Kepler - Bug #4013

Make sure duplicate jars are not included in the installer

04/22/2009 03:29 PM - Chad Berkley

Status:	Resolved	**Start date:**	04/22/2009
Priority:	Immediate	**Due date:**	
Assignee:	Chad Berkley	**% Done:**	0%
Category:	installer	**Estimated time:**	0.00 hour
Target version:	2.0.0	**Spent time:**	0.00 hour
Bugzilla-Id:	4013		

Description

to the greatest extent possible, try to reduce the size of the installer by removing duplicate jars.

Related issues:

Blocked by Kepler - Bug #3949: Get the installer working with the new build s...	**Resolved**	**04/06/2009**

History

#1 - 05/13/2009 10:38 AM - Chad Berkley

Duplicate jar list from Christopher:

bash-3.2$ find . -name "*.jar" -print | awk -F / '{print $NF}' | sort | uniq -c | sort -nr
3 xalan.jar
3 jdom.jar
3 ij.jar
3 batik-all-1.6.jar
2 xml-apis.jar
2 xdoclet-1.2.2.jar
2 wsdl4j.jar
2 wmsd.jar
2 tar.jar
2 soaplab.jar
2 servlet.jar
2 saaj-impl.jar
2 saaj-api.jar
2 qaqc.jar
2 mysql-connector-java-5.1.6-bin.jar
2 mail.jar
2 lsid-client.jar
2 log4j-1.2.8.jar
2 jython.jar
2 jts-1.4.0.jar
2 jsch-0.1.31.jar
2 jena.jar
2 jaxrpc.jar
2 jaxrpc-spi.jar
2 jaxrpc-impl.jar
2 jaxb-impl.jar
2 jaxb-api.jar
2 jargon_v2.0.jar
2 jacorb.jar
2 gt1.jar
2 gnu-regexp-1.0.8.jar
2 forester.jar
2 concurrent.jar
2 commons-net-1.2.1.jar
2 commons-logging.jar
2 commons-logging-1.1.jar
2 commons-httpclient-3.0.1.jar
2 cog-jglobus.jar
2 cipres_framework.jar
2 castor-0.9.5.jar
2 axis.jar
2 antlr.jar
2 antelope.jar
2 ant.jar
2 alltools2.jar

2 alltools.jar
2 ImageJ.jar
2 HelloWorld.jar
2 GeoVista-PCPVis.jar
2 CipresKeplerRegistry.jar

#2 - 05/13/2009 11:26 AM - Matt Jones

This is probably an underestimate of the duplicates because I'll bet that some of the jar files that are in only once are actually duplicates that have a different filename but the same contents as other jars in this list. Of course, eliminating these is a good first start, but we might also want to check these. I found it useful to simply sort them by name which shows similarly named jar files even if they aren't exact matches. For example, for xerces there are 4 unique filenames but 7 jars, and its hard to tell from the filename alone which are dups:

xerces-2.4.0.jar
xerces-2.4.0.jar
xercesImpl-2.7.1.jar
xercesImpl-2.8.1.jar
xercesImpl.jar
xercesImpl.jar
xercesImpl.jar

Aaron Schultz did a pretty comprehensive analysis of the jar files and created a list with unique names for each version. That might be useful in helping resolve this problem.

#3 - 05/13/2009 04:11 PM - Shaun Walbridge

Perhaps better yet, use a checksumming algorithm to check for duplicates. Something like:

```
find . -name "*.jar" | xargs sha1sum | sort > jar-list.txt
```

Then:
cut -d" " -f1 jar-list.txt | uniq -c

Should give you a listing of hashes which are duplicated.

#4 - 05/14/2009 04:58 PM - Chad Berkley

I wrote a build task to identify all duplicate jar files within the current suite. I identified many jar files that had the same checksum. Most of them existed both in util and core. I moved a single copy of the jar file to common/lib/jar and removed them from core and util. This has significantly reduced the number and size of jars within kepler.

The ant task can be run with 'ant analyze-jars'. It creates two log files. One analyzes the jars by filename and the other analyzes them by checksum. The filename analyzer is still finding some duplicates that are not the same file, but probably contain much of the same content. I'm going to leave these alone for now until I'm sure my changes today haven't messed anything up.

#5 - 07/10/2009 02:26 PM - Chad Berkley

At least most jars have been checked. There are probably still a few duplicates, but the major ones have been removed. Closing this bug for now. If any major duplicate issues pop up, we can re-open this bug.

#6 - 03/27/2013 02:25 PM - Redmine Admin

Original Bugzilla ID was 4013