data.melbourne

@DrStevenManos
University of Melbourne

Figshare Fest, UTS, Thursday 9 Feb 2017

THE UNIVERSITY OF MELBOURNE

RDSI Research Data Storage Infrastructure

nectar

NCRIS National Research Infrastructure for Australia
An Australian Government Initiative

VicNode

ands AUSTRALIAN NATIONAL DATA SERVICE

Data story at Melbourne

context

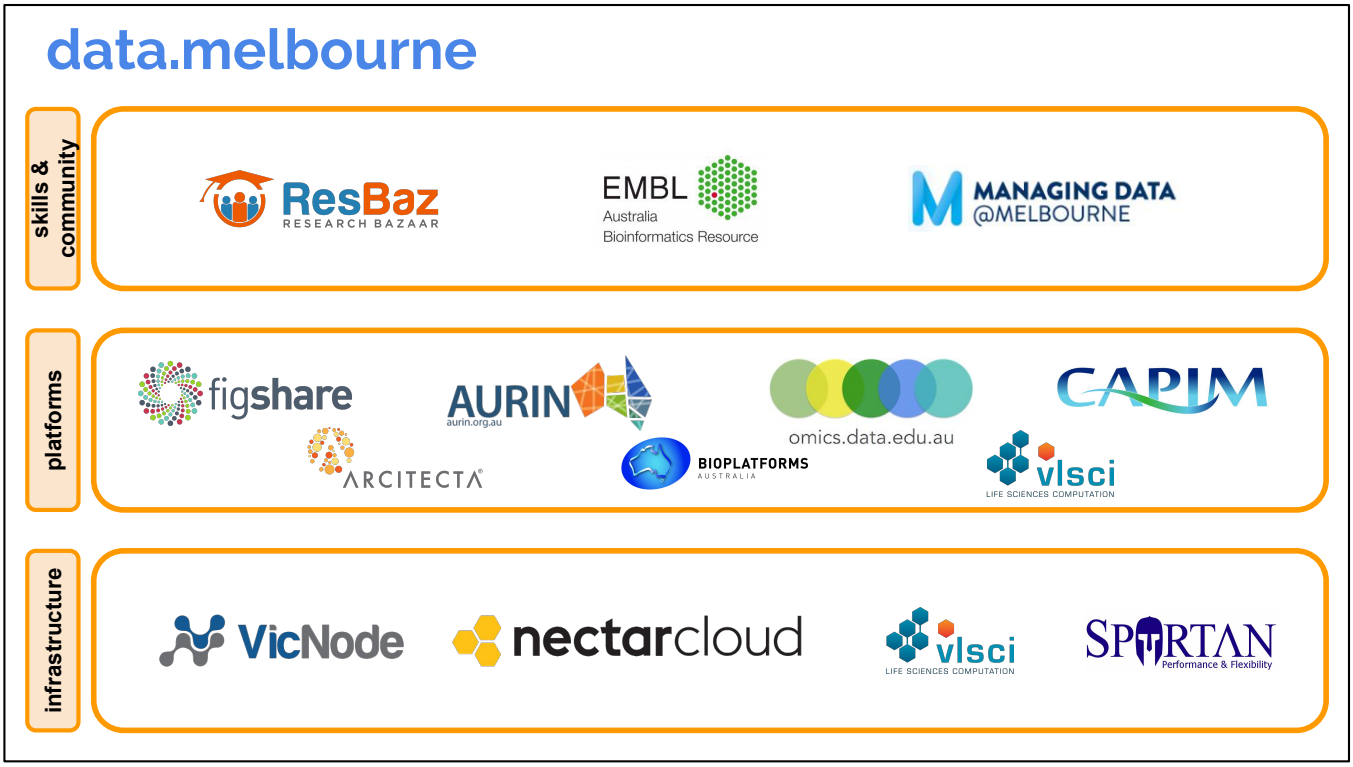
- Set within the Parkville and wider City of Melbourne Knowledge precincts, with strong connections to hospitals and research institutes
- A comprehensive and research-centric University made up of 10 diverse discipline based faculties
- 140+ ‘research ICT’ support staff spread across a diverse range of support units
- The university also sits within a strong social sciences precinct, including • Melbourne School of Government, Melbourne Institute of Applied Economic and Social Research, the Melbourne Social Equity Institute and the Centre of Advancing Journalism, Oxfam, The Conversation, and so on..

Context that is relevant to the shaping of the data storage and how it will evolve over the coming years.

The setting results in partnerships and collaborations that are often complex. Getting to the bottom of how data plays a role in those engagements and collaborations is a key thing we need to do.

Wide range of support units have evolved over the last decade, with more than 140 support staff spread across a diverse array of groups. This evolution reflects a dizzying array of discipline needs, specialised skills in data, compute, research software development and so on..

