Data Down-Under

- work in progress

Current Practice in Australia

• Currently no common data or metadata management standards/practices are being followed. Metadata schema and content is vendor dependent.

• Data storage is not coordinated at a national level and storage is left to the individual laboratory – typically DVD based with some use of tape.

• For instance, at Sydney there are 2 X-ray machines in Chemistry, a SMART 1000 and an APEXII-FR591, and in biochemistry there is a Rigaku rotating anode machine operating 2 ports. In both labs data is backed up to DVD.

• Some anarchy in chemistry as storage is decentralised across the groups who undertake their own crystallography, in addition to the service storage. Reciprocal Net installed but not actively used

• Currently no coordinated approach to data and metadata for the beamlines at the Australian Synchrotron

But there are changes underway …

Summary of Developments

• Multi-institution collaboration (DART-MMSN projects) for ‘end to end’ X-ray data management. Collaboration includes extensions to Indiana CIMA model to allow SRB and MCAT use, intention to adopt CCLRC Scientific Metadata model. Adoption or interoperability with ICAT.

• Southampton e-Bank system.

• Support use of imgCIF.

• ANSTO led proposals for changes to nature of NeXus – aimed at NeXus 3.0

• National Collaborative Research Infrastructure Scheme (NCRIS) – includes e-Research infrastructure (cyberinfrastructure) and national data storage service. The Australian Parnership for Advanced Computing is to be ‘re-born’ through the NCRIS process.

• Victorian e-Research Strategic Initiative (VeRSI) to provide medium term storage for the synchrotron. Will link with the NCRIS process.

DART: Dataset Acquisition, Accessibility, Annotation & e-Research Technologies

- A DEST funded project (http://www.dart.edu.au & http://plone.jcu.edu.au/dart) to develop tools to handle the data & information management needs of diverse research environments.
- DART is developing tools to handle typical research data & information management needs, such as:
    1. collecting data
    2. managing data
    3. analyzing that data to produce useful information
    4. managing that information
    5. collaboration & annotation on the information
    6. publishing information
    7. searching on the information
- DART has developed three demonstrators:
    - Climate Research
    - **X-ray crystallography**
    - Indigenous Digital History

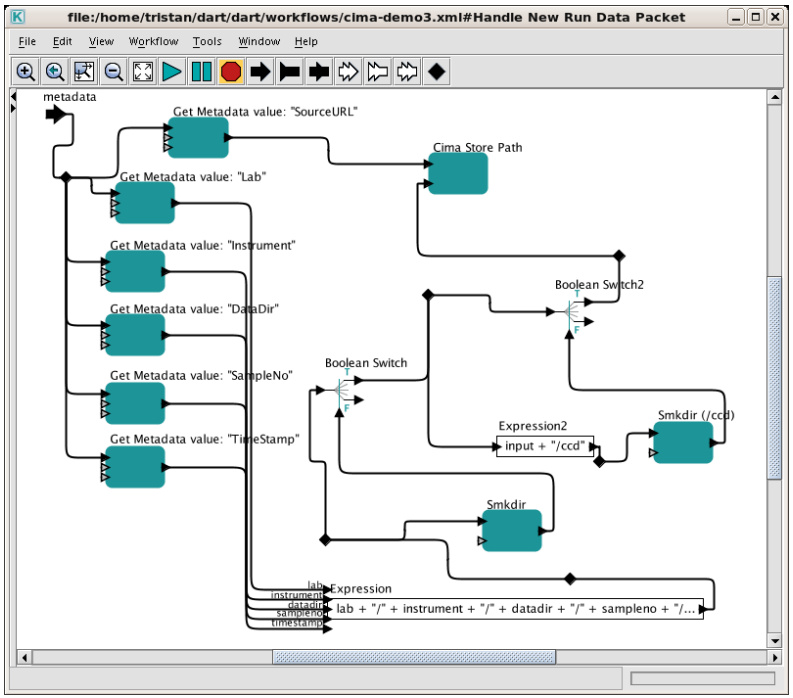
CIMA Data Management Changes

- Introduction of use Storage Resource Broker (SRB)
  - Grid storage with uniform interface to heterogeneous resources over a network.
    User defined metadata structures

