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.

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).

A domain can hold multiple so called 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 the selective / event harvest).

A configuration consists basically of two things

- A harvester template (predefined templates for the Heritrix web crawler)
- A number of seedlists to use with that template
  A domain will always have a default configuration (selectable) and that configuration will be used when starting a snap shot harvest. The snap shot harvest therefore takes all domains known in the database.

Contents

Search manual

Download as pdf