A Collaborative Australian Characterisation Informatics Strategy

DRAFT for open comment (25th August 2017)

> *“Australian characterisation infrastructure encompasses a wide selection of instruments and capabilities that are united by the need to address common informatics challenges. The multi-modal and distributed nature of the research, science and supporting instruments is a challenge that has been united in the past by the Australian characterisation community being able to successfully coordinate across key informatics initiatives.”*

Characterisation refers to the general process of probing and measuring the structures and properties of materials at the micro, nano and atomic scales. It is essential across natural, agricultural, physical, life and biomedical sciences and engineering. Characterisation facilities, as outlined in the 2016 National Research Infrastructure Roadmap, provide researchers in Australian universities, research centres, and industries with critical infrastructure, including both instrumentation and expertise, to enable quality research outcomes in an efficient and cost-effective manner.

These facilities are a key capability that underpin flagship Australian research collaborations including ARC Centres of Excellence which are both significant users and partners in the development of future characterisation techniques and applications. The Australian Characterisation community and our partners bring together thousands of researchers who are driving the future of Australian imaging and innovation.

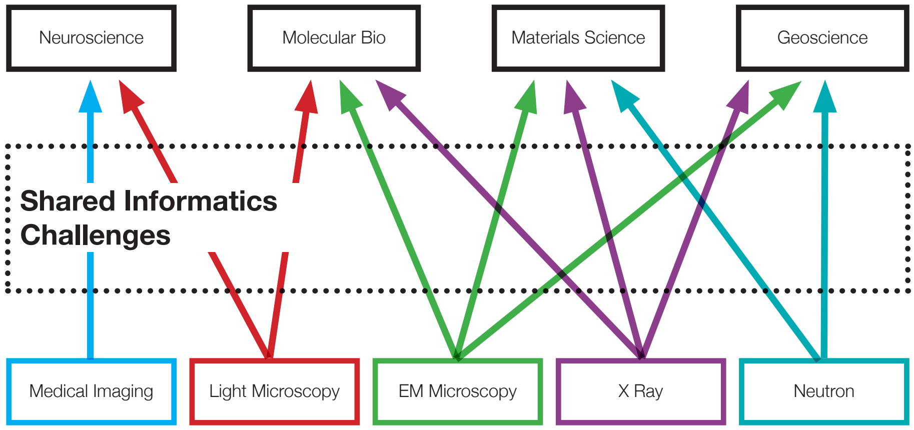
Characterisation has become a capability where informatics infrastructure, expertise and best practice is essential to turning data into new discoveries.

As a collective, the Australian Characterisation community shares a number of significant informatics challenges, and we are working together to plan and implement strategies to overcome these challenges.

Emerging informatics challenges in the field and the way we plan to address them are:

Challenge:	Requires:
Scale and complexity • Data volumes are increasing with new detector technology • Processing requirements are increasing – the “science is being affected by compute” • Opportunity offered by multi-site, Australia-scale data capture introduces specific challenges • Custom and specialised instrumentation requires custom workflows • Increasing need to perform analysis “in-experiment” • Analysis of of data from multiple instruments requires specialised skills and familiarity with the data	**A national infrastructure program that supports:** • **Community driven instrument integration and data management initiatives** to capture data from the point of generation • **Rich online environments** for characterisation in the cloud and on HPC platforms • **Simple and seamless access** across instruments, repositories and analysis environments • Programs for **specialised and big data producing instruments**
Working with digital objects • Data remains unpublished, is difficult to reuse, and it is often unclear whether it can be trusted • Data curation is a second priority to publication and data is often non reusable • Research software is often closed source and impossible to validate, can be challenging to newcomers and is often very specific to particular problems • Research outcomes are difficult to validate and are often unreproducible • To increase return on investment research outputs need to be machine readable	**Making Characterisation digital objects Findable, Accessible, Interoperable, and Reusable (FAIR)[1].** **To achieve this requires:** • **Community** efforts to increase application of FAIR principles • **Coordination** across Australia to provide leadership and organisation • **Commitment** by data producers, in partnership with research communities and tools developers to increase uptake of FAIR principles
Expertise is rare • The characterisation community is increasingly reliant on data-science skills • Digital expertise coupled with applied characterisation knowledge is rare • Cross modality analysis requires multiple areas of expertise to facilitate new insights and discoveries • Where knowledge is available it is often in accessible beyond a local node or institution	**A national program to spread knowledge and underpin change, which includes:** • **national training** to uplift data skills across characterisation users • a **national network of characterisation informatics experts** with expertise in research software engineering, and specialist skills in specific modalities, as part of an overarching Australian characterisation experts network

