2016 National Research Infrastructure Roadmap
Capability Issues Paper

Name	**A/Prof Shawn Ross, Dr Brian Ballsun-Stanton, Dr Adela Sobotkova**
Title/role	**Director/Technical Director/Deputy Director FAIMS Project**
Organisation	**Field Acquired Information Management Systems (FAIMS) Project, Macquarie University (a NeCTAR- , ARC-, and NSW Research Attraction and Acceleration Program-funded e-research infrastructure project.**

Question 1: Are there other capability areas that should be considered? (Parts of this section were also included in Macquarie's response.)

Discovering already existing data. An endemic problem observed at many research institutions across the country is: data exists in desk drawers and file cabinets and in CDs in the back of theses. Bringing this data online represents a huge wealth of research which has already been completed -- but is not usable in its present form. While making new data for research available, knowing what we have and organising it into a linkable, sharable, and extendable format is something which promises remarkable gains and which much of the expensive discovery work has already been completed.

We should also promote methodological replication and data sharing. Exploring, participating in, or creating our own pilot all sciences reproducibility project in the vein of the NWO's "Replication Studies pilot programme"1 could provide for a wonderful "Hallmark national laboratory"2 providing an incentive system for sound methodologies and analytical methods.

Question 5: Should research workforce skills be considered a research infrastructure issue? (Parts of this section were also included in Macquarie's response.)

Absolutely. Investing into hardware and software and "hoping that if you build it they will come" ignores the core problem with the introduction of innovation: the sociotechnic barriers to the adoption of innovation. The diffusion of innovation is not a technical matter, but instead a social process, a special kind of communication3 . For an innovation [new research infrastructure] to be adopted, resources need to be invested into information dissemination, workforce training and upskilling so as to alleviate anxiety/uncertainty arising from novelty and allow for more widespread and therefore

[1] http://www.nwo.nl/en/news-and-events/news/2016/nwo-makes-3-million-available-for-replication-studies-pilot.html

[2] http://www.nature.com/nature/journal/v537/n7618/full/537034b.html

[3] Rogers, Everett M. 2003. Diffusion of innovations. 5th ed. New York: Free Press. http://www.simonandschuster.com.au/books/Diffusion-of-Innovations-5th-Edition/Everett-M-Rogers/9780743222099

more sustainable deployment of new technologies. Lack of investment into the building of workforce skills hinders and generates a long lag in adoption, or leads to non-adoption. Costs of deployment can easily exceed the costs of hardware and software development, but a well managed and funded programme of change [e.g. digital literacy upskilling] will help achieve better utilisation of research infrastructure, and produce a world-class workforce.

Main points:

1. Building a skilled [digitally literate] workforce (and not just a few researchers) that can take advantage of national eresearch infrastructure should be a national priority
2. Building digital literacy should be a joint responsibility of projects involved in eresearch, and educational institutions, but it requires coordination and integration that might be best achieved at national level [standards, guidelines, and funding].
3. Separate budget line is needed for outreach, training and upskilling of workforce in digital literacy and technologies, so that educational institutions and relevant eresearch projects, can develop and deliver training programmes. Such budget lines are currently not prioritized in research infrastructure or pure research funding, leading to a glaring gap between the few technically competent specialists and that majority who are unable to utilize available eresearch resources.

Question 7: What responsibility should research institutions have in supporting the development of infrastructure ready researchers and technical specialists? (Parts of this section were also included in Macquarie's response.)

What responsibility do institutions have to researchers?

With regards to institutional responsibilities to researchers, those responsibilities are significant and not fulfilled. Wide gaps exist between disciplines and between researchers in disciplines; competence in an arcane and complex software sometimes accompanies ignorance of basic data management, or a profound lack of interest in data modelling. Without solid data planning skills and broad-based basic technological awareness Australian researchers are subject to claims of poor research practices. This lack of training stretches from undergraduate classrooms to senior researchers -- not simply a lack of knowledge of how to code (which is excusable in most study areas), but a lack of knowledge of data lifecycle and of eresearch infrastructure, constraining their ability both to assess a particular digital solution, or apply it to a problem.

Propositions:

1. Require a long-term data management plan to avoid a time tax on researchers and research data
2. Discourage boilerplate eresearch policy from all institutions. Require answers to:
    a. How do their infrastructure resources specifically support research goals,
    b. What internal and external resources for data management are provided and how accessible they are to researchers
    c. Articulate how their choices will keep research data available and usable in 20 years.

d. How will continuity in expertise, personnel and equipment be provided over time (say a 5 year timeframe).
e. How will basic data literacy, digital competence and good data practice be promoted within research community?

