Wayback usage in the Danish webarchive

The Danish Webarchive implementation of wayback has the following primary components.

**ArcRepository**

All access to the harvested data and metadata is via the ArcRepository interface from NetarchiveSuite. This ensures that we follow our own guidelines for bitpreservation and restricted access and also allows us to leverage the distributed ArcRepository architecture for the purpose of high-performance indexing.

**Indexing Component**

The Indexing Component consists of

1. The Wayback Indexer which generates raw index data using wayback-supplied and custom code packaged in NetarchiveSuite BatchJob "wrappers" which allow the code to be executed on the distributed ArcRepository.
2. The Aggregator which sorts and merges the raw index files. The actual sorting/indexing process is delegated to the native linux "sort" function which performs extremely well even on very large (> 100GB) files.
3. A database which simply records which files in the archive have already been indexed and which are awaiting indexing.

Currently all these components run on the same physical machine as the wayback tomcat server.

**Access Component**

The access component consists of a wayback installation under tomcat in Proxy Access Mode using a composite local CDX index (see wayback documentation for details). In addition, the installation includes the NetarchiveSuite wayback plugin which enables wayback to extract harvested data from the archive via NetarchiveResourceStore or NetarchiveCachingResourceStore.

Because of specific security requirements there are two further layers between the wayback server and the end user. These are

1. A kerberos proxy responsible for logging of all requests, and
2. An F5 FirePass browser-plugin based VPN solution, to ensure encrypted access when browsing from outside the institutional firewalls of the two home institutions.

**Limitations**

The wayback proxy is http only. It is essentially impossible to use such a proxy solution for browsing of https-harvested sites, nor does the solution work for ftp-harvested material.
For more information in general see Tools in the Wayback module and Wayback Configuration.

Users

Only a very limited number of researchers are currently using the Wayback access to the Danish webarchives. The Viewerproxy is used for Curator access to the Archive.