Heritrix3 Configurations

Contents

For configuration related to NetarchiveSuite, please refer to section on Detailed Configurations#Configure Heritrix process.

For more specific Heritrix configurations, please refer to Appendix B2: Managing Heritrix 3 Crawler-Beans and Migrating H1 templates to H3 to use with NetarchiveSuite 5.1+

The crawling in NetarchiveSuite uses by default Deduplication.

How to configure which Heritrix report has to be uploaded in the metadata ARC/WARC file

Three settings properties control which heritrix reports are added to the metadata ARC or WARC file:

- `settings.harvester.harvesting.metadata.heritrixFilePattern` is a java pattern that allows you select which files in the crawl dir (not recursively) to include in the metadata ARC.
- `settings.harvester.harvesting.metadata.reportFilePattern` is also a java pattern that controls which subset of the files selected by heritrixFilePattern are to be considered as report files All the other files will be considered as setup files.
- `settings.harvester.harvesting.metadata.logFilePattern` is a third java pattern that controls which files in the logs subdirectory of the crawldir are to be added as log files to the metadata ARC.

It is possible to control the naming of the metadata file.

- `settings.harvester.harvesting.metadata.filename.versionnumber` controls a version-number which is used in the metadata file name, so the filen gets a name like 12345-metadata-4.warc.gz, where "4" is the version number
- `settings.harvester.harvesting.metadata.compression` is a boolean which determines whether the metadata files are generated in gzip format or not
- `settings.common.metadata.fileregexsuffix` determines which metadata files are recognised by the indexing tools, for both the Viewerproxy and Wayback, and for deduplication. The full regex string to be searched consists of the string <jobid>-<harvestid> followed by this suffix. The default value is `-metadata-[0-9]+.(w)?arc(.gz)?` which matches all metadata files using the standard NetarchiveSuite naming scheme.