



## Web archiving

The purpose of the Web Archive of the Koninklijke Bibliotheek/National Library of the Netherlands (KB) is to collect and preserve a selection of Dutch websites, thereby ensuring long-term access to the Dutch cultural heritage as it is published online.

### The Dutch web archive

The KB is responsible for collecting and preserving Dutch publications and for providing permanent access to them in the Netherlands Deposit Collection. In order to meet the challenge posed by the ephemerality of digital media, the KB developed the e-Depot, the KB's digital archive, which enables the KB to extend its traditional responsibility to digital objects, including websites.

The purpose of the KB Web Archive is to collect a selection of Dutch websites, emphasizing permanent storage and representation of archived websites. This means that websites are not only being harvested, but a strategy for long-term access is also being developed.

### Selection procedure

Websites are selected by the KB on the basis of its general collection policy, which focuses on Dutch history, culture and society. The archive should give an impression of the Dutch Web as we use it today. Since the Netherlands do not have legal deposit legislation, a procedure had to be developed to seek the owners' permission. The KB decided to use a practical opt-out approach. This is based on implicit permission to archive if the owner of a website does not declare otherwise. The first experiences with this approach are promising; very few website owners have raised any objections.

Although it was the absence of legal deposit legislation that first prompted the KB to choose the selective approach over a snapshot approach whereby the full Dutch web domain would be archived, it was also discovered that the Dutch web domain is far too large for a large-scale snapshot method: the .nl domain consists of over 2.5 million registered domain names, making the Netherlands the fourth largest 'country code Top Level Domain' in the world.

### Tools

The KB uses a common set of Web Capture tools, the IIPC toolset, consisting of a set of open source tools developed under the colours of the International Internet Preservation Consortium. The acquisition tool is the Heritrix crawler. The KB uses the WERA and Wayback Machine access tools and the NutchWax search engine. For managing the selection, harvesting and cataloguing, the Web Curator Tool is used. The archive will be accessible by means of a full text index and through the KB catalogue.

### Quality

Significant effort is invested in ensuring the authenticity and integrity of each archived domain. When storing a website into the archive, the policy is to maintain its 'look and feel', that is, its appearance and functionality, as well as its contents, to the fullest extent possible. This means that websites will be checked for completeness and functionality before being stored in the archive. During the ingest procedure, single files of websites are validated. Although it may not be possible to correct possible errors found in the files, it is important to store as much information about them as possible. The KB is currently working on a generic file validation procedure using JHove and DROID.

### Preservation

After crawling and quality control, websites are stored in the e-Depot and they become subject to a long-term preservation regime. As hardware and software platforms evolve and the need arises for preservation action, the KB will strive to maintain the 'look and feel', as well as the content. For technical and resource reasons this may not always be possible but the KB will aim for the ideal situation on a 'best efforts' basis.

The presentation of a website depends to a great extent on the browser being used, as well as the plug-ins needed for the presentation of specific aspects of a website (such as Flash, video and audio). Therefore, the KB is actively researching and developing techniques and methods that will enable the e-Depot to migrate or emulate digital objects in order that they can be viewed on current computer systems.

>>>

---

### Further information

Project Manager: Marcel Ras, Koninklijke Bibliotheek,  
+31 70 3140 180 [marcel.ras@kb.nl](mailto:marcel.ras@kb.nl)  
[www.kb.nl/e-depot](http://www.kb.nl/e-depot)

---



