netarchive.dk

Templates

Lars Clausen
Netarkivet

The role of templates

- Hides the complexity of Heritrix
- Allows added options without UI changes
- Part of domain configurations
- Modified on a per-job basis
- One job one template
- Input to the HarvestController

Typical templates

- With/without form posting
- With/without honouring robots.txt
- With special handling of site-specific URLs
- [more examples from BJA]

Modified parts of template

- Templates are order.xml, modified at
  - *Frontier queue-total-budget (max objects)*
  - *QuotaEnforcer group-max-all-kb (max bytes)*
  - *ExcludeFilter filters (crawler traps)*
  - Deduplicator index-location
  - Controller seedsfile
  - WriteProcessor path
  - WriteProcessor prefix

Assumed parts of templates

- QuotaEnforcer
- QueueAssignmentPolicy
  - DomainNameQueueAssignmentPolicy
  - Makes queues correspond to domains

Creating new templates

- Run Heritrix standalone
- Set up the options you want
- Grab the order.xml from the interface
- Upload it to NetarchiveSuite
- Templates are checked for required Xpaths
- No other checks yet – plans for more checks

Template Tool

- Command-line tool for templates
- .harvester.datamodel.HarvestTemplateApplication
- Arguments:
  - `create <template> <xml-file>`
  - `download <template>`
  - `update <template> <xml-file>`
  - `showall`
- Templates are checked for required Xpaths