Photo is of the systems garden on the Parkville campus, which is nestled behind Royal parade and Bio21. New buildings planned for this part of campus will play host to data oriented life sciences and biology research.



We can talk about the Melbourne data activities as an ecosystem of things grouped in infrastructure, platforms and skills and community development.

A comprehensive approach to dealing with data required consideration of all of these.

These reflect spectrum of needs, from

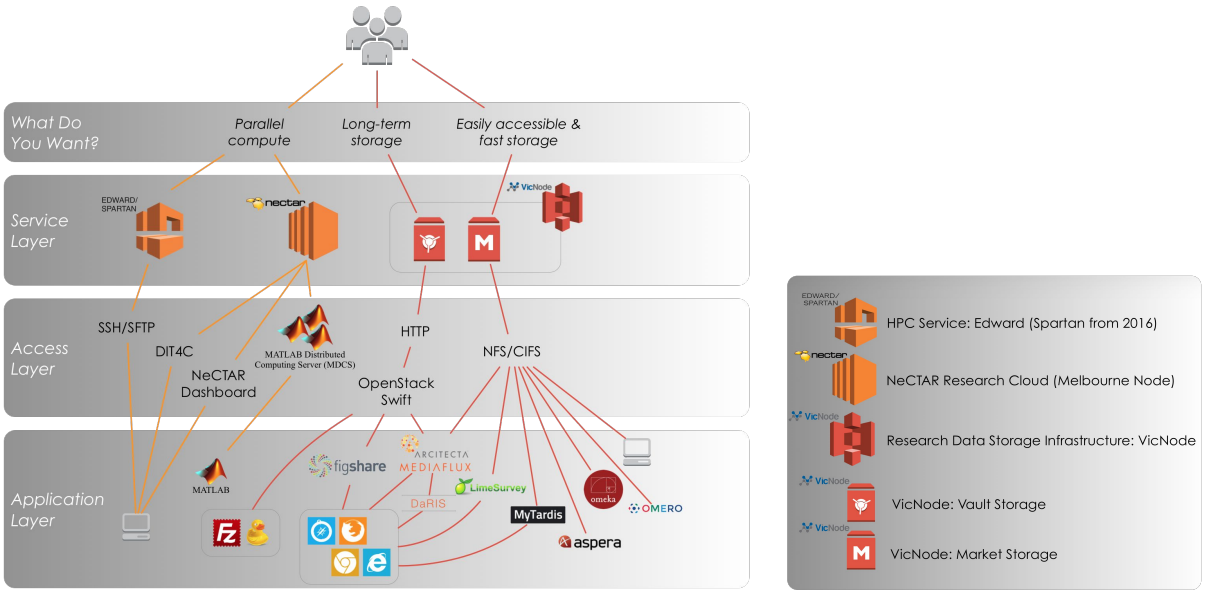
‘active data’ through to archive data (e.g. nectar cloud/spartan platforms versus vicnode vault..)

discipline specific to generic data platforms (e.g. omics and figshare)

Training and community development for all and for particular areas… (resbaz versus EMBL-ABR for example..)

compute & data foundation services

infrastructure



The foundation infrastructure piece really starts with an answer to the questions of where are we going to hold the data? The intent is a single place that acts as an integration point to make that data useful, connected and available.

The data storage piece represents about 13 PB of installed and 5PB of usable infrastructure across a range of technologies.

This integrated set of data management platforms which offer further capabilities in assisting researcher to address the challenges of research data lifecycle.

active data - cloud and compute platforms

infrastructure

Cloud

- 11,000 cores
- Self-managed virtual machines on demand, 24/7
- Connected to petascale storage
- “Urban Analytics Data Infrastructure”
- “Gana Burrai Health Outcomes in Goulburn Murray Region”

HPC

- Hybrid bare-metal and virtualized compute
- Expands to fill unused cloud resources
- Combination of performance and scale
- GPGPU capability (expanding in 2017)
- “Human Proteome Structural Prediction”
  - Needed 3 million core hours over 3 months

