

Stemformatics Data Portal Linux Australia Grant Submission

Project Name

Stemformatics Data Portal

Aim of the Project

Stemformatics is a like a web based pocket dictionary for stem cell biologists. It is based on the principles of open source software and collaboration.

One of the strengths of Stemformatics is that we have manually curated and processed 300+ public datasets from the original raw data. Each dataset can take a large amount of time, compute and biological understanding to process.

This is even more significant when you realise that over 29% of the datasets we have processed have failed our quality checks. By sharing this information with the rest of the research public, we can help to reduce duplicated effort and speed up the research process.

Some of our overall aims are to :

- Make the current Stemformatics data findable, accessible, re-usable while separating it from the web hosting instance
- Develop an open data portal around the Stemformatics data ecosystem
- Make the portal easy to distribute
- Create all of these under an open source license (Apache License, Version 2.0)

You can look at more details here in these slides from December 2016 :

https://drive.google.com/open?id=0B9dIBx_Elj7wblRGNHhwYW9Ib0U

Measurements of success

As this is a long term goal, we are aiming to ensure we can build an easily maintainable and distributable system from the start. To do this we aim to:

1. Build a web layer to:
 - a. search for datasets and/or samples that goes to a dataset page and
 - b. show a Dataset page with links, PCA plot and metadata processing information
2. Ingest 5 public datasets manually potentially using Object Store (on [NeCTAR](#))

* *Findable, Accessible, Interoperable and Reusable*

3. Store the metadata in a database to enable easy searching
4. Create Ansible scripts to build on top of the NeCTAR cloud
5. Interoperability with Stemformatics instance
 - a. Stemformatics web hosting to import dataset from this data portal

Estimated cost breakdown

\$5000 to be provided as a bursary for a student in the Masters programs [Master of Information Systems, Master of Information Technology and Master of Engineering (Software)] at University of Melbourne. This will be invoiced before the commencement date of Mid-Late July.

Project Team

- Rowland Mosbergen, University of Melbourne, helped develop the Stemformatics web instance and wrote BioJS libraries
- Othmar Korn, University of Queensland, developed pipeline interfaces
- Steve Englart, University of Queensland, new team member working on Ansible and system administration
- Isha Nagpal, University of Melbourne, helped develop the Stemformatics web instance and integrated the BioJS libraries into Stemformatics

List of open source repos for Stemformatics

- [Web instance](#)
- [S4M Pipeline for Next Generation Sequencing pipelines](#)
- [Stemformatics_tools for microarray pipelines](#)
- [BioJS website \("search for Stemformatics"\)](#)

Person responsible

Rowland Mosbergen can be contacted at rowland@stemfomatics.org

Project updates

Stemformatics will commit to provide a project upgrade in the middle and at the end of the student's project. This project will be for the Semester 2 intake, with commencement date being Mid-Late July with a conclusion in Late October.