

AUSTRALIAN CANCER GRID

Marianne Hibbert

Bio21:MMIM from a pilot to a national Grid

The Bio21 Molecular Medicine Informatics Model (Bio21:MMIM) is a virtual repository of clinical and genetic data sets. Physically they are located at various organisations, but are 'federated', that is they are able to be integrated, searched and queried seamlessly via a federated data integrator. Bio21:MMIM was established as a pilot (funded by the Vic Government) where a variety of data sets have been integrated (clinical, genomic, etc) for colon cancer, diabetes and epilepsy, and is now about to expand. It recently received DEST funding to increase the data in the neurosciences, oncology and diabetes areas, add a respiratory network and add additional sites, including one Tasmania. There is a proposal to extend this layer to create a National Life Science Grid with Oncology as the first area.

1. National or regional infrastructure and support services

A National grid structure is proposed where the APAC grid services are utilised.

2. Access Policy Issues relating to authentication, authorization, practices, right management

The infrastructure of Bio21:MMIM enables discovery research to be accessible via the Web with security, intellectual property and privacy addressed from any location within Australia. However, researchers must gain authorisation to access data, and inform/obtain permission from the data owners, before the data can be accessed. The project will be examining more effective and efficient processes for authorisation, tracking, and audit of user access across the collaborating universities and research institutes.

3. Legal and ethical issues, such as privacy, intellectual property, liability

The legal and ethical issues surrounding health data have been addressed in the pilot. The challenge is data long-term data curation within an environment where privacy is paramount and budgets are constrained.

4. Sustainability, data curation and preservation

Almost all current clinical research databases in Australia suffer from limitations including: inaccessibility (locally accessible only), lack of good data validation and curation, and a general lack of robustness and good design. This is because such databases were typically set up to support local research projects – the databases have grown and were never designed to support distributed access and the higher quality of data demanded for longitudinal studies. The solution is to "refactor"¹ each database in a way that supports data entry, validation, sharing, security and privacy. Each research area will need its own database conversion and federation effort and the software expertise, tools and techniques used for these conversions and federation will be "reused" across projects.

The Bio21:MMIM project is committed to sustainable technology and migration and has a business model designed to maintain its viability. This project aims to use the expertise and protocols developed as part of the APSR (Australian Partnership for Sustainable Repositories) project to develop and implement best-practice procedures for the long-term sustainability of the Bio21:MMIM repositories.

5. Data usage issues, such as discovery, authentication, representation, etc

To enable discovery of relevant data and resources within the MMIM infrastructure, we propose to add new publish and discovery mechanisms, as well as supporting metadata services, to the Bio21:MMIM architecture. These mechanisms and services will draw, where appropriate, on the mechanisms used for publishing, discovery, and metadata services used by the grid itself. We also propose to extend the techniques used by MMIM for handling privacy, ethics, and provenance to the general grid architecture.

¹ That is make changes and improvements incrementally, preferably minimizing changes to existing databases.

6. Data management issues – provenance, cleaning, merging, format conversion, workflow, archiving etc

Data management issues are the key to usefulness of the repository and there is no magic. The processes to ensure the integrity of the source data through to its availability in the repository in a consistent format at a national grid level will be developed and documented in this project.

7. System requirements - (software, storage, communications, etc) and capabilities (reliability, resilience etc), system support

The system requirements and proposed support structure will be outlined during the talk.

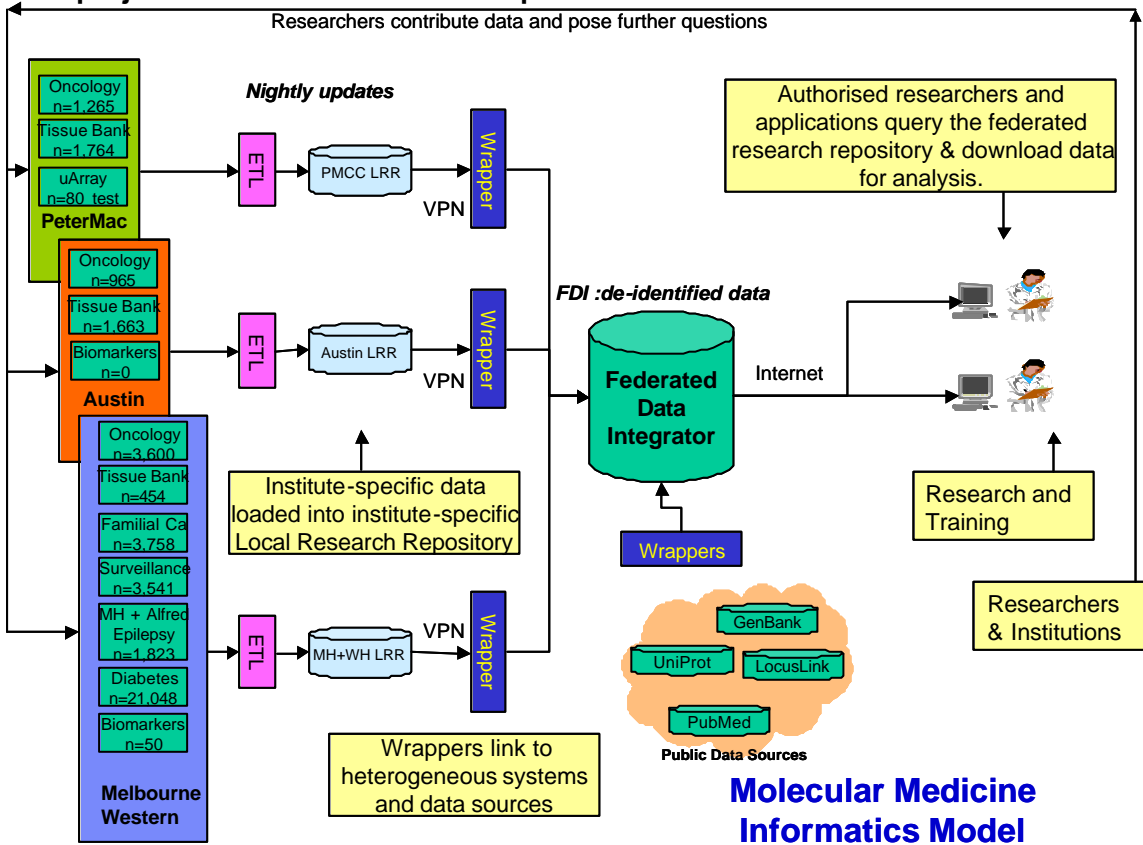
8. Skills development

Training of researchers will occur in the Bio21:MMIM systems as well as in Statistical and Bioinformatics analytical applications. In addition the Information technology personnel in effective and efficient use, development and maintenance, however there is a need to identify the skills development required for the grid services.

9. Data content issues – data quality, metadata, reliability

Covered under data management issues.

Pilot project – current status before expansion:



In summary this project will establish the infrastructure and processes to provide a life science repository and grid where the sustainability are addressed within the privacy requirements of the data. The project is at the start of a major expansion and intends to work collaboratively with APSR and APAC to establish a robust system that will meet the needs of biomedical and clinical researchers into the future.

Marianne Hibbert
6th October 2005