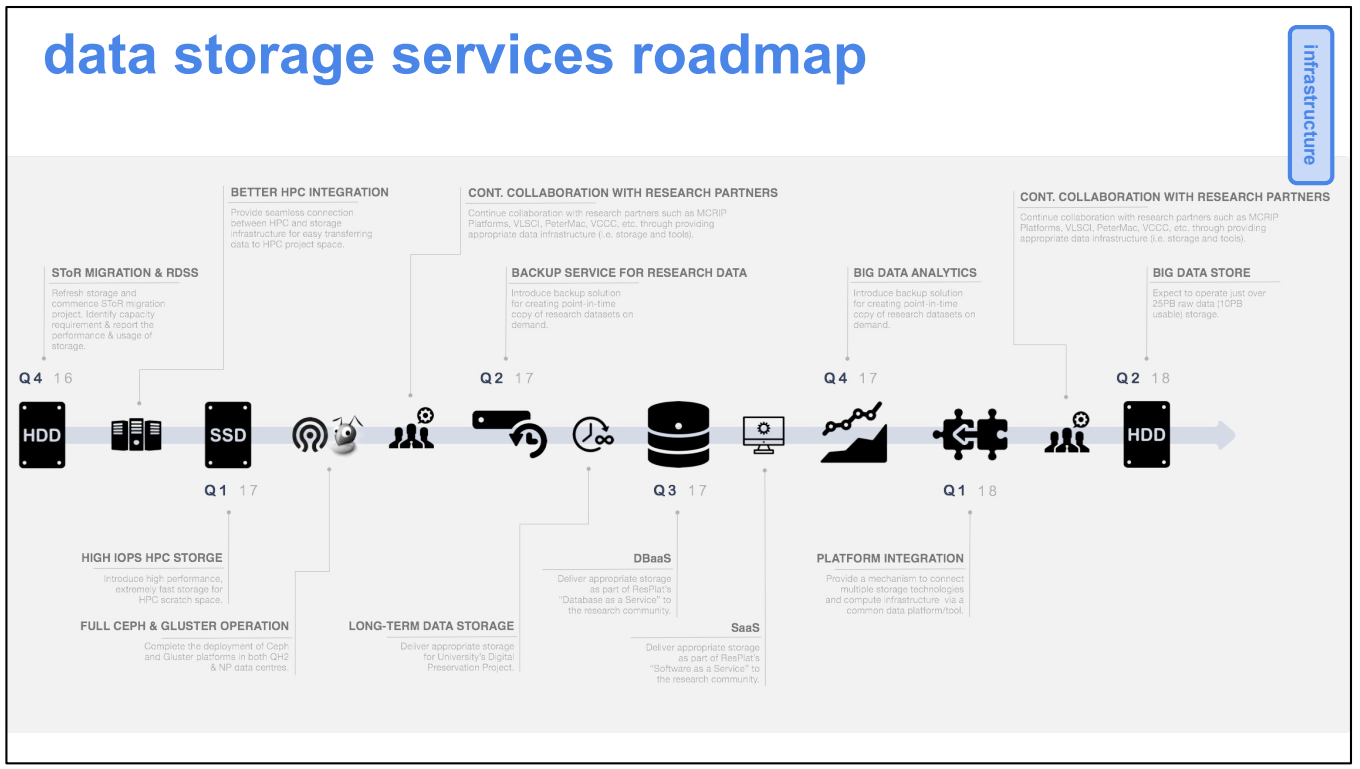
nectarcloud

SPARTAN Performance & Flexibility

Urban analytics - This project aims to develop the critical digital infrastructure required to underpin the next generation of data driven modelling and decision-support tools to enable smart, productive and resilient cities by bringing together expertise from six leading urban research centres across Australia.

Gana Burrai - Gana and Burrai means ‘mother and child’ in the language of the Yorta Yorta people. The project will undertake a trial to determine whether we can safely link together birth outcome and early childhood health information collected about Aboriginal and non-Aboriginal mothers, babies and young children living in the Goulburn Murray Region.

Human Proteome - 3 million CPU hours of computation to predict structure and function for the entire collection of human genes/proteins, of which a significant percentage (25-50%) have an unknown function.



"Big Data" usually refers to datasets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time and the "size" of Big Data is always a constantly moving target. In order to solve Big Data challenges, new set of techniques and technologies need to be developed in order to uncover large hidden values from those large datasets such as those generated by high-throughput instruments such as NGS, etc.

Big Data poses 3 main challenges, **Volume** (i.e. the amount of data), **Variety** (different data types) and **Velocity** (speed of the data being processed) and this data services roadmap has attempted to address those upcoming challenges with the followings:

Volume (amount of data):
Operate SDS solutions to address the ever-increasing needs of large-scale, extendable and secure storage.

Variety (different data types):
Provide variety of data management platform, e.g. Mediaflux and OMERO, etc. as a way of managing complexity of multiple data types, both structured and unstructured.

Velocity (speed of the data being processed):
To introduce analytic frameworks, e.g. MapReduce, Hadoop and Spark to help speed up the data processing over those large datasets using the underpinning HPC facility.

