What is web archiving?

Why is web archived?

Archived internet will serve as a basic source of information for future researchers. Vast amount of scientific and cultural information is nowadays published only in a digital form. Web content, meanwhile, is short-lived: it quickly changes, links rot, and information that was online yesterday is gone today. This is why various institutions interested in preserving data harvest and archive also internet content.

Web archiving technology

To harvest or scrap the content of internet pages, the Webarchiv of the National Library of the Czech Republic, like many other institutions, uses the Heritrix web crawler. Smooth and efficient harvesting, however, requires further extensions and scripts. The crawler browses the web, harvests content, and creates snapshots of pages at a particular point in time. It also creates an index, which is then uses to emulate archives pages in order to make them accessible.
Archived content is stored in ARC or WARC XML containers, which not only store web content but also supplement it with technical and administrative metadata.

<iframe src="//www.slideshare.net/slideshow/embed_code/key/AgMgbvBq0gR4Ks" width="595" height="485" frameborder="0" marginwidth="0" marginheight="0" scrolling="no" style="border:1px solid #CCC; border-width:1px; margin-bottom:5px; max-width: 100%;" allowfullscreen> </iframe>

What can be archived using the web archiving technology?

In principle, web archiving amounts to the downloading html and css files, images, pdf, doc and other objects, as well as audio and video files, eventually also javascript.
Web archiving technology enables the harvesting of only a fraction of internet. What remains inaccessible is a large part of the deep web, paid content or contents which require login, contents of databases, and problematic is also the harvesting of social networks or sites which contain streamed content. It is also impossible to harvest the content of digital libraries and similar applications.
In addition to technological limitations, web harvesting has also organisational and financial limits. Webarchiv of the National Library of the Czech Republic does not have unlimited resources, so that for instance a comprehensive harvest of Czech internet can only be carried out several times a year. Other limitations are user-defined, such as the number of links a crawler follows, maximum size and number of downloaded objects, and the like.

Comprehensive versus topic and selective harvests

Harvesting as such is implemented by either comprehensive or topic and selective harvests. Regular comprehensive harvest of an entire domain creates a snapshot of Czech internet at a particular moment. Topic harvests focus on documenting, for instance, the impact of a particular event on the internet information space. Additionally, there are some important sources which the Webarchiv of the National Library of the Czech Republic archives selectively, that is, over and above the regular comprehensive harvests.

What is the difference between web archiving and a local backup of a web page and a database?

The description of web archiving technologies listed above clearly indicates that web archive cannot replace the backing up of files which make up a web page, its CMS system, and its database. It does, however, make it possible to access an image of internet pages at a particular time even after they die and can no longer be accessed by the usual methods.

For webmasters

For webmasters, harvesting by the National Library crawler usually represents no risk. Robot of the Webarchiv of the National Library can be identified in access logs and its access to some particular web objects can be denied in robots.txt. You can inform the Webarchiv of the National Library of the Czech Republic about your interesting pages using the https://www.webarchiv.cz/en/add form.