Quality Assurance

Quality Assurance is an essential activity on the curation of any digital collection. In NetarchiveSuite, the main tool for manual QA is the Viewerproxy. With Viewerproxy one can

1. Browse in harvested material from a single harvest job or from a single run of a harvest definition
2. Collect a list of URLs missed by the harvest, which can be added as seeds to a future harvest

In order to use Viewerproxy, your webbrowser needs to be set up to read data from the NetarchiveSuite archive, rather than from the living web. The details of the setup will depend on precisely how your installation of Netarchivesuite is configured. You will need to know i) the machine on which the Viewerproxy application is installed, and ii) the port number the Viewerproxy uses. Your web browser should then be set to use this machine and port as a proxy for all requests, except those to the machine where the NetarchiveSuite GUI is running. Most browsers have plugins which can help with managing proxy settings.

To use the Viewerproxy, select any Run number or JobID from the list of jobs or harvests and click on the line "Select these jobs for QA with viewerproxy". You should see a page which looks like

Generating index 'Harvest snapshot_wO5, run 1' for jobs [10, 11, 12, 13]

If you don’t see this then your proxy setup is not correct. After a while, this page should redirect to something like this:

Viewproxy Status

Current Viewproxy status:

*Currently does not collect missing URLs.*

Current list of missing URLs contains 0 URLs.

Using index 'Harvest snapshot_wO5, run 1', built on jobs: 10, 11, 12, 13.

If the frame is empty, either the viewerproxy hasn’t been started, or your web browser has not been configured to use it.

Missing URL collection

Start collecting URLs
Stop collecting URLs
Clear collected URLs
Show collected URLs

Browsing jobs in the viewerproxy

Use these pages to select the index for viewerproxy browsing:
Selective harvest history
Snapshot harvest history

Quality assurance is done by browsing the archive for selected domains. For example, simply open another tab and type a URL you expect to be present in the harvest you are studying. The page should be visible and navigable.

The various links under "Missing URL collection" can then be used to help you find material missed during harvesting.

Start collecting URL’s This starts the collection of URL’s. The Current Viewproxy status textbox shows if the system is currently collecting URL’s or not – and how many URL’s are currently collected.

Stop collecting URL’s Stops the collection of further URL’s.
Clear collected URL’s The list of URL’s can be cleared at any time e.g. when investigating a new domain starts. **NB! This function can not be undone.**

Show collected URL’s The list of URL’s can be viewed at any time. The list can be copied and added to relevant Seed lists for the relevant domains.