- Software-defined storage (SDS)
  - consolidate and fine-tune Gluster & Ceph storage platforms
  - expect to expand capacity to over 7PB usable storage

- Connecting with computation facility
  - deploy large high-performance scratch space for HPC
  - establish a better data management capability with HPC

- Big Data: management, analysis, publishing & sharing
    - build new analytics framework, e.g. spark, hadoop, etc.
    - better data management with long-term preservation capability (via Mediaflux)
    - provide mechanism to publish and sharing large data-sets

figshare

platforms

Operational since April 2015.. being used by Melbourne University researchers to:

promote their papers through their data

- https://figshare.com/articles/UQV100_An_IR_Test_Collection_With_Query_Variability/3180694

collect related research outputs – not yet published

- https://figshare.com/projects/Improving_the_design_of_a_conservation_reserve_for_a_critically_endangered_species/18313

publish evidence of reproducible research

- https://figshare.com/articles/Reference_environment_bootable_ISO_for_the_paper_Predictive_modelling_of_gene_expression_from_transcriptional_regulatory_elements_/2002317

provide supporting data to open access journals publications

- https://figshare.com/articles/Soil-atmosphere_methane_exchange_data_for_AU-WOM_and_AU-WRR_with_soil_environmental_variables/4578640

publish large datasets

- https://figshare.com/articles/Horsham_reflectance_dat/4596853

A range of data management platforms sit atop our data storage platforms. One of these platforms is Figshare, which has been operational at Melbourne since April 2015. The data itself, data files, image files, presentations, CDROM ISO images, etc. - are stored within the University of Melbourne data centres, but the front door, the Figshare service, is operated and managed by Mark and his mob.

figshare provides way for researchers to make their research data outputs discoverable and citable through DOIs, and allows external references to their journal publications. This gives researchers and easy way to meet journal requirements around publishing data supporting their papers and potentially allowing multiple vectors for other researchers to discover and cite those papers. All this is done through an interface that is designed to be self service with very little requirement for training or additional support. The institutional figshare subscription allows all of this to be done with our own institutional storage as well.

There are a range of examples of use at Melbourne.. But here are some which showcase the diversity of applications.

promote their papers through their data:
A nice example that links to the related published paper in the “References” section, Alistair Moffat is the Melbourne University researcher, the others are external I think. Also they request a paper citation if the data is used, so hopefully their open data will lead to more paper citations.
http://dl.acm.org/citation.cfm?doid=2911451.2914671

collect related research outputs – not yet published:
A collection of data, not yet openly available to support a research paper “Modelling species responses to extreme weather provides new insights into constraints on range and likely climate change impacts for Australian mammals”
http://onlinelibrary.wiley.com/doi/10.1111/ecog.02850/full

publish evidence of reproducible research:
Daniel Hurley has created a bootable ISO that supports his paper and put that up here.

provide supporting data to open access journals publications
An example used to support an open access journal publication “Soil methane oxidation in both dry and wet temperate eucalypt forests shows a near-identical relationship with soil air-filled porosity” , paper also links to the data.
http://www.biogeosciences.net/14/467/2017/bg-14-467-2017-discussion.html

publish large datasets
Not a really great example for something well described but they’ve published almost 10GB of data

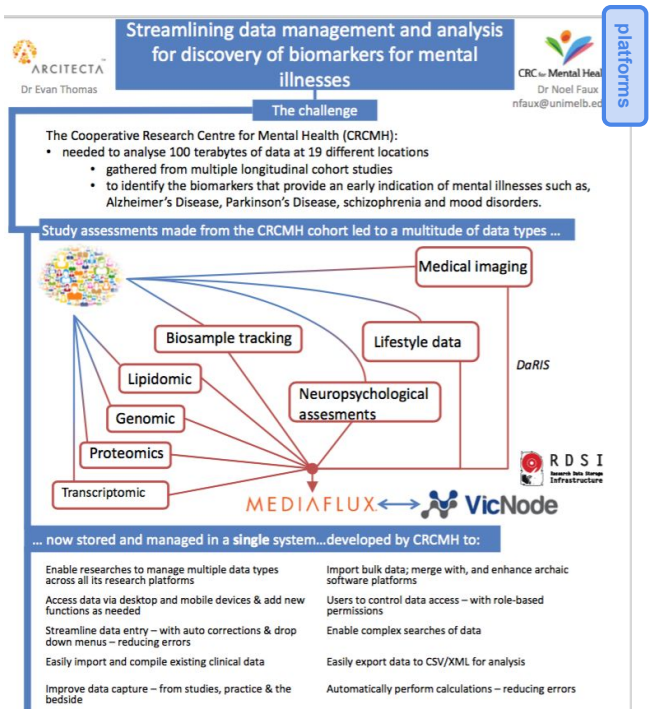
ARCITECTA mediaflux

A very capable, general and extensible data management platform for managing and preserving digital assets.

35 collections ~ 120 TB under management

Provide mediaflux capability in both generic and specialised domain-specific contexts:

- CohortConnect app - Jointly developed by Arcitecta and CRC for Mental Health. Holds Parkinson's disease data and now being re-used to hold Schizophrenia data
- VCA MCM film and TV archive
- Centre for Aquatic Pollution Identification Management (CAPIM) application for managing field samples
- TenToMen (an Australia-wide study of the health and lifestyles of a large group of Australian males)



A very capable, general and extensible data management platform for managing and preserving digital assets. It has advanced capabilities in security, meta-data and data modelling, big data, flexibly managing storage resources, data resilience and multi-protocol inter-operabilty.