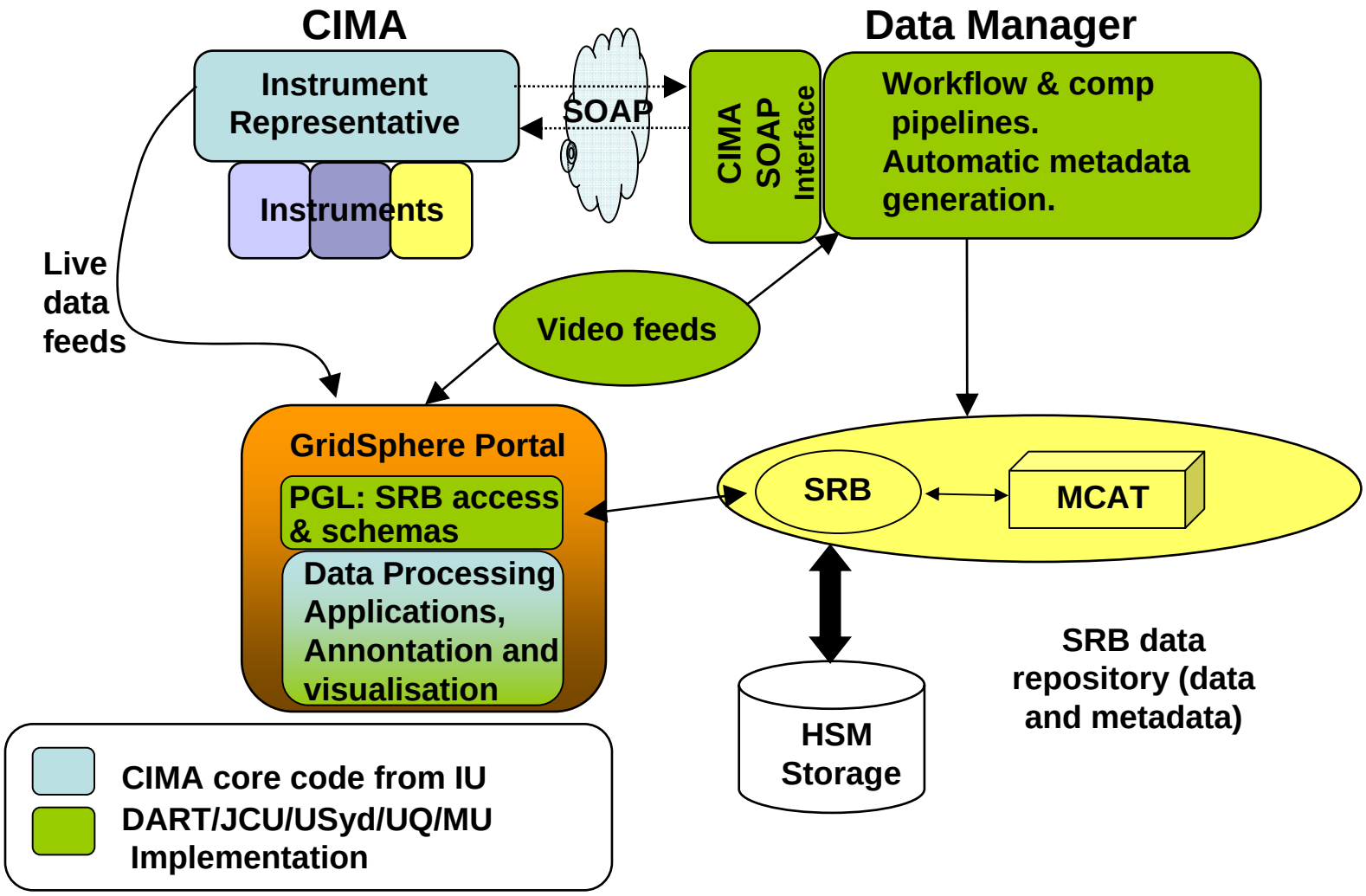
- Kepler workflow manager:

