Transdisciplinary eResearch-Reflections on Providing eResearch Services from Centralised ICT

Jonathan Padavatan[1]

[1]University of the Witwatersrand, Johannesburg, South Africa, jonathan@padavatan@wits.ac.za

EARLY DAYS

The centralised ICT services of the University of the Witwatersrand (WITS -ICT) recently embarked on an ambitious journey to offer cluster computing and data management services to the research community at Wits University. A small heterogeneous cluster is currently hosted by Wits-ICT. The lead partners, who built the 1340 core, 1PB cluster, Wits Core Cluster (WCC), have been the Electrical and Information Engineering, Physics and Wits Bioinformatics, with support from Computer Science and Mathematical Sciences. Currently the climate change group have invested a further a further 240TB of storage and local compute to the existing infrastructure as a first meaningful step to expanding research computing. Being a one-person division, in its infancy, the challenges and opportunities on the research computing arena at the university are starkly apparent in the era of data deluge. This paper will lay out a “lessons learned” journey in migrating the support services from the "sole expert” in the bioinformatics research unit, to more sustainable centralized support role, while strengthening the relationship between researchers and support staff.

Decisions on software platforms and providing user support for specialized research computing requirements carry the bias of a “what can be rolled out”. This bias invites a tension in the fact that researchers compute needs are usually complex. Hence, designing workflows transforms into a creative collaborative process that attempts to balance this tension in a resource stressed environment. Current challenges include escalating costs (and unreliability) of electricity; acquiring specialised skills on large data handling and HPC support while learning on the job; and an overall funding-stressed environment in the tertiary education sector in South Africa. Innovation and collaboration are now necessities for a sustainable research computing community. A few significant use cases will be discussed in this paper with a special mention of the data management platform IRODS. Its features include high flexibility, expandability and excellent support that are most suitable for a diverse range of research data from a centralized infrastructure perspective.

The role of iRODS in metadata handling is that iRODS provides for abstract annotation of every data object, collection and storage resources within the system. Data objects are organized into collections, which are very similar to sub-directories. To enforce a schema, iRODS validates the metadata before, during, or after insertion into the catalog. This can be done programmatically, to allow the user to build policy around 'good metadata'.

The allows for other systems to trust the metadata downstream in the data flow process. IRODS allows the user to create powerful, customised workflows in large data set environments.

EVERYONE NEEDS STORAGE - PRIORITISING DATA MANAGEMENT

The bioinformatics group at the Sydney Bremmer Institute for Molecular Biology are the power users of the cluster. Establishing the right data management tool proved a critical initial step in fielding these requests. Provisioning access to storage through the FreeNAS operating system installed on KVM configured virtual machines (vms) was the optimal solution, based on existing expertise for support and recovery. Being a shared resource, current storage requests are received from variety of research fields with highly valued data sources. This is where the complexity resides. Issues of duration of storage, physical and cyber related security and ease of data access are concerns discussed with the researchers. A "one size fits all" approach must be initially discarded to determine the optimal solution.

Table 1 summarizes the diversity in requests

Table 1: Current Storage Requests

Date Sources	Storage	Metadata Handling
Hominid fossils	15TB	no
Geosciences	80TB	yes
Climate Change	240TB up to 1PB	Yes , with visualisation

iRODS (Integrated Rule Orientated Data System) was selected for its scalability and highly configurable data management features. Geosciences initially expected network file share, but on demonstration of metadata handling tools such Mytardis and CKAN, they were keen to build a metadata schema from scratch. However, the primary power user, the Bioinformatics group were also in need of a metadata handling system. Support for IRODS from www.renci.org is excellent and played a key factor in the roll out component of the architecture. We choose to split the installation into 3 different vms: a database server, a zone server and a web server for the Metalnx web interface to the data. The rule engine resides in the zone server; hence each research group would initially have one zone. iRODS allows for compute to data and vice versa. Hence iRODS enables the research computing specialist to provide high performance computing, big date storage and data management to a cross disciplinary audience. Transdisciplinary eResearch skills are required to meet the needs of this variety of end users.

TRANSDICIPLINARY ERESEARCH – UBUNTU AS A PLATFORM

Recently I attended a strategic planning session with WITS-ICT management team, near the Kruger Park. We resolved to adopt Ubuntu as a value. Ubuntu is the philosophy that the operating system (which many eResearch professionals use daily) is based on. The philosophy states that person is a person through other persons. This value is particularly apt for the complex field of research computing support which brings together the skills of a diverse group of experts in order to serve the needs many very different but equally important research projects. Instead of operating in their traditional separate realms, technical support staff and academic researchers are now walking on a journey of discovery together.

REFERENCES

1. Dessus, Sebastien C.; Hanusch, Marek. 2019. *South Africa Economic Update : Enrollments in Tertiary Education Must Rise (English)*. South Africa Economic Update; no. 12. Washington, D.C. : World Bank Group. http://documents.worldbank.org/curated/en/173091547659025030/South-Africa-Economic-Update-Enrollments-in-Tertiary-Education-Must-Rise
2. Lech Nieroda, Lukas Maas, Scott Thiebes, Ulrich Lang, Ali Sunyaev, Viktor Achter, Martin Peifer *iRODS metadata management for a cancer genome analysis workflow.* BMC Bioinformatics2019 **20**:29 https://doi.org/10.1186/s12859-018-2576-5
3. https://renci.org/research/irods/