The University delivers Mediaflux capability in both generic and specialised domain-Specific contexts:

Supplies general cloud-based project resources for distributed research teams

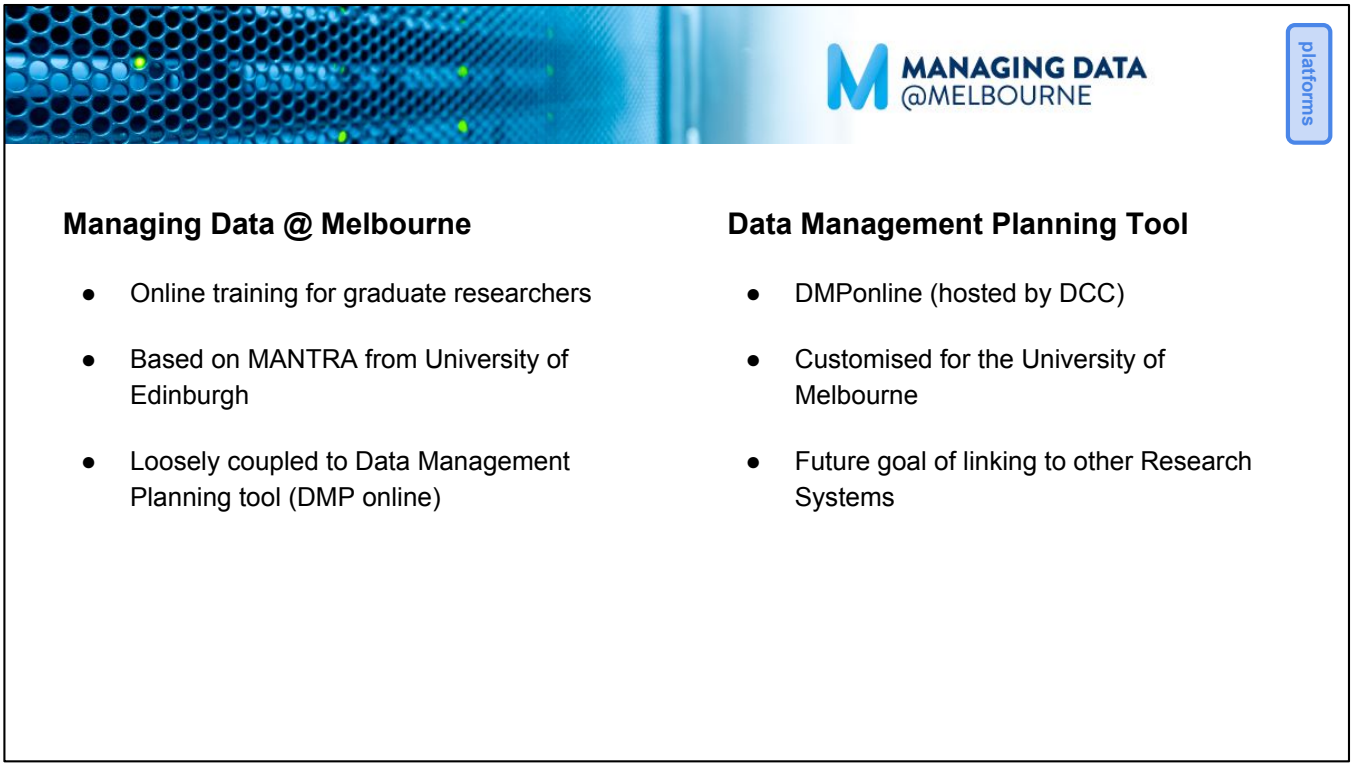
- Can utilise multiple account types (e.g. AAF, local, LDAP)
- Meta-data overlays for query-based discovery
- Easy-to use Java client for uploads and downloads
- **Data can be shared with links, downloaded or easily despatched to other end-points (sinks: e.g. Spartan HPC scp server)**
- **Collections can be accessed via multiple protocols (e.g. http, NFS, SMB, SSH)**
- **Mix of actively used and archival collections**
- Flexible management of storage resources with built-in disaster recovery processes
- **35 collections (120TB)**

<u>Domain-Specific</u>

- *DaRIS/Mediaflux*
  - Mature platform using Mediaflux for managing bio-medical imaging data with a focus on instrument integration
  - E.g. DaRIS is at the heart of data operations for the Melbourne Brain Centre Imaging Unit (7T MR & PET/CT scanners) for which DaRIS holds 13TB of data
  - DaRIS also holds many other collections such as the School of Dentistry's high-value Femur collection
  - Open source platform led by UoM with deployments at other nodes
- *Combine various Mediaflux frameworks* (e.g. extensible services, data models, triggers, queries) as needed
  - CRC for Mental Health CohortConnect app. for Parkinson's disease data (being extended to Schizophrenia)

