User Manual

This is a manual for end-user setup and control of harvests and controlling storage and QA. The audience for this manual will typically be curators.

Search manual

Download as pdf

Fundamental Concepts in Harvesting

The basic concept in the NetarchiveSuite harvesting module is the notion of domains.

A domain has a two part name host.top-level-domain|top-level-domain (e.g. netarchive.dk) or is an IP-number. What is considered a top-level domain is configurable. For most countries it makes sense that the top-level domain is simply the country code (like .dk or .fr), while for others it makes sense to go one level further down (like .co.uk). (The proliferation of new top-level domains is an additional challenge for anyone involved in "national" webarchiving.)

A domain can hold multiple harvest configurations. A configuration describes how to harvest the domain or a part of the domain. So one configuration could harvest the whole domain (used by the snapshot functionality) and other configurations could take different minor parts of the same domain (for selective or event harvests) - for example just the front page of a newspaper site.

A harvest configuration is defined by

1. A harvester template (a predefined crawler-bean template for the Heritrix web crawler)
2. A number of seedlists to use with that template - ie a list of which urls to use in initiating the crawl process

One of the harvest configurations defined for each domain is designated as the default configuration, and that configuration will be used when starting a snapshot harvest of all domains in the database.

Contents