Institutional Usage of NetarchiveSuite

At the KB-Denmark Netarkiv we are working on some quite radical changes to our backend architecture - replacing our ArcRepository storage with bitrepository.org software, and implementing a new mass-processing architecture probably based on hadoop. As part of this process we would like to know what parts of NAS are actually in use at our partner institutions so we can develop a strategy for future support.

NAS Applications

Which of the following NAS applications (services are in use in your production environment?

Application	Denmark	France	Austria	Spain	Sweden	Comments
HarvestControllerServer	y	y	y			
GUIWebServer	y	y	y			
HarvestJobManager	y	y	y			
ChecksumFileServer	y	n	n			
ViewerProxy	y	y	n			BnF: we only use the ViewerProxy to get access to warc files
WaybackIndexer	y	n	n			
AggregationWorker	y	n	n			
IndexServer	y	y	y			
ArcRepository	y	y	y			
BitarchiveServer	y	n	y			
BitarchiveMonitorServer	y	n	y			
AccessBitarchiveServer	y/n	n	n			This is a special read-only server which is used in a specific data-extraction system in DK, outside the main Netarkivet installation.

Plugins

Which of the following plugins are used in your production setup? Those marked with a ⭐ are default values set in the packaged settings file.

Interface	Implementation	Denmark	France	Austria	Spain	Sweden	
AbstractRemoteFile	HTTPRemoteFile			y			
	HTTPSRemoteFile						
	FTPRemoteFile ⭐	y	⭐				
ActiveBitPreservation	DatabaseBasedActiveBitPreservation			y			
	FileBasedActiveBitPreservation ⭐	y	⭐				
Admin	UpdateableAdminData						
	DatabaseAdmin ⭐	y	⭐	y			
arcrepositoryadmin.DBSpecifics	DerbyServerSpecifics ⭐						
	DerbyEmbeddedSpecifics						
	MySQLSpecifics						
	PostgreSQLSpecifics	y	y	y			
ChecksumArchive	FileChecksumArchive ⭐	y	⭐				
	DatabaseChecksumArchive						
JMSConnection	JMSConnectionSunMQ ⭐	y	y	y			
ArcRepositoryClient	JMSArcRepositoryClient	y					
	LocalArcRepositoryClient		y				
MonitorRegistryClient	PrintMonitorRegistryClient						
	JMSMonitorRegistryClient ⭐	y	y	y			
JobIndexCache	IndexRequestClient ⭐	y	y	y			
Notifications	EMailNotifications ⭐	y	y				

	PrintNotifications			y			
FreeSpaceProvider	DefaultFreeSpaceProvider ⭐	y	⭐	y			
	FreeSpaceProvider						
	OnbFreeSpaceProvider			y			
datamodel.DBSpecifics	DerbyServerSpecifics ⭐						
	DerbyEmbeddedSpecifics						
	MySQLSpecifics						
	PostgreSQLSpecifics	y	y	y			
JobGenerator	DefaultJobGenerator ⭐	y		y			
	FixedDomainConfigurationCountJobGenerator		y				
ArchiveFileNaming	LegacyNamingConvention ⭐	y		y			
	CollectionPrefixNamingConvention		y				
FrontierReportFilter	TopTotalEnqueuesFilter ⭐	y	y	y			
	ExhaustedQueuesFilter						
	MaxSizeFrontierReportExtract						
	RetiredQueuesFilter		y				
HeritrixLauncherAbstract	HeritrixLauncher ⭐	y	y	y			
IHeritrixController	HeritrixController ⭐	y	y	y			
HarvestReport	LegacyHarvestReport ⭐	y		y			
	BnFHarvestReport		y				
IndexRequestServerInterface	IndexRequestServer ⭐	y	y	y			

Command Line Tools

Over the years, the NetarchiveSuite codebase has accumulated *a lot* of command line utilities. Some of these were probably developed for a single specialised use-case or for test purposes, but others may have become part of the normal workflow at the various repositories. Here is a partial list of those that look most likely to be of general interest. Please mark any of those you know of that are used as part of your workflows.

