Annotation of the Auburn Directory Plugin

What is a plugin?:

In the LOCKSS software, a plugin provides all the information needed to preserve a particular publisher's content. A plugin directs the daemon to crawl a defined set of pages with a shared URL structure. Plugins contain such information as how often the daemon will crawl a website, what URLs will be preserved, and how to divide a website into AUs.

Plugins are represented as XML files which contain key-value pairs. These pairs begin with <entry> and end with </entry>. A key is an assigned variable, e.g., "au_start_url" or "au_crawlrules" and a value is website specific information needed to direct a crawl.

Some Useful Terms:

AU = Archival Unit. A portion of content that we preserve as a unit. Usually a book or a volume or year of a journal.
AU Configuration Parameter = The information that defines AUs. The information varies between plugins, but typical examples are the base url and volume number. Configuration parameters are supplied by the title database.
Crawl = The automated process of collecting pages from a website.
LOCKSS Daemon = Our term for our central software that we commonly refer to as "the daemon." The daemon manages plugins, crawls sites, and stores and preserves collected information from crawls.
Key-value pair = A standard data representation used in computer languages. A key is a string which is the name of a variable. The value is its defined value, which does not have to be a string. It could be an integer, a decimal, or a complex object. It is whatever information is needed to direct a crawl. A key-value pair could just as easily be called a word-definition pair.

Example of a key-value pair:
```
<entry>
 <string>plugin_name</string>
 <string>Auburn Directory Plugin</string>
</entry>
```

Manifest page = A web page at a known location containing links for a crawl to follow. A manifest page is created by the publisher and must contain a permission statement for LOCKSS to crawl the AU. Every AU has its own manifest page which serves as a starting place for crawls.
Permission Page = The page which contains a defined permission statement, sometimes a creative commons license, which permits LOCKSS to crawl the website. The manifest page and the permission page are usually the same page.
printf = printf refers to "print formatted" and is a way to assemble a string from constant parts and variables. printf statements are used in the plugin to substitute AU config parameter values into strings.

Example of a printf string: "%s%s/manifest.html", base_url, directory.

Inside the quotation marks are one or more variables represented by *%s.* The variables' meanings are listed after the comma in the order in which they are replaced. The variables here are *base_url* and *directory*.
Regular Expression = A rule that matches patterns, strings of characters, or individual characters. One way LOCKSS uses regular expressions is in crawl rules to match URLs.
String = A common computer programming term meaning a sequence of characters. A string does not have to be inside <string></string> to be a string.
Title Database = A collection of AU descriptions. Information from the title database (e.g. AU configuration parameters) is copied onto the LOCKSS box when an AU is added to it.

COMMENTS are in black and are placed above the section they describe.
TRANSLATIONS are in blue and are placed under or next to the line they translate.

XML Plugin text is in red.

Note: This annotation contains line breaks in the XML code that do not exist in the actual plugin. This is to make the annotation and comments easier to read.

Here begins the Auburn Directory Plugin:

COMMENT: All key-value pairs are inside <map></map>
<map>

COMMENT: A unique name for the plugin. It usually corresponds to the file path where the plugin is stored. Here it is written as a Java class name and adheres to Java naming conventions.
<entry>
<string>plugin_identifier</string>
<string>edu.auburn.adpn.directory.AuburnDirectoryPlugin</string>
The identifier is the path to the XML file of edu/auburn/adpn/directory/AuburnDirectoryPlugin.xml
</entry>

COMMENT: The human readable name of the plugin.
<entry>
<string>plugin_name</string>
<string>Auburn Directory Plugin</string>
</entry>

COMMENT: Each time a plugin is edited and released, its version number is incremented by 1. It is important to keep track of plugin versions for best practices.
<entry>
<string>plugin_version</string>
<string>4</string> This plugin has been edited and released 4 times
</entry>

COMMENT: A printf string for a name of the AU that can be displayed in the daemon interface. AU name is used if no name has been supplied in the title database. AU name should contain enough information to uniquely identify the AU based on its AU configuration parameters.
<entry>
<string>au_name</string>
<string>”Auburn Directory Plugin, Base URL %s, Directory %s”, base_url, directory</string>
Auburn Directory Plugin, Base URL http://digdata1.lib.auburn.edu/, Directory AlabamaCommunityPlans
</entry>

COMMENT: The URL of the AU’s LOCKSS manifest page represented as a printf string.
<entry>
<string>au_start_url</string>
<string>“%s%s/manifest.html", base_url, directory</string>
http://baseurldirectory/manifest.html
If the base url is “http://digdata1.lib.auburn.edu/” and the directory is “AlabamaCommunityPlans” then the AU start URL is http://digdata1.lib.auburn.edu/AlabamaCommunityPlans/
</entry>

COMMENT: Crawl rules are specific, customized rules which tell the daemon what to collect and what to ignore. The crawl rules determine which URLs will be included in the AU. Crawl rules are written as printf strings that generate a regular expression.

