DATA CURATION FROM HIGH PERFORMANCE COMPUTING FACILITIES – A CASE STUDY

Venki Balasubramanian
Ian Thomas
Amir Aryani
Heinrich Schmidt

eResearch Office            RMIT University, Melbourne, Australia

Project Details

- Developed at eResearch Office @ RMIT University, Melbourne
- An Australian National Data Services (ANDS) funded Data Curation (DC) Project.
- Data Curation (DC) for:
  - Physics Simulation Packages:
    - VASP, CRYSTAL, SIESTA and GULP
  - High Performance Computing (HPC) Facilities
    - NCI-NF, VPAC, RMIT HPC
- The aim of the DC application is to harvest research data and publish metadata.
- Project Team:
  - 1 Project Director
  - 1 Project Manager
  - 2+ Developers
  - 2 Domain Users
DC Application from HPC

In an abstract view the components of DC project are systems by themselves.

DC application have at least 5 systems separately.

Have limited control over external legacy systems in terms of reliability and availability.

Relationships are often ad-hoc.

Functionality of each individual systems may not be known exactly.
Interoperability

- Key challenge is to make these systems work together
- e-Research deals with different domain-specific software systems
- Lack of established e-Research standards for data representation and transfer.
- Need a adapter (convertors) to allow data to flow between these systems.

Reusability

- Reusability saves much in time and development cost.
- The use of open source software can give more opportunities for researchers.
- DC uses myTardis – an open source metadata repository by extending its functionality.
- Uses external handle server
Security

- Security is always an issue in any application that involves data transfer between systems.
- It is important in eResearch to create reliable data repositories and process for who can access them.
- It is crucial that the user authentication and authorisation happen seamlessly.
- The Australian Access Federation and other federated services are a solution.

Domain Users

- The project would not be successful without the collaboration of the researchers (domain users).
- They want the software to be effortless to use as much as possible.
- It should assist them in storage, retrieve and access to their research data.
- Our experience is that researchers wish to fit new software into their existing workflow and not change unless the solution is widely used.
- Unlike most users, some researchers are conservative in that they care little about technology.
What We Learned!

*eResearch* - of the Researcher, for the Researcher and by the Researcher (!)

- It's all about the users (Researchers)
- One solution for all domain researchers is not possible
- eResearch solution should evolve not develop in one step
- Some researchers are conservatives
- Data curation is not just BACKUP
Cont…

- Lack of knowledge on eResearch by some academics/researchers.
- Should make use of domain users’ software knowledge (such as Unix scripting in DC)
- Collaboration with Researchers is essential
- Researchers just need good enough solutions (this is both good and bad!).
- Developer have to explore open source while building eResearch tools

Thank You!

Questions?