    - Collaboration with the Victorian College of the Art's Film and Television School to create a platform for managing and sharing their film archive.
    - Omics for managing multi-omics data and inter-operating with the Genomics Virtual Laboratory
    - Centre for Aquatic Pollution Identification Management (CAPIM) application for managing field samples
    - TenToMen (an Australia-wide study of the health and lifestyles of a large group of Australian males)

- We operate a suite of tools and capabilities, each with its own strengths and foci. We are exploring, for example, ideas around managing valuable resources via Mediaflux whilst presenting those collections via other views such as Omeka.



New online training rolling out in March 2017 - based on University of Edinburgh MANTRA course

- Focussed on essentials of RDM for graduate researchers (1 hour)
- More localised Melbourne content from researchers and experts
- Linked to DMPonline (this is probably the key innovation) - participants use DMPonline while they do the online course so by the end of the training they have produced the start of a data management plan.

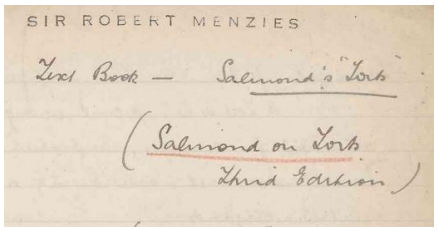
The Managing Data @ Melbourne program was developed after Melbourne piloted the University of Edinburgh's MANTRA program. The pilot was successful, but participants found it a bit too long and wanted more local and relevant content. While based heavily on the MANTRA training, we have reduced the number of modules down from 8 to 6 and the overall time to complete the course has gone from 2-3 hours for MANTRA to a bit over an hour for Managing Data @ Melbourne.

Data and Digitisation Services

platforms

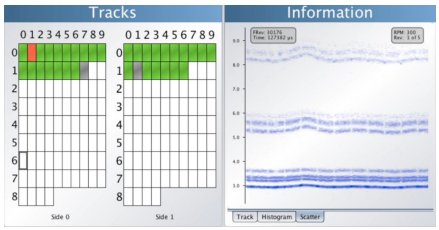
University Digitisation Service

- Expert digitisation of archives, manuscripts, documents, books, artworks, research records
- Self service in supported environment
- Value adding and metadata capture through advanced workflows



Data Forensics Service

- Analyse, identify, recover and audit research data in contemporary or legacy computer storage media
- VALA 2016 Innovation Award winner
- State of the art equipment for processing floppy disks, zip drives etc



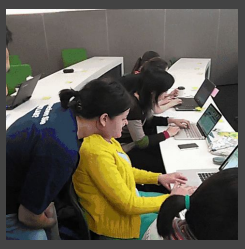
Digitisation service is transforming paper-based content into digital form, allowing it to be captured, archived and re-used in new ways.

Data forensics is capturing files and data on legacy, at-risk media so that it can become part of the research data collection of the University. Using state of the art equipment, old media can be imaged, indexed and ultimately discovered - potentially providing the basis for new research. Two significant projects currently underway are the Bill Mitchell archive (a pioneer in Computer Aided Design) and the Germaine Greer archive.

Image on left is from Robert Menzies student notebook - University of Melbourne Digitised Collections
https://digitised-collections.unimelb.edu.au/handle/11343/55401
Image on right is a disk being forensically imaged using a system called Kryoflux

The Research Bazaar digital skills training program

skills & community



1. Community
   a. A community of support for researchers
2. Campaign
   a. A campaign to empower researchers in digital tools and literacy
3. Conference
   a. A conference to bring the the community & campaign together

Runs early Feb every year.
2015 was the first year.
2016 went international, joined by Auckland, Perth, Dunedin, Brisbane, Sydney, Guayaquil, Vancouver, Oklahoma & Wellington running simultaneous conferences globally.

Key Values for ResBaz

1. Diversity
   a. Research area, institution, gender, ethnicity, career stage, department…
2. Open Access
   a. Reproducible research, teaching materials, live-blogging, open story-telling, no (low) cost fee…
3. Community
   a. Social events, loneliness, challenge-based learning, peer-to-peer, digital community, 'breaking bread together'…

The key values are absolutely critical here and largely represent themes of open science reproducibility, and so on, that in particular our RHD students and ECRs are heavily exposed so. Building a digital skills conference around these values is imperative…

The next ResBaz conference is on in two weeks, February 22 to 24. The first two conferences focussed on a particular training modality

But some changes will be made for **ResBaz Melbourne 2017**. Rather than an intensive three days of teaching/learning a digital tool, we're putting the 'Bazaar' back into ResBaz. We'll be offering a number of 'ResPitches' in a variety of digital tools and services. Attendees, grouped through digital tool preferences, will move around the Bazaar and listen to the stories of these tools: how can they be used to do research.. They'll also have the opportunity to sign-up to training workshops/meet-ups for each tool scheduled for the weeks following the conference.

