Arch Quick Start Guide

1. Introduction

Arch 1.7 and later versions use own Jetty servlet engine and embedded RDBMS (H2) by default. This significantly simplifies deployment if you don't need advanced features of Arch. All you have to do to get Arch working is run the Ant build script and insert your seed URLs into Arch crawling script.

2. System requirements

- A Linux/Unix Operating System, or Windows Vista/7 with Cygwin installed;
- Java 7, or later version;
- Apache Ant and Ant-options packages. These are required for building Arch binaries.
- Apache Ivy;
- At least 2 GB RAM. The amount of memory required will depend on the size of the web site(s) and the number of web pages to crawl at one iteration. As a very rough guide 2 GB of RAM should be sufficient to crawl approximately 10,000 web pages at one iteration. There can be several iterations in each crawl.
- Free disk space for index data and temporary files. The space for temporary files is used in the indexing process and then released. Again, the amount required will depend on the size of the web sites. As an indication around 50 GB of temporary disk space should be sufficient to crawl about 10,000 pages at one iteration. The size of produced index is small compared to the amount of temporary space required.

3. Quick Start

Download Arch package from Arch home page:

http://www.atnf.csiro.au/computing/software/arch/

Switch to directory where you've downloaded the package (e.g. arch-1.7-src.tar.gz) and do the following:

```
#$> tar -xzf arch-1.7-src.tar.gz
#$> cd arch-1.7
#$> ant
#$> cd ArchHome/bin
#$> vi arch
#$> ./arch
```

When you type `vi arch` (see above), Vi editor will open Arch crawling script. Go through it and edit the parameters. As a minimum, you must provide seed URLs to start crawling with. It is also a good idea to do a trial crawl first, with `crawling.depth = 2`.

This is it!

It is a bit more complicated for Cygwin, see Section 4.6 of Arch Deployment Manual.

When Arch finishes crawling, you can search the index at this URL (unless you've changed the related configuration parameters):

http://<your-arch-host-address>:8993/arch/search

4. Using Arch Advanced Features

Starting Arch quickly as described here is sufficient for a trial, a demo, or for a relatively small set of sites with simple structure and up to a few thousand URLs to index. Using more advanced features of Arch requires further configuration. The table below describes essential Arch features and provides references to Arch Deployment Manual sections that contain relevant instructions.

Almost all features below require providing dedicated configuration directories for each web site. This is explained in Sections 4.1-4.4 of the Deployment Manual.

Feature	Benefits	What to do	References
Use statistics available in web servers logs **(Highly recommended)**	1. Dramatically increases user satisfaction with search by ranking most used (popular, useful) documents higher 2. Helps to find and index isolated pages that can't be found by "normal" crawling	Provide log files	Section 4.7
Differential crawling	Control how often particular areas in web sites are crawled	Split your sites into areas	Section 4.4
Limiting (filtering) search to a particular site area	Easier, more effective search	Split your sites into areas. Switch faceting on areas on.	Section 4.4, 5.10 and sample configuration files
Security	Make pages visible only to users and groups who allowed to see them	Configure security parameters	Sections 4.4, 5.2 and sample configuration files
Crawling password protected sites	Ability to index password protected content	Enable required Nutch plugin and provide authentication parameters	Section 5.7
Pruning documents, excluding from indexing common parts (such as menus) and advertising	Clean and effective index	Configure relevant parameters	Section 5.11 and sample configuration files
Watch mode	Instant indexing of new pages	Switch on watch mode	Section 5.13
PHP front-ends	Multiple search gateways, each with own filters and authentication	Deploy PHP front-ends	Section 6.1

Getting help

For help with installing Arch, please contact:

CSIRO Astronomy and Space Science

Arkadi Kosmynin

Phone +61 2 9372 4633

Fax +61 2 9372 4444

Email Arkadi.Kosmynin@csiro.au