NetarchiveSuite Overview

software to harvest, archive and browse large parts of the internet.

Printer friendly version

Contents

1. Introduction
2. The modules in the NetarchiveSuite
    1. The Common Module
    2. The Harvester Module
    3. The Archive Module
    4. The Access (Viewerproxy) Module
3. For developers

Introduction

The primary function of the NetarchiveSuite is to plan, schedule and archive web harvests of parts of the internet. We use Heritrix as our web-crawler. NetarchiveSuite was released on July 2007 as Open Source under the LGPL license and is used by the Danish organization Netarkivet.dk (http://netarkivet.dk). This organization has since July 2005 been using NetarchiveSuite to harvest Danish websites as authorized by the latest Danish Legal Deposit Act.

The NetarchiveSuite can organize three different kinds of harvests:

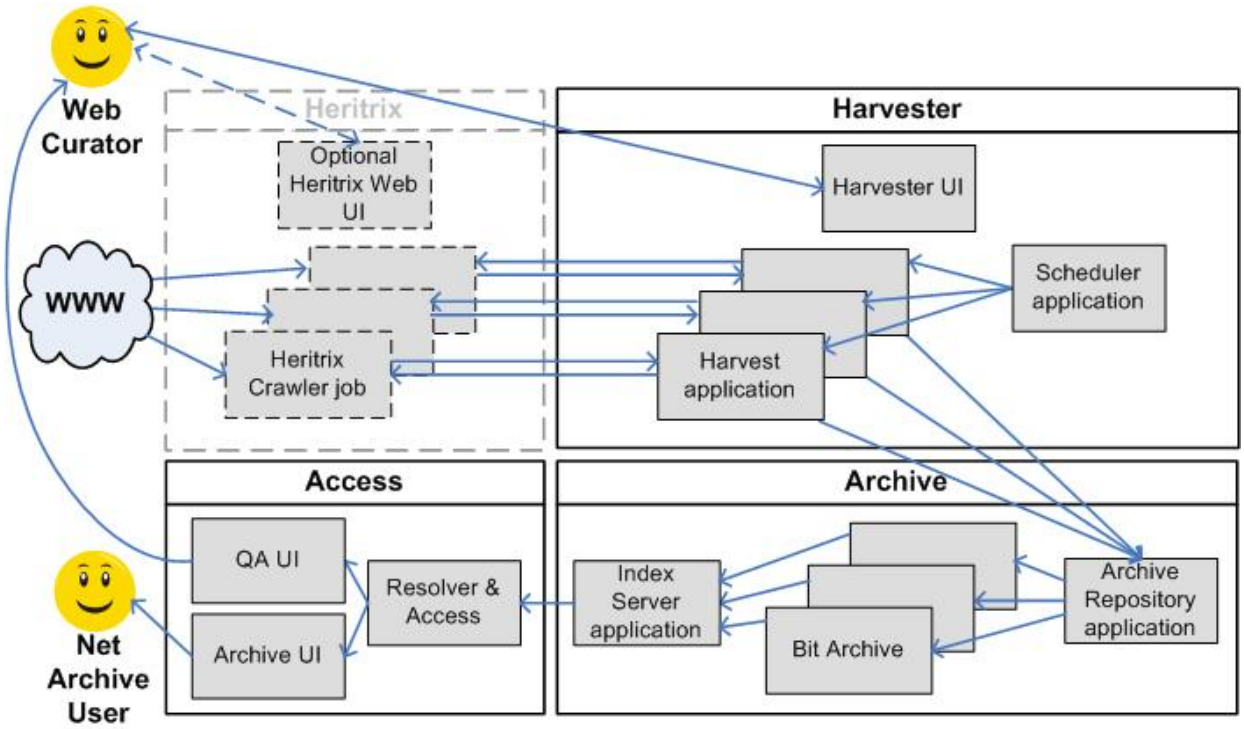
- Event harvesting (organize harvests of a set of domains related to a specific event e.g. 9/11, Elections and so on).
- Selective harvesting (recurrent harvests of a set of domains).
- Snapshot harvesting (organizing a complete snapshot of all known domains).

The software has been designed with the following in mind:

- Friendly to non-technicians - designed to be usable by librarians and curators with a minimum of technical supervision.
- Low maintenance - easy setup of automated harvests, automated bit-integrity checks, and simple curator tools.
- High bit-preservation security - replication and active integrity tests of large data contents.
- Loosely coupled - the suite consists of modules that can be used individually, or be used as one large web-archiving system.

The modules in the NetarchiveSuite

The NetarchiveSuite is split into four main modules: One module with common functionality and three modules corresponding to processes of harvesting, archiving and accessing, respectively.



The Common Module

The framework and utilities used by the whole suite, like exceptions, settings, messaging, file transfer (RemoteFile), and logging. It also defines the Java interfaces used to communicate between the different modules, to support alternative implementations.

The Harvester Module

This module handles defining, scheduling, and performing harvests.

- Harvesting uses the Heritrix crawler developed by Internet Archive. The harvesting module allows for flexible automated definitions of harvests. The system gives access to the full power of the Heritrix crawler, given adequate knowledge of the Heritrix crawler. NetarchiveSuite wraps the crawler in an easy-to-use interface that handles scheduling and configuring of the crawls, and distributes it to several crawling servers.
- The harvester module allows for de-duplication, using an index of URLs already crawled and stored in the archive to avoid storing duplicates more than once. This function uses the de-duplicator module from Kristinn Sigurdsson.
- The harvester module supports packaging metadata about the harvest together with the harvested data.

The Archive Module

This module makes it possible to setup and run a repository with replication, active bit consistency checks for bit-preservation, and support for distributed batch jobs on the archive.

- The archiving component offers a secure environment for storing your harvested material. It is designed for high preservation guarantees on bit preservation.
- It allows for replication of data on different locations, and distribution of content on several servers on each location. It supports different software and hardware platforms.
- The module allows for distributed batch jobs, running the same jobs on all servers at a location in parallel, and merging the results.
- An index of data in the archive allows fast access to the harvested materials.

The Access (Viewerproxy) Module

This module gives access to previously harvested material, through a proxy solution.

- The viewerproxy component supports transparent access to the harvested data, using a proxy solution, and an archive with an index over URLs stored in the archive.
- Support for browsing an entire crawl (like a snapshot or event harvest) or a single job (what one machine harvested).
- Allows for collecting unharvested URLs while browsing, for use in curation, and to include these URLs in the next crawl.

For developers

- The modules are loosely coupled, communicating through Java interfaces, with the implementation replaceable without recompiling.
- A rich number of settings in an XML structure allows for numerous ways of tweaking the applications for special needs.
- All design and code is peer-reviewed.
- There are Javadoc and implementation comments throughout the code.
- The code is tested with unit tests (coverage of 80%) and thorough release tests.
- Development happens in a well-defined development model (originally based on evolutionary prototyping).
- NetarchiveSuite is available under a well-known, integration-friendly, open source license (LGPL).

Overview 3.8 (last edited 2009-05-26 15:02:01 by KaareChristiansen)