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 or later with Cygwin installed;
- Java 9, 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:

Switch to directory where you've downloaded the package (e.g. arch-1.15-src.tar.gz) and do the following:

#$> tar -xzf arch-1.15-src.tar.gz
#$> cd arch-1.15
#$> ant
#$> 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/solr/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.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Benefits</th>
<th>What to do</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use statistics available in web servers logs <strong>(Highly recommended)</strong></td>
<td>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</td>
<td>Provide log files</td>
<td>Section 4.7</td>
</tr>
<tr>
<td>Differential crawling</td>
<td>Control how often particular areas in web sites are crawled</td>
<td>Split your sites into areas</td>
<td>Section 4.4</td>
</tr>
<tr>
<td>Limiting (filtering) search to a particular site area</td>
<td>Easier, more effective search</td>
<td>Split your sites into areas. Switch faceting on areas on.</td>
<td>Section 4.4, 5.10 and sample configuration files</td>
</tr>
<tr>
<td>Security</td>
<td>Make pages visible only to users and groups who allowed to see them</td>
<td>Configure security parameters</td>
<td>Sections 4.4, 5.2 and sample configuration files</td>
</tr>
<tr>
<td>Crawling password protected sites</td>
<td>Ability to index password protected content</td>
<td>Enable required Nutch plugin and provide authentication parameters</td>
<td>Section 5.7</td>
</tr>
<tr>
<td>Pruning documents, excluding from indexing common parts (such as menus) and advertising</td>
<td>Clean and effective index</td>
<td>Configure relevant parameters</td>
<td>Section 5.11 and sample configuration files</td>
</tr>
<tr>
<td>Watch mode</td>
<td>Instant indexing of new pages</td>
<td>Switch on watch mode</td>
<td>Section 5.13</td>
</tr>
<tr>
<td>PHP front-ends</td>
<td>Multiple search gateways, each with own filters and authentication</td>
<td>Deploy PHP frontends</td>
<td>Section 6.1</td>
</tr>
<tr>
<td>---------------</td>
<td>---------------------------------------------------------------------</td>
<td>---------------------</td>
<td>-------------</td>
</tr>
</tbody>
</table>

**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