The Sourcerer Project

Exploring Open Source software

search.....

Home Datasets Database Tutorials Publications People Contact Us

Eclipse Tutorial

This tutorial is designed to show how to download, compile, install, and use the Sourcerer infrastructure to build a database of Open Source Software. It is designed to be run using Eclipse. The requirements for this tutorial are:

Oracle's Java SE Development Kit 8, here (it should work with alternatives, such as OpenJDK).
Apache Ant, here.
Eclipse Mars, here (also known to work with the previous version, Luna).
A Mysql server, here.
A test repository, here.

Start by clonning the Sourcerer repository:

```
[sourcerer-path]$ git clone https://github.com/Mondego/Sourcerer.git
```

where `[sourcerer-path]` is the directory where you cloned Sourcerer.

Preparing your Environment

Start by installing the Plug-in Development Environment (PDE) for Eclipse (here). The simplest way to do this is to, inside Eclipse, go to **Help → Eclipse Marketplace** and search for **"pde"**.

Next, inside Eclipse, go to **File → Import → General → Existing Projects into Workspace**. Press **Next** and under **Select root directory** (first option) browse to `[sourcerer-path]`. A group of projects should appear. Make sure they are all selected, and press **Finish**.

Wait a few seconds for Eclipse to build the workspace, and then you should see a list of projects (if you do not see anything, go to **Window → Show View → Project Explorer**).

> Do not worry, for now, with possible problems Eclipse allert you about when importing the projects. We will fix them later.

You will need to create three folders in your computer, as seen below:

```
~$ mkdir crawled-projects
~$ mkdir extracted-projects
~$ mkdir db-import-output
```

We created these folders in the home directory. This means, for example, that the folder **crawled-projects** is accessed by typing `cd ~/crawled-projects/`. We will follow this path, but there is no restriction where you create these folders, just be sure to adjust this tutorial accordingly.

The folder `~/crawled-projects/` should contain Java projects. To start, you can download our test repository (above) and move the contents to this folder.

Follow the instructions here to set-up mysql and create a database. Create a user for the database and give it writing permissions. Save the **DATABASE-URL** (for ex.: localhost if the server is running in your machine), **DATABASE-NAME**, the **DATABASE-USER** and **DATABASE-PASSWORD**.

Building your Environment

From all the projects you can see in Eclipse, the most important is **bin**. In this project, inside the only existing folder, **launch**, you have the projects split in four groups. Open the folders so that your environment looks exactly like this figure:

Latest News

JAVA 8
July 31, 2015

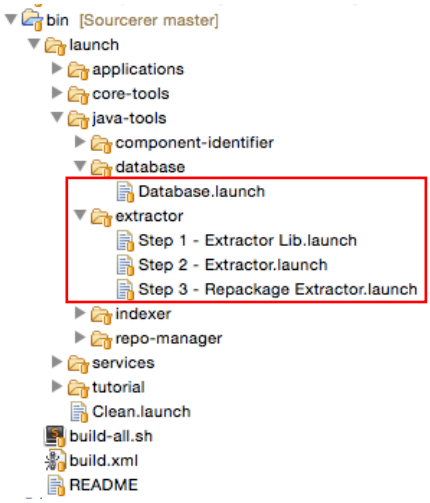
The Sourcerer tools can now handle Java 8's lambda.

NEW WEBSITE
June 11, 2015

2015 sees the redesign of Sourcerer website. Take a look around and let us know what you think.

Quick Links

- Sourcerer public repository on GitHub.
- Mondego group - we like large systems and large data
- Donald Bren School of Informatics



The launch files under **database** and **extractor** are the only ones you need to build. This is easily done by **Right Click on *.launch** → **Run As** → **[launcher-name]**. For the **extractor**, you will need to run them in the order they appear.

You can track the process with the **Console** window on Eclipse (**Window** → **Show View** → **Console**).

Now, go to the **extractor** project, and **Right Click** → **Refresh**. The path errors this project had should now disappear.