Question 24: Are the identified emerging directions and research infrastructure capabilities for Understanding Cultures and Communities right? Are there any missing or additional needed? (Parts of this section were also included in Macquarie's response.)

Digital humanities

It is important to note that the Digital humanities are not simply textual analysis of documents, but any computer-enabled research out of the humanities, arts, and some of the social sciences. Supporting computer-assisted humanities research beyond textual analysis is of critical importance, e.g., regarding cultural heritage and archaeology. Other arts and humanities disciplines (aside from literature, linguistics, and archaeology) have particularly underdeveloped e-research infrastructure (and practice by researchers), offering a major opportunity for improvement.

With regards to textual analysis, it is critical to ensure that large-document harvesting is not criminalised to the degree it is in the United States. While intellectual property protections are important, the effective illegalisation of textual analysis on large, available, corpora is hurting research efforts in the textual "digital humanities". The UK offers a better model for balancing IP and research needs.

'Small sciences' problems

Archaeology, anthropology, sociology, linguistics, oral history, and other disciplines require infrastructure (and ongoing support of that infrastructure) for the collection, management, and dissemination of field-acquired data, and thus have much in common with some field sciences (geosciences, biology, ecology, etc.).

Disciplines in this cluster suffer from 'small science' problems of diverse, heterogenous data. 'Small data' has its own set of challenges and opportunities4, but is often overlooked since 'big data' analyses has been fashionable for some time. 'Understanding Cultures and Communities', outside of economics and few other disciplines, mostly deals with this sort of 'small data' - or the only path to 'big data' is the amalgamation of many such 'small data' datasets. As noted above, field sciences face similar challenges, as was highlighted in a Science editorial in March 20165. The sort of infrastructure needed for 'small data' includes access to or training in data modelling (since each project has to model their data), software that is built for research but generalised and customisable, active and archival storage accommodating diverse data, data curation services, and domain-specific repositories. Federated architectures are generally to be preferred, to allow evolution of specific tools and provide sufficient flexibility, while also maintaining data interoperability. National support for linked open data datasets within open government initiatives can help

4 Kansa, E. C., & Bissell, A. (2010). Web syndication approaches for sharing primary data in "small science" domains. DSJ, 9, 42-53. https://www.jstage.jst.go.jp/article/dsj/9/0/9_009-012/_article

5 McNutt, M., Lehnert, K., ... & King, J. L. (2016). Liberating field science samples and data. Science, 351(6277), 1024-1026. http://science.sciencemag.org/content/351/6277/1024.full

provide the infrastructural basis for ontological mapping of small-science terms and datasets.

State cultural heritage registers

With regards to: “The ability to compare, contrast, manipulate, link and integrate the holdings of national and state institutions, particularly via digital technologies, enables researchers, regardless of their physical location, to conduct research on national cultural holdings, “ Data potentially useful - even critical - to historical and archaeological research that is held by state agencies is often difficult to access and impossible to reuse effectively. [redacted]

Future role of data institutions

'Future role of cultural and data institutions' section is very static; it speaks of existing collections, but archaeology and cultural heritage, for example, are only now beginning to generate significant digital datasets, so data from these fields are not in anyone’s remit. Our own e-research infrastructure project had to discontinue a domain-specific repository in Australian archaeology, and merge its contents into an American repository, because long-term data curation infrastructure and services were not available in Australia at a price that our research community could afford (see Appendix A, where we reproduce our letter to contributors, for a fuller explanation).

Question 25: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond? (Parts of this section were also included in Macquarie’s response.)

For archaeology and cultural heritage, Australia needs to look at the Archaeology Data Service in the UK as a model[6]. Should that prove too ambitious, the DINAA project in the US offers a model for sharing data housed in multiple state registers[7]. Part of the problem of unavailable and poor-quality data in the state registers is infrastructure, but part of it is also policy. Ownership of this data, licensing, and reuse terms and conditions must be developed and publicised, taking into account the sensitive nature of some data (especially but not exclusively Indigenous archaeology / heritage data) - but that problem has been solved by other jurisdictions.

Question 26: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Understanding Cultures and Communities capability area? (Parts of this section were also included in Macquarie’s response.)

With regards to: “National and state cultural collecting institutions are a vital set of national research infrastructure to researchers”, it is critical to allow these institutions to expand their remit. For example, the state library of NSW should be encouraged and funded not only to capture topics of interest to NSW, but the outputs of researchers and research projects headquartered in NSW.