Tool	Purpose	Denmark	France	Austria	Spain	Sweden	
DeployApplication	Creates deploy scripts from a deploy-config	y	y	y			
HarvestdatabaseUpdateApplication	Updates HarvestDB schema	y		y			
BuildCompleteSettings	Merges module settings files in NAS to one large global default settings file. Run as part of release process.	y					
GetFile	Retrieves a file via the ArcRepository interface	y					
GetRecord	Retrieves a (w)arc-record via the ArcRepository interface	y					
LoadDatabaseChecksumArchive	Migration tool from file-based checksums to database-based checksums		n(?)				
ReestablishAdminDatabase	For reestablishing the admin database from a 'admin.data' file						
RunBatch	Runs a batch job from the command line	y					
Upload	Uploads a file to the ArcRepository from the command line. (Handy for testdata.)	y		y			
ReestablishAdminDatabase	Should be deprecated ❓ Reads old admin.data file.						
ClassDependencies	Non NAS Utility (license is not ours)						
CreateIndex	CLI to talk to IndexServer via IndexClient						
RunChecksum	CLI to get all checksums from a Bitarchive (deprecated)		n(?)				
SendDedupIndexRequestToIndexserver	Asynchronously starts a dedup indexing on an IndexServer and then exits. Tue Hejlskov Larsen is this what you use to generate deduplication indexes?	i don't know ...					
MakeIndex	Runs a CDX extraction on a single file in a remote ArcRepository						
FindRelevantCrawllogLines	Finds crawl-log lines matching a given domain name in a local metadata file						
JMXProxy	"This tool will simply reregister all MBeans that matches the given query from the JMX hosts read in settings, using* its own platformmbeanserver. It will then wait forever."						
DeduplicateToCDXApplication	Extracts CDX records for deduplicate annotations from a local crawl log file						
ResetFailedFiles	Utility for WaybackIndexer to reset files that have failed more than 3 times so they can be retried		n(?)				

ARCReaderUtils	Splits an arcfile (not warc) and dumps results to a directory						
ArcWrap	Creates an arcfile by wrapping a file						
ExtractCDX	Extracts CDX records, unsorted, from a list of local input arcfiles (not warcs)						
JMSBroker	Checks that a JMS broker (as specified in NAS settings) is up and running.						
WriteBytesToFile	Just creates large files full of null bytes						
FTPValidator	Tests if an ftp server configuration in a NAS settings file points to a NAS-compliand ftp server.						
ArcMerge	Merges several arcfiles into one arcfile						
ArchiveExtractCDX	Extracts CDX records, unsorted, from a list of local input (w)arcfiles						
WARCExtractCDX	Extracts CDX records, unsorted, from a list of local input warcfiles						
ReformatTranslationFile	i) reorders a translation file so keys are in the same order as a reference file, and ii) allows the encoding of the output file to be changed						
MailValidator	Checks the validity of a mail-server configured in NAS settings by sending a test-mail						
MakeNewMetadataFile	Creates a metadata file. For use when postprocessing fails. Is this used?						
FindDomainsForCrawllogExtraction	?						
CheckDuplicateReduction	Validates deduplication by comparing a crawl log with a collection of arcfiles. (not warc)						
StandaloneApplicationReduced	Creates a standalone NetarchiveSuite in a single JVM						
MigrateDefaultHarvestDatabase	This just initialises a SiteSection object which is supposed to upgrade the harvest database as a side-effect						
CreateCDXMetadataFile	Complex tool that takes a set of filenames and runs a batch job to extracts the cdx'es from each files and pack them in a metadata arc or warc file, one record per input file						
HarvesterQueueControl	Tool to count the number of messages in a given JMS queue						
HarvestDatabaseValidator	Validates whether you can connect to the harvest database with the settings in a given settings file						
HarvestTemplateApplication	Utility for uploading and updating heritrix templates	y (in test)					
CheckDomainCrawltraps	Runs through all domains in the harvest database and checks whether each crawlertrap regexp can validly be included as text-content in an xml document	y					
CheckTrapsInFile	Runs through a list of crawler-trap regexes in a fileand checks whether each crawlertrap regex can validly be included as text-content in an xml documen	y(?)					