netarchive.dk

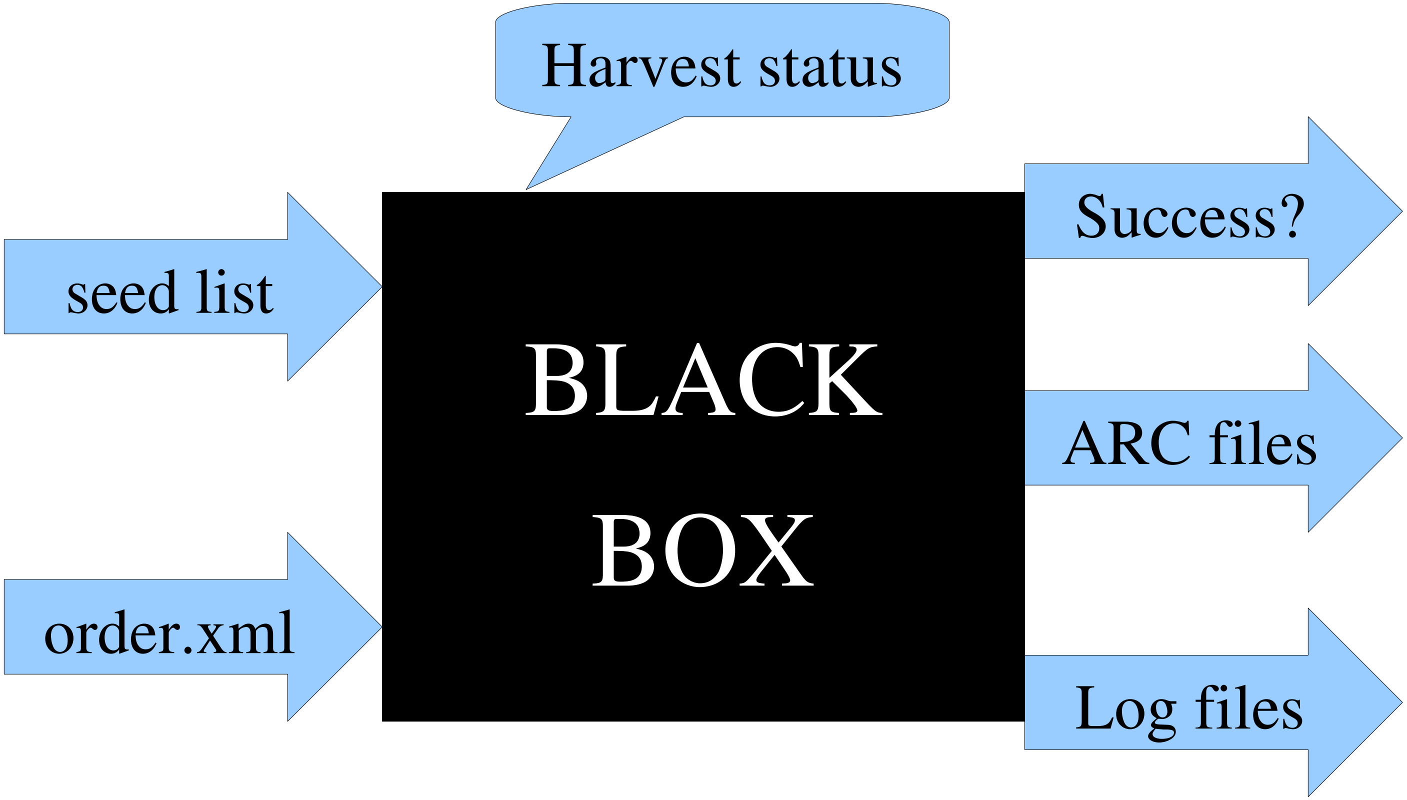
Heritrix Integration

How do we integrate with Heritrix

Outline

- Heritrix integration overview
- Input to Heritrix
- Output from Heritrix
- Starting/stopping Heritrix
- Uninvestigated scenarios

The NetarchiveSuite view of Heritrix



The generation of the seed list

- A concatenation of the seed lists of all domain configurations in the job

The generation of the order.xml

- order.xml template from domain configurations
- Updated with specific configuration options during scheduling (byte/object limits, crawler traps)
- Updated with deduplication info just before crawling
- Updated with file path info just before crawling

The handling of the output

- ARC files are index in a CDX file
- Log files and CDX files are packed up in a metadata ARC file
- All ARC files are uploaded to ARC repository
- Log files and success/failure is analyzed and the result sent back to the scheduler

The Black Box

- Current practice
  - Our HarvestControllerServer wraps an instance of Heritrix, using an instance of CrawlController to start and monitor server
  - A harvest is considered done if the harvest ends, throws an uncaught exception, or no activity is seen for a number of minutes (setting)

The Black Box

- Life cycle of HarvestControllerServer
  - Receive job on JMS
  - Wrap Heritrix crawl
  - Handle output
  - Suicide (Heritrix leaks memory)
- The application SideKick
  - Monitors the state of HarvestController
  - Restarts HarvestController after its suicide

The Black Box

- Upcoming practice
  - Instead of wrapping Heritrix CrawlController, start standalone Heritrix instance and monitor it with JMX
  - This allows the Heritrix UI to be up while harvesting
  - We do not need to restart HarvestControllerServer, only Heritrix

Uninvestigated scenarios

- Clustered crawls
- New built-in deduplication (we use Kristinn Sigurðssons)
- DecidingScope
- WARC
- What happens in 2.0?