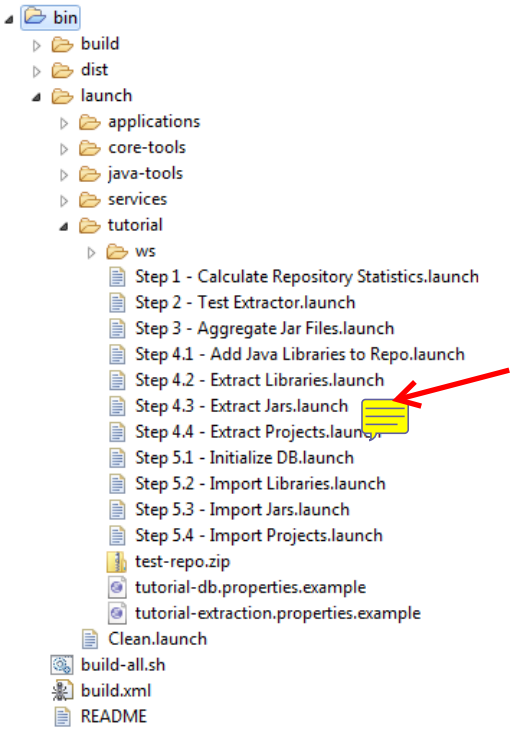
> Sometimes running a process on Eclipse that uses PDE is tricky, because even though the process has not finished yet, Eclipse might tell you so. This is a non-solved bug (check here). If, for example, you are trying to run **Step 3** but you can not, it might be because **Step 2** has not finished yet, even though Eclipse is telling you so. Just wait a few seconds after each step and everything should run smoothly.

Want to know more?

Under the project **bin** you find **launch** files to build all the tools in the Sourcerer infrastructure. They are as easy to build as the ones we just showed. Give it a try!

Running your environment

Open Eclipse so that your projects window looks like this:



Before running anything, you need to copy or rename the files **tutorial-extraction.properties.example** and **tutorial-db.properties.example** to proper .properties files **tutorial-extraction.properties** and **tutorial-db.properties**. Then you should edit these and adjust the data in them for your local environment. These files should look something like this:

tutorial-extraction.properties

```
input-repo=~/crawled-projects/
output-repo=~/extracted-projects/
report-to-console=true
force-redo=true
```

tutorial-db.properties

```
input-repo=~/extracted-projects/
output=~/db-import-output/
database-url=jdbc:mysql://DATABASE-URL/DATABASE-NAME
database-user=DATABASE-USER
database-password=DATABASE-PASSWORD
thread-count=1
report-to-console=true
```

The next step is to run the steps under **tutorial** sequentially, from **Step 1** to **Step 5.4**. The first two steps are just to give you some statistics about your repository and tell you if the **Extractor** is configured correctly.

If you just want to run the extractor, yo only have to run until the **Step 4.4**, and you just have to edit the properties file **tutorial-extraction.properties**.

> These steps imply a very large number of pre-processing tasks, as a huge volume of information is extracted for each projects and later imported into the database. The steps are, therefore, very slow. Please be patience and be sure each task finishes before starting the next one.

Please refer to the tab **Database** to see some examples of queries you can do.

Want to know more?

The raw dataset **sourcerer_repo_2011** contains the same type of information that we find in `~/crawled-projects/`, and the extracted dataset **sourcerer_repo_2011_extracted** contains the same type of information that we find in `~/extracted-projects/`, although on a much larger scale.

The creation of the database shows the steps we took when creating the **Database** of Sourcerer. In particular, the raw dataset extracted dataset **sourcerer_repo_2011_extracted** contains the same type of information that we find in `~/extracted-projects/`, and the database we generated has the same schema and the same type of information found in the Sourcerer database.

> Did you have any problem running this tutorial? Do you have a sugestion or a comment? Please reach us on: pribeiro@uci.edu.