  - ability to customise data storage *via* the workflow system

  - highly extensible, not restricted to SRB as repository



DART Implementation



DART Portal SRB Utilities

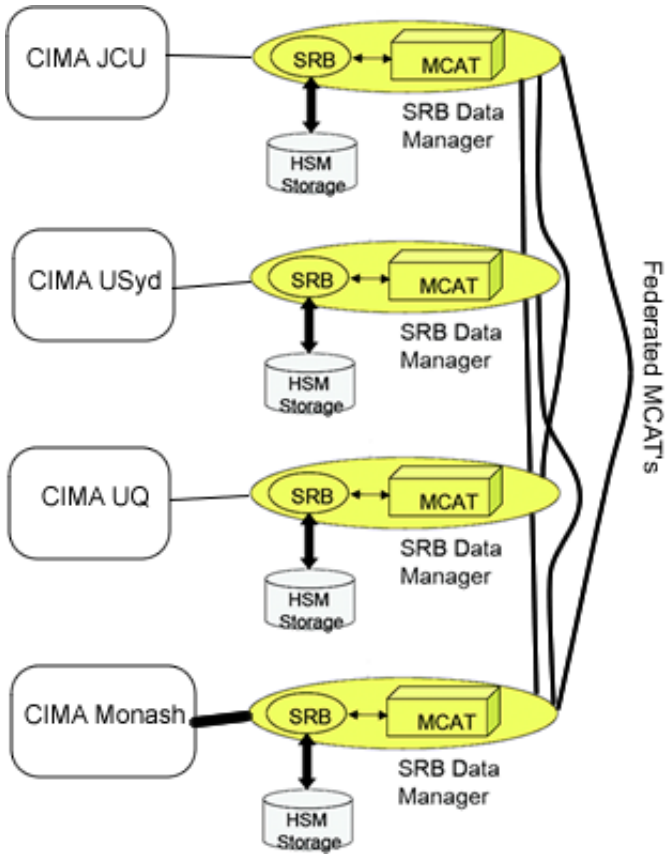
- Use of Personal Grid Library (PGL) for SRB data manipulation / metadata display
- Metadata schema definitions applied to experimental data – via JCU ‘chitter chatter’
- Stored experimental data is able to be easily retrieved, secured and annotated

Data Federations with SRB

- Initial Instrument Representatives deployed at JCU, USyd, Monash & UQ

Goal

- Each site having their own Data Manager and SRB storage facility which is federated across sites
- Shibbolith based AAA and Virtual Organisations
- Federate/Replicate data into a a national SRB store…