Why the change? We figured that in terms of pedagogy learning a complex digital tool works best when it's staggered over time. Learning is a process, and at Melbourne we have the resources to provide a community of dedicated support year round. We wanted to shift ResBaz from being a "one-stop-shop" to an "induction" - a kickstarter - for the year to come. ResBaz 2017 will be a chance to explore, learn about and engage with the community.

This year… running across 14 sites internationally including.. Hobart, Brisbane, Dunedin, Christchurch, Wellington, Auckland, Perth, Vancouver, Tucson, Cuenca, Oslo and here in Sydney, right here at UTS in July.

Research Support Community of Practice

skills & community

What it is

- Community of practice for Liaison Librarians (research)
- Managed with support from digital scholarship
- Utilises domain knowledge and university expertise to facilitate research support across schools and faculties

Example Activities

- Tailored training in research data management
- 23 Research Data things
- Researcher@Library week (40+ activities over a week)
- Project-based collaboration

RESEARCHER @LIBRARY

Of course, community networks of support extend elsewhere through the University….

The Research community of practice provides a network of support across the University. Deeply embedded librarians are able to provide messaging around important data activities - such as data management, ORCIDs, digital tools, publishing etc

The CoP started as a way of upskilling librarians to support researchers, including requirements around data management. A tiered model outlines responsibilities for librarians and clear delegation pathways to support networks across the University for supporting more complex queries.

A large part of the CoP has been upskilling librarians, but leveraging existing knowledge across the University from Research Platforms, Digital Scholarship RIC, Legal etc

The CoP organised bi-weekly meetups for staff participating in 23 Research Data Things (mostly librarians, but not exclusively). The meetings allowed participants to discuss their progress and any issues they were having. At each meetup an Invited expert gave their perspective on a particular topic (eg data storage, copyright and licencing, linked data). These experts were sourced from the University and externally. ANDS were also heavily involved in meetups.

skills & community

EMBL Australia Bioinformatics Resource

- Admin and executive hub located at University of Melbourne, currently 10 network nodes around Australia

- Hub training events in 2016:
  - Data management for biologists and bioinformaticians
  - Bioinformatics infrastructure workshops on non-model organism annotation and curation, bioinformatics software, open and scalable training, registries
- Node training events:
  - Numerous end-user training workshops run locally at nodes

Of course, ‘data management’ makes a lot more sense in discipline speak.. And recently we have seen this exemplar of national activities based out of Melbourne… really the sort of thing we should head towards.

4 domain-specific workshops on Best Practice data life-cycle: plants, animals, microbes, health

Collected participants’ experiences/challenges and focussed on repositories, resources and issues specific to that research domain.

Most participants had very limited initial awareness about the numerous different repositories for different biological datatypes, but left with a much better understand of where their data should be stored/shared.

E.g. for Health there was a strong focus on legal and privacy obligations for patient data and how this affects data management and sharing

For plants several participants had commercialisation obligations that also limited their data sharing capabilities

Data size (transfer to and from repositories, storage, and sharing) was a fairly common ‘pain point’ for participants in the plant/animals/health areas

Domain-specific hands-on examples of searching for datasets in public repositories made it clear just how important it is to include rich, useful metadata.

Digital Preservation 2015-2025

Our ten year strategy establishing Digital Preservation at the University of Melbourne.

The digital material in scope:

- ➔ Research outputs
- ➔ Research data
- ➔ University records
- ➔ Cultural collections

Through 2014 to 2015 the University developed a digital preservation strategy, and its implementation began in 2016 as a funded university project. Its development was a collaboration between the Library and the Research IT support unit.

The picture here shows how we visualise the data preservation landscape, where the size of these blobs has meaning. Culture is the larger and more important area we need to address over the 10 years, with infrastructure coming second. The policy and organisation areas act in support of these two key outcomes.

