMF (Internet Memory Foundation 2013), in contrast, provide open access to their contents via the web. Two types of access interfaces prevail in web archives.

URL-based interfaces, such as the Internet Archive Wayback Machine shown in Fig. 1, assume that users have an idea about the URLs whose past content is of interest. When prompted with a URL (e.g., http://www.un.org), these interfaces show the user an overview of available snapshots of the resource from the past. These then serve as entry points to the web archive, and the user can browse its contents by following hyperlinks much like when browsing the live web at the time when snapshots were taken.

Search-based interfaces, such as the one provided by the Portuguese Web Archive (2013) (PWA) shown in Fig. 2, allow users to search web archives by means of keyword queries (e.g., united nations) much like when using web search engines such as Google or Bing. Optionally, keyword queries can be extended by a time interval of interest (e.g., the year 1996).
Web Archives, Fig. 1  The Internet Archive Wayback Machine showing http://www.un.org as of January 4, 1997

Web Archives, Fig. 2  The Portuguese Web Archive showing results for the keyword query united nations from 1996
(Berberich et al. 2007) to indicate a period in the past from which snapshots are sought. Again, search results serve as entry points to the web archive. Both types of access interfaces need to deal with near-duplicate snapshots to improve user experience. This is done, for instance, by highlighting representative snapshots in URL-based interfaces and grouping near-duplicate results in search-based interfaces.

In addition to interfaces targeted at human users, several web archives including IA and PWA offer application programming interfaces (APIs) to access their contents as a building block in other software.

**Institutions**

Institutions with a stake in web archiving cover a wide spectrum and include:

- **Nonprofit organizations**, e.g.:
  - Internet Archive [http://www.archive.org](http://www.archive.org)
  - Internet Memory Foundation [http://www.internetmemory.org](http://www.internetmemory.org)
  - International Internet Preservation Consortium [http://www.netpreserve.org](http://www.netpreserve.org)
- **National libraries**, e.g.:
  - British Library [http://www.webarchive.org.uk](http://www.webarchive.org.uk)
- **Universities**, e.g.:
  - Columbia University Libraries [http://library.columbia.edu](http://library.columbia.edu)
  - Harvard Libraries [http://lib.harvard.edu](http://lib.harvard.edu)
- **Publicly funded research projects**, e.g.:
  - Living Web Archives [http://www.liwa-project.eu](http://www.liwa-project.eu)
  - Longitudinal Analytics in Web Archives [http://www.lawa-project.eu](http://www.lawa-project.eu)
- **For-profit companies**, e.g.:
  - Hanzo Archives [http://www.hanzoarchives.com](http://www.hanzoarchives.com)
  - Iterasi [http://www.iterasi.com](http://www.iterasi.com)
  - Smarsh [http://www.smarsh.com](http://www.smarsh.com)

More complete overviews of institutions involved in web archiving, including a comparison of their provided services, can be found in Gomes et al. (2011) and Niu (2012a).

**Key Applications**

Key applications of web archives already common practice today comprise the following: the preservation of web contents as part of cultural heritage, the archiving of company websites for reasons of legal compliance and accountability or for the sake of documentation, as well as the retrieval or reconstruction of individual web pages or entire websites that have – sometimes accidentally – been deleted and are no longer available on today’s live web.

Beyond these, web archives have potential applications in different branches of science. Web science, for instance, studies the web itself in terms of its properties and evolution, making web archives an ideal data source for it. Also, more traditional sciences such as sociology and linguistics can potentially profit from the data treasured in web archives, for instance, to analyze the spread of ideas in a society or to study the evolution of language. While these applications are already feasible today, for them to become common practice, web archives must become more widely known and improve in terms of accessibility. More details on established and potential applications of web archives can be found in Dougherty et al. (2010) and Thomas et al. (2010).

**Future Directions**

Web archives, like the web itself, are still a relatively new phenomenon. It will impact their role in the future, how they address the following challenges:

- Coping with the ongoing evolution of web technology towards more complex client-server interactions (e.g., AJAX/JavaScript), tailoring of web contents to post-PC era devices (e.g., tablets and smartphones), and increasing personalization of web contents
- Appearance of semantic-web data and linked open data, as a complement to traditional web contents, in formats such as RDF/S
- Encompassing online social networks, such as Twitter, Facebook, and Google+, that are
becoming substantial parts of the web but restrict access by crawlers

- Dealing with other contents that do not build on standard web technology such as smartphone applications or virtual worlds like Second Life
- Enhancing their accessibility by means of search functionality tailored to web archives and becoming more widely known

Possible scenarios for the future of web archives are delineated by Meyer et al. (2011).

Cross-References

- HTML
- Web Science

References


Recommended Reading

Masanès (2006) remains the key reference on web archives. While web technology has evolved since its publication, the majority of issues discussed therein are still current. More recent accounts on the state of the art in web archiving can be found in Niu (2012b) and Dougherty et al. (2010). Meyer et al. (2011), as a final recommendation, give a glimpse of web archives’ future.