Australian Partnership for Advanced Computing

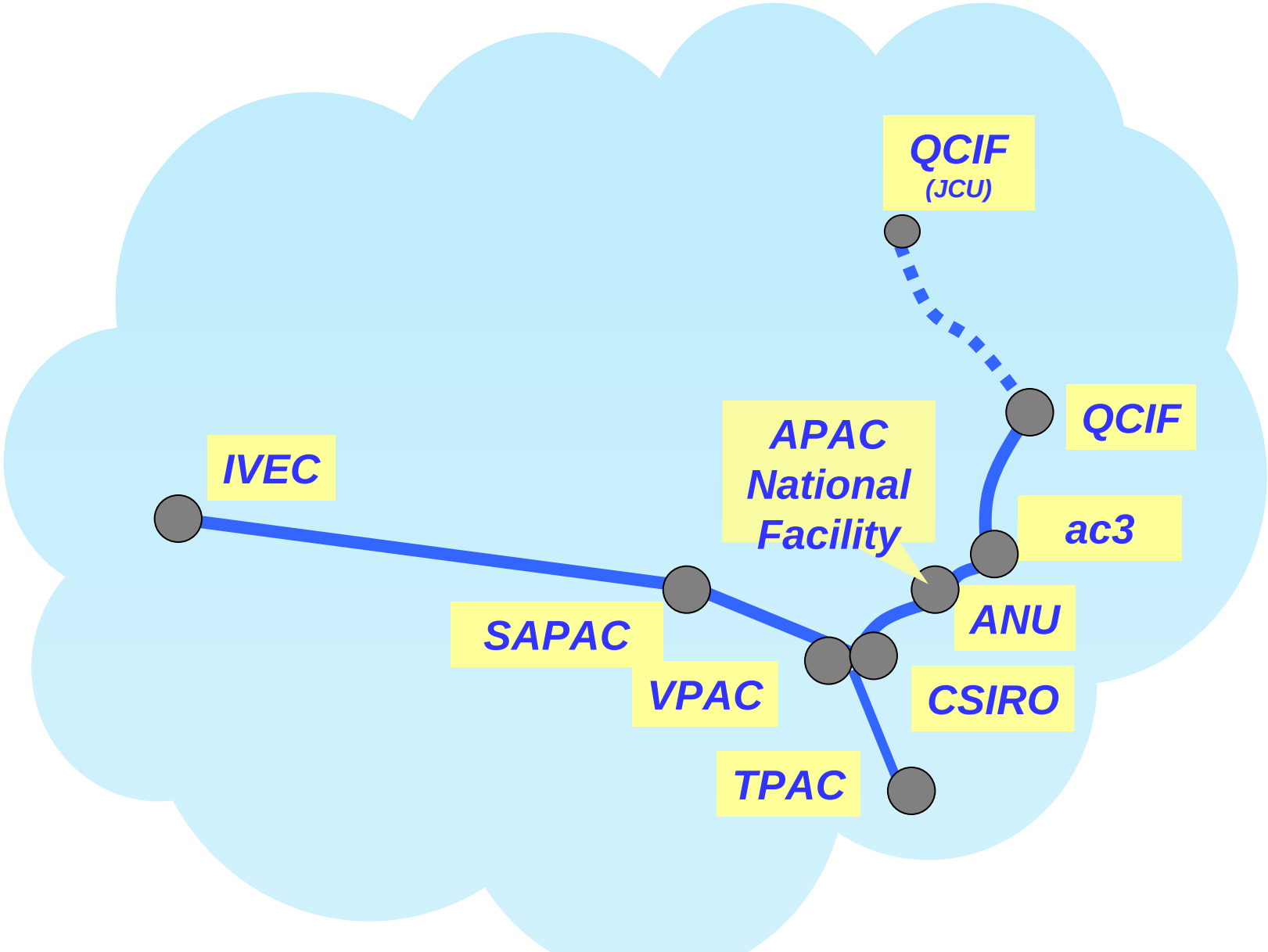
"providing national advanced computing, data management and grid services for eResearch"

Partners:

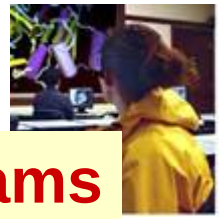
- Australian Centre for Advanced Computing and Communications (ac3) in NSW
- CSIRO
- iVEC, The Hub of Advanced Computing in Western Australia
- Queensland Cyber Infrastructure Foundation (QCIF)
- South Australian Partnership for Advanced Computing (SAPAC)
- The Australian National University (ANU)
- The University of Tasmania (TPAC)
- Victorian Partnership for Advanced Computing (VPAC)

4500 CPUs, 3PB storage

APAC Partner Sites



Concept of the APAC National Grid



Research Teams

Sensor Networks

Data Centres

APAC National Grid

a virtual system of computing, data storage and visualisation facilities

Instruments

Other Grids:

Institutional

International

NCRIS - National Collaborative Research Infrastructure Scheme

National Plan to invest AU$500M in medium scale collaborative research infrastructure across 5 years 2007-2011

15 Investment areas - including: structural *characterization –neutron and X-ray*

NCRIS investments are expected to develop and execute plans to ensure e-Research (cyberinfrastructure) tools and practices are embedded into their practices and data management

Data management is now a hot topic in Australia!