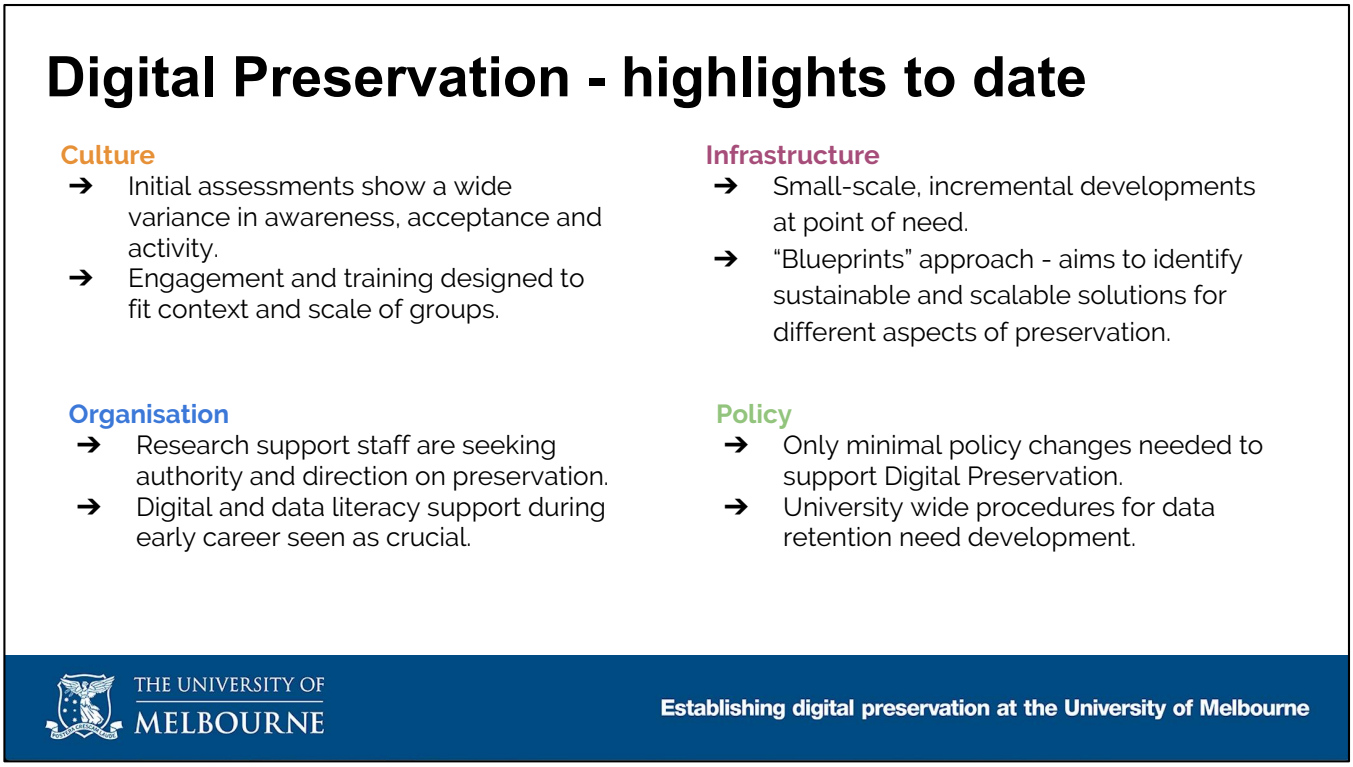
The cultural element considers training, education and community building.

Infrastructure is really about the data archive, preservation toolkits and any kind of tech to support people in research practice. We have lots of bits of this are here already as you've seen, but the technical integration and cross organisation integration and support is not quite there. For example, we already have FigShare and the institutional repository (Minerva based on DSpace), but they are not set up for preservation. We currently lack the workflows and integration to do this.

The organisational element is really about awareness building and policy and integration, integration not through a technology perspective but rather through people. How would we support and implement cross-organisational workflows (e.g. preservation in the institutional repository)?

And finally, on the policy area, it's really to see what policy gaps we have and plug them once we have data on what needs to change.

The biggest challenge is the willingness to get people to engage in the process. It is extremely difficult to engage the academic community, even if preservation problems are close to them.



Some of the highlights and learnings to date

Policy

On hold for now until we have workflows and infrastructure in place.

Culture

There is massive variance across research domains and within domains. MDHS very much aware of preservation, but other areas need more support and education, including Engineering and Cultural Collections.
The approach here is to put something in place for base level of awareness raising, targeting graduate researchers initially through the Managing Data @ Melbourne training program mentioned earlier. Engaging with their supervisors and support staff is essential as well, supporting them through providing materials that helps them understand the issues.

Organisation

We did in depth interviews with research support staff, getting thoughts on digital preservation for their particular research area. The key messages here were that data literacy support through early career stages is crucial, and that research support staff are seeking authority and direction on preservation, as well as data issues more broadly.

Infrastructure

We are taking an experimental and proof of concept approach here, and looking at blueprints to do small investigations into different areas, different workflows. The starting point is small scale and sustainable infrastructure or services that could grow with demand. Much of the research support and infrastructure is funded on a project basis.. But this needs to be all funded operationally, where we really want to build this into existing support teams and infrastructure.

Integration at infrastructure and organisational levels is absolutely key.

- finis -