From Stars to Baht: Broadening the economic impact of astronomical data handling techniques in Thailand

Data management plan

We will ensure from the outset that protocalls are in place to systematically manage all data associated with the project. By the nature of the research, the project will collect and produce diverse datasets. The raw data we will receive from our external partners will be in the form of digital values and strings in either ascii (.csv) or excel spreadsheets. One of the aims of this project is to make this data more searchable and accessible for our external partners by setting up (No)SQL databases to contain this data. The data derived from our ML analysis can be stored more simply, in ascii files, for example. No raw data will be published, but data arising from the research will be published from the research where it is not commercially sensitive.

Because of their potentially commercially sensitive nature, any data transferred between the external partners and our team will be done so in an exncrypted state (whether by physical media such as USB stick or over the internet). Once on site at MFU, the raw data will immediately be backed-up on MFU's multiple backup servers. Once this has been done, raw data will be sent (via ssh, which is encrypted once the connection is established) to the UK to be backed up on Sheffield's secure servers. A Structured Query Language (SQL) database will be created to track all digital files.

Once the original raw data are backed-up on both sites, work will start on researching the best means to incporporate the data into a distributed non-relational database. The data management system we will use for this is *Hadoop*, which is widely used throghout the IT sector, thus ensuring the long-term readability of the data. Most of our data analysis will take place withing the *Orange* data mining software environment, which is a open-source and under constant development. These steps ensure that the data data collected and produced by the project will be readable for at least 10 years.

The main data that will be of use to other researchers are descriptions of the technologies we will develop to address the needs of the external partners. As with any research, our project will involve trial and error - testing different databasing and analysis techniques. The main publishable outcome of the work will be the results from these tests in terms of accuracy, reliability and speed of the database system and the ML-based analysis. By signing the Letter of Support, our external partners already have an understanding of this. Where not commercially sensitive, all raw data will be made publically available, together with any non-commercially sensitive derived data. This includes data from Pibulsongkram Rajabhat University (User behaviour and service provision) and MFU (Academic and educational development office and M-Store).

The outcome of our research – the completed databse design and working model – will be made available by enacting the Release to Public policies of Mae Fah Luang University. We will package our data analysis algorithms in an *Orange* module as well as providing a self-contained software package complete with user interface to the external partner. Any non-commercially sensitive code will be shared publically using the GitHub software sharing website.