

Data Capture from High Performance Computing Multi-User Environments

- Software to interface with simulation packages (VASP, CRYSTAL, SEISTA, GULP) at NCI-NF, VPAC and RMIT HPC facilities to collect, store and extract metadata from very large data outputs of simulation runs.
- Satisfying data retention responsibilities of the university
- Key requirement: No (or minimal) changes to existing workflows

Stakeholders:

- Prof. Salvy Russo's Group in RMIT School of Applied Sciences (users)
- ANDS (funders),
- RMIT e-Research Office (developers),

VASP
CRYSTAL
SEISTA
GULP

Package
generates
data...

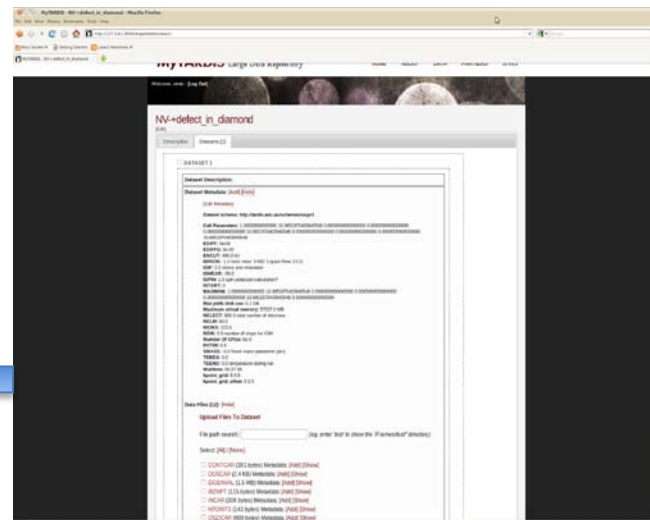
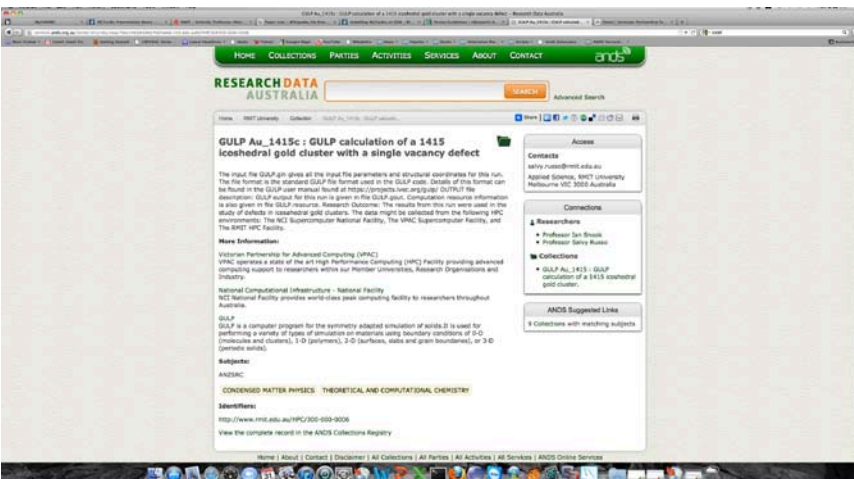
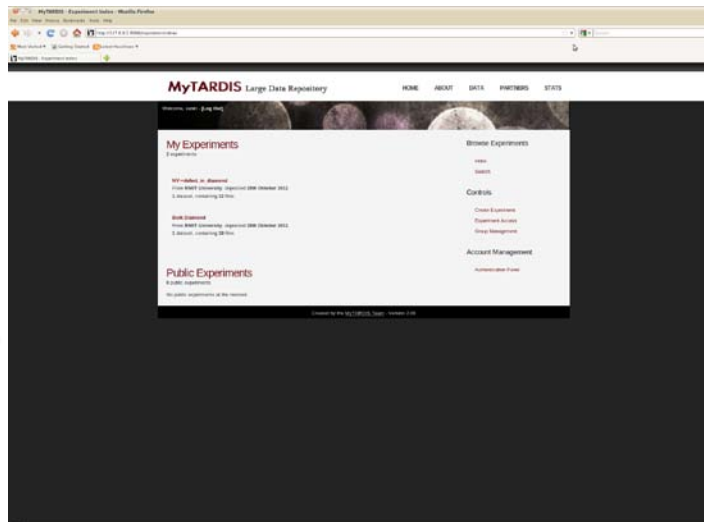
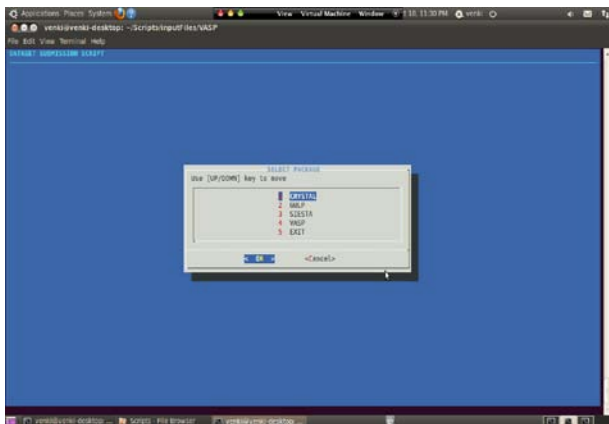
user scripts
get metadata
from users...

...metadata/datasets ingested
into myTardis

...data to
institutional
repository

..metadata published to
ANDS

...automatic extraction
of metadata ...



Where does it go next?

As is	To be (2012)
Pilot	Production Quality
Under eResearch Office	Under ITS/University
1 domain, 1 user group	Multiple HPC user groups, additional packages
Single-discipline	Muti-discipline
≈0 support	ITIL, training materials, manuals, helpdesk support, security audit etc.

Challenges

- Heterogeneity
- Multiple facilities at the same time
- Data from commercial packages (so no plugins)
 - Need for connectors
- Tension between:
 - standardised environments and discipline-specific metadata and research practice
 - different types of metadata, data and workflow
- Compliance with data curation policies

Observations

- Lack of technical support: home grown versus standard solutions
- "One system fits everyone" versus discipline specific approaches
- Data curation \neq backup
- Researchers want to create own workflow: resist having workflows fixed by others