A crawl rule consists of a regular expression and an action. The regular expression is matched against a URL to determine what action to take. There are four actions that can be taken. The number preceding the regular expression indicates the action. The possible actions are:

1 The URL will be included if it matches the regular expression.
2 The URL will not be included if it matches the regular expression.
3 The URL will be included if it does not match the regular expression.
4 The URL will be excluded if it does not match the regular expression.

The order of crawl rules is important. Each URL encountered during the crawl is tried in turn until a condition is satisfied. The URL matches or does not match the regular expression, at which point the URL is either included or excluded according to the action. If the URL is included, the page is collected.

```
<entry>
 <string>au_crawlrules</string>
  <list>
   <string>1,“^https?://%s/.*\.(bmp|css|gif|ico|jpe?g|js|png|tif?f)$”, base_url_host</string>
    Include the URL if it matches “http or https ://baseurlhost/anysequenceofcharacters. any of these files: bmp, css, gif, ico, jpg, jpeg,
    js, png, tif, tiff”
   <string>4,“^%s", base_url</string>
    Exclude URL if it does not begin with the baseurl
   <string>2,“^%s.*\?.*;O=[AD]$”, base_url</string>
    Exclude URL if it starts with the baseurl and has a query string that ends with with “;O=A” or “;O=D”. A query string is the part of the
    URL following the “?”
   <string>1,“^%s%s$”, base_url, directory</string>
    Include URL if it matches the baseurl  followed by the directory
   <string>1,“^%s%s/”, base_url, directory</string>
    Include URL if it starts with the baseurl followed by the directory followed by “/”
  </list>
</entry>
```

COMMENT: plugin_config_props defines the AU configuration parameters that are used to define AUs. The block that begins with org.lockss.daemon.ConfigParamDescr is repeated. Two au config parameters are defined here.

```
<entry>
 <string>plugin_config_props</string>
 <list>
  <org.lockss.daemon.ConfigParamDescr>
   <key>base_url</key> One parameter for the AU is a base URL
   <displayName>Base URL</displayName>
   <description>Usually of the form http://&lt;journal-name&gt;.com/</description>
   <type>3</type> Data type of configuration parameter. 3 indicates a URL
   <size>40</size> Indicates the size of the text field in an internally used interface
   <definitional>true</definitional> Indicates that the parameter is required to fully define an AU
   <defaultOnly>false</defaultOnly> Necessary for configuration structure, used internally
  </org.lockss.daemon.ConfigParamDescr>
  <org.lockss.daemon.ConfigParamDescr> 2nd ConfigParam block begins
   <key>directory</key> Another parameter for the AU is a directory name
   <displayName>Directory name</displayName>
   <type>1</type> Data type of configuration parameter. 1 indicates a string
   <size>20</size> A smaller text field is needed here
  <definitional>true</definitional>
  <defaultOnly>false</defaultOnly>
```

```
   </org.lockss.daemon.ConfigParamDescr>
  </list>
 </entry>
```

Comment: The amount of time the daemon pauses between successive page fetches. The value is given in milliseconds.

```
 <entry>
  <string>au_def_pause_time</string>
  <long>6000</long> The daemon pauses for 6000 milliseconds or 6 seconds
 </entry>
```

COMMENT: Determines which pages are checked for changes on re-crawls. The depth of a page is defined as the length of the shortest path of links from the manifest page to page X. (i.e. the smallest number of “clicks” necessary to navigate from the manifest page to page X). All pages with a depth up to the au_crawl_depth will be re-fetched to check for changes during re-crawls.

```
 <entry>
  <string>au_crawl_depth</string>
  <int>99</int> The depth and re-fetch depth of the crawl is 99 levels
 </entry>
```

COMMENT: The minimum amount of time before a re-crawl.

```
 <entry>
  <string>au_def_new_content_crawl</string>
  <long>5356800000</long> 5,356,800,000 milliseconds or 62 days
 </entry>
</map>
```

End of ADPN Plugin.