New Names and Structures

E-Reseach services

National Compute Infrastructure (NCI)
- APAC Nat. Fac. >1600cpu Altix
- Shoulder clusters

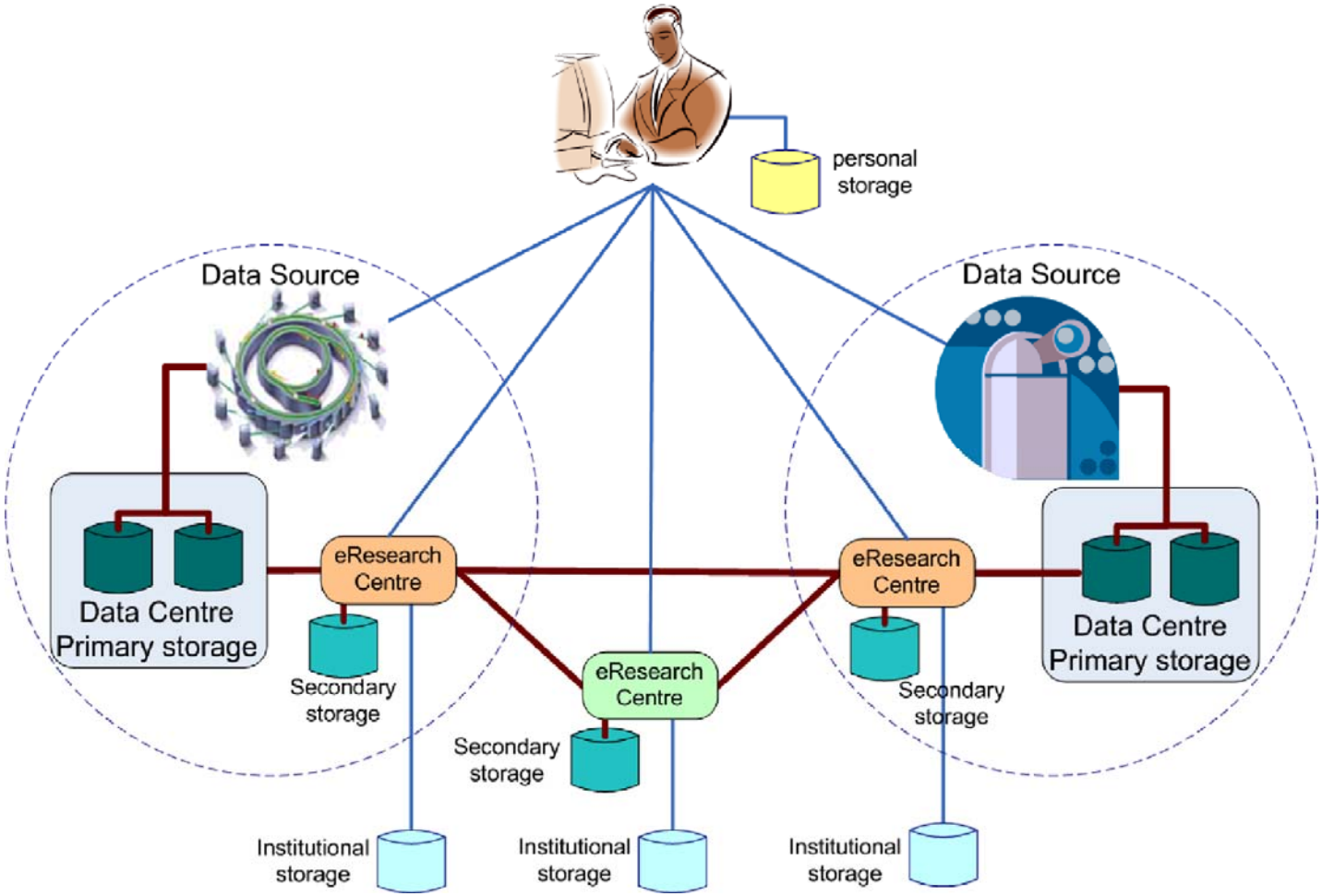
Interoperation and Collaboration Services (ICS)
- Old APAC Grid

Aust. Nat. Data Service (ANDS)
- Federation of Mass Data Stores
- Long term archiving and curation

Australian Access Federation (AAF), AREN - Network

National Coordination Council

VeRSI proposal national data architecture



Synchrotron Data Storage

Users want …

- Raw data archived for an absolute minimum of 5 years
- Ingest system for other data (eg neutron diffraction, lab X-ray, CD…)
- Data security including audit trail
- Metadata should include machine readable beam & instrument logs, annotated video recordings and experimenter's notes
- At least two storage sites – one at sync and one nearby. Linkage to institutional repositories

Summer in Australia?

Thanks ….