[1] doi:10.1038/sdata.2016.18



The Australian characterisation community provides a wide range of techniques that are applied across a variety of scientific domains. Common across these are a set of shared informatics challenges.

Scale and complexity

Researchers in the biological, biomedical, earth and physical sciences are acquiring ever-larger multi-dimensional data sets and are increasingly using multiple instrument platforms; e.g. X-ray CT together with EM and Raman spectroscopy. This presents a Big Data challenge not only in terms of managing and curating these data over the research life cycle, but also in terms of analysing data to facilitate new discoveries.

Government supported characterisation infrastructure:

AMMRF, ANSTO, NIF, and flagship institutional facilities - supply instruments with diverse capabilities, along with associated technical expertise, to over 5,000 researchers annually.

Australian teams have a history of successful collaboration to deliver a range of data and informatics solutions for the sector.

These include:

Multi-modal Australian Sciences Imaging and Visualization Environment (MASSIVE)	Specialised High Performance Computing for peak instrument data processing, including ANSTO-Australian Synchrotron, CryoEM, and data collections generated at NIF and AMMRF sites. Seed funded by NCRIS. Used by 2000+ researchers across 100+ institutions.
NIF Informatics	Supports users of NIF instrumentation in the areas of managed storage systems, online visualisation and automation of workflows. The goal of these systems is to make it easy for users to achieve the F.A.I.R principles. 1400 users have been supported by the NIF infrastructure, including 85 international collaborations.
MyScope (AMMRF)	Provides eLearning for advanced microscopy, used by 100k people annually across Australia and Internationally.
The NeCTAR Characterisation Virtual Laboratory	A NCRIS-funded national program to make research computing more accessible to a new generation of scientists which has involved 7 Go8 Universities, 3 NCRIS funded capabilities, CSIRO and other organisations. Tools, environments and workflows developed by the CVL are used by 2,446 researchers.
MyTardis and ImageTrove	MyTardis and its national deployments, such as ImageTrove, integrate 50+ instruments across Australia, including major deployments at NIF, Monash University, UQ, RMIT, UoN, ANSTO. It provides easy mechanisms for facilities to integrate instruments with centralized storage, management and the cloud.
RDS Image Publication	A community of 5 Australian eResearch nodes working together to integrate 100+ instruments with cloud based data management platforms across 20+ Australian instrument facilities.
RDSI ReDS Australian Coordinated Characterisation Data Space	Provides over 3 petabytes of storage space across three RDS nodes to characterisation researchers.
ANSTO Tools	Technique specific tools developed by national and international collaborators and partners are used in data pipelines to reduce, analyse, simulate and visualise instrument data.
DaRIS	DaRIS (Distributed and Reflective Informatics System) is a subject-oriented informatics framework that is deployed at NIF facilities and other locations.

This document has been produced by a writing group with representatives from Monash University, AMMRF, ANSTO and NIF, based on the outcomes of a series of open Characterisation Informatics workshops held on the 28th of February 2017, and the 12th May 2017, involving stakeholders from AMMRF, ANDS, ANSTO, BPA, Monash, NeCTAR, NIF, RDS, UMelbourne, UNSW, UQ, UoW, UWA, USydney, Agilent and NVIDIA.