“Real-world” Information Infrastructure Issues

Markus.Buchhorn@anu.edu.au
*Head, **ANU** Internet Futures*
*Lead, **APAC** Information Infrastructure Program*
*Grid Services Coordinator, **GrangeNet***

Overview

- Common and Uncommon Issues from Diverse Application Areas
  - Technical issues from e-research activities at ANU, APAC and GrangeNet
    - Also relevant to education
    - Large range of community ICT literacy
  - Scholarly Input and Output
  - Slice by issue, dice by application ☺

Application Areas - 1

- Geosciences
  - Minerals, oils and gases
  - Govt, Surveys, Industry
  - Many data sources (spatial and physical) and simulations
- Bioinformatics
  - Genetics, proteomics, …
  - Public datasets, private queries, private annotations
- Chemistry
  - Simulation, need data *services* mainly

Application Areas - 2

- High Energy Physics
  - Large expensive instruments, projects
  - Massive data, computation and simulation
- Earth Systems Sciences
  - Massive remote sensing data set, large and complex simulations
- Astronomy
  - Big data, complex reduction process, big simulations, long-term research

Application Areas - 3

- Linguistics, Musicology
  - Archives of digitised cultural material
  - Complex analyses
- Social Science Data
  - Census, health, surveys, …
  - Complex data structures, qualitative data
- Archaeology
  - Digitised physical materials, spatial and chronological data

Application Areas - 4

- Financial
  - Many sources, SX, FX, news, …
  - Timeliness and long time scales are important
- Music, Arts, Sports
  - Performance, formal and practice
  - Education focus

Longevity

Geo, Bio, ESS,
Astro, Ling, SS,
Arch, Fin, Mus.

- Sustainability
  - Metadata
    - Varied research schemas
    - preservation, curation and valuation
  - Data formats
    - Descriptions, Compression, lifetimes
  - Software
    - Algorithms, implementations, OS
  - Versioning
    - Recalculation, interpretation, validation
- Community valuation and quality
- Underlying infrastructure, technologies
  - E.g. mirroring for protection

Geo, HEP, Ling, SS, Arch, Fin, Mus.

Movement

- Performance requirements
  - *Mirroring/Caching*
- Collision with authorisation
  - *Some data cannot move from its host (in bulk)*
- Appropriate Delivery needs
  - *Remote/field access to data*
  - *Clients in a different 'circle'*
    - *Bandwidth, compute, language, culture*
- Movement Protocols
  - *Access and inter-repository*
  - *One standard is great – ten are not*

Rights

Geo, Bio, HEP, ESS, Astro, Ling, SS, Arch, Fin, Mus.

- Needs AAA to be working
  - Authentication, Authorisation and Accounting
- Privacy, Security
  - Anonymised data, needs to stay usable
- Ownership
  - Not always with the researcher
- Time-varying
  - Data sourced under old agreements
  - Rights vary by status of source
    - people die, agreements expire, …

Types

- Digital
- Non-Digital
  - *Paintings, Objects, Manuscripts*
- Semi-Digital
  - *Books, texts, images, video*
- Quantitative and Qualitative
  - *Describing, searching and finding useful qualitative data is hard*

Geo, Bio, Chem,
HEP, ESS, Astro,
Ling, SS, Fin

Processing

- Data fusion
  - *Single or multiple repositories*
- Data slicing, latitudinal searches
  - *Impacts technology choices*
- Interfaces to
  - *computing,*
  - *collaboration,*
  - *visualisation*

Summary

- Common and Uncommon Issues from Diverse Application Areas
- One size (infrastructure) does not fit all (yet)
  - But 3-4 sizes may fit most (for now)
- Some domains have very different definitions of sustainability, rights issues, data movement needs
- User and developer education is still needed