---

6 Beagrie, N., & Houghton, J. (2013). The value and impact of the Archaeology Data Service: A study and methods for enhancing sustainability. Charles Beagrie Ltd: Salisbury. http://archaeologydataservice.ac.uk/blog/2013/11/the-value-and-impact-of-the-archaeology-data-service-final-report/

7 http://ux.opencontext.org/archaeology-site-data/

We especially applaud and commend: “Research infrastructure-like activities currently undertaken at national cultural institutions need to be supported and recognised as core national infrastructure, as important as any other research infrastructure holding, and just as irreplaceable.” This phrase is entirely accurate.

Question 27: Are the identified emerging directions and research infrastructure capabilities for National Security right? Are there any missing or additional needed?

It is important not to apply Wassenaar style arms-export regulations to computer security research, nor should we restrict crypto research (see the discussion in the US about arms control on 0days and the “make a new math that agrees with what we say” arguments on crypto.

Appendix A:

The NeCTAR- and ARC-funded FAIMS e-research infrastructure project had to discontinue a domain-specific repository in Australian archaeology, and merge its contents into an American repository, because long-term data curation infrastructure and services were not available in Australia at a price that our research community could afford. This is a letter we sent to our contributors on that occasion. We believe that the roadmap committee may use this letter to highlight existing deficiencies and inform the 'future role of data institutions' for Understanding Cultures and Communities.

Migration of the FAIMS Repository to tDAR

In May 2016, the FAIMS Project entered into a partnership with the Center for Digital Antiquity at Arizona State University in order to transfer all resources in the FAIMS Repository into The Digital Archaeological Record (tDAR).

Background

The FAIMS Repository was established in 2013 as part of the National eResearch Collaboration Tools and Resources (NeCTAR) program to store data sets, documents, images, and sensory data produced by archaeological research in Australia or collected by Australian archaeologists working abroad. The primary focus of this first phase of the FAIMS project was the creation of an Android mobile application for archaeological field recording, and much of the development of the Repository was concerned with automating the ingest of data created on the mobile app. The Repository also subsumed the Australian Historical Archaeological Database (AHAD).

The Repository is, essentially, a slightly modified Australian implementation of tDAR ('The Digital Archaeological Record') which is an open-source repository system developed by Digital Antiquity - a not-for-profit organisation at Arizona State University committed to the long-term digital preservation of archaeological data. The FAIMS Repository was developed by VeRSI (later VPAC, then V3), working in collaboration with Digital Antiquity. While some features were never implemented, the Repository was running live from repo.fedarch.org and monitored in-kind by V3 staff until the closure of the organisation in 2015. The FAIMS team has been managing the repository in-house since then, while exploring options for funding more sustainable system administration.

Why does the FAIMS Repository need a new home?

The short answer is a lack of funds and resources to administer an Australian repository for archaeological data.

The long answer is a range of circumstances has made it more desirable to merge with a larger, international repository, rather than secure the resources needed to sustain an Australian repository. These circumstances include but are not limited to:

1. Development challenges

The initial development of the Repository was hampered by technical difficulties on a number of fronts, including adaptation to NeCTAR, RDSI, and other Australian national infrastructure). Meanwhile, the developers (VeRSI, then VPAC, then V3) went through two major restructures doing development, eventually ceasing operation in 2015.

2. Limited cost-effective infrastructure

The tDAR software is complex, highly customised to the task of long-term data archiving. Few local providers have the expertise or inclination to learn the system and take on system administration at a cost the Australian archaeological community can fund from ingest fees. Digital Antiquity knows the software better than anyone, but cannot practically take on the administration of Australian-based servers.

Furthermore, appropriate data curation infrastructure (essentially, organisations that are willing to guarantee the survival and availability of data over the long term - say, a 100-year horizon) are lacking in Australia at any price. We have approached several libraries and museums, the sort of organisations who often play that role in North America or Europe, but none were interested in data curation that included datasets beyond their relatively narrow research and collection policies. We have raised this broader problem with ANDS, NeCTAR, and AARNET, who have acknowledged the need for curation services, but do not offer any short-term solutions.

3. Lack of funding streams for infrastructure of this kind

While we benefited from generous NeCTAR and ARC grants to build the FAIMS infrastructure, both explicitly excluded expenditure on maintenance and operation. This is typical of many grants which fund infrastructure: tertiary institutions are expected to run and maintain such infrastructure in perpetuity from central funds. This arrangement is unsustainable in today's funding environment.

4. Insufficient demand

While the Repository was running live from repo.fedarch.org, it attracted lots of curious users - but few depositors. This may in part reflect the inevitable a lag in data deposition which often comes a year or two after fieldwork, or more likely the lesser urgency for deposition vs data collection, but it has proved a difficult to justify funding bids with such low rates of adoption.

5. Economy of Scale

By returning the records to the tDAR main branch, we are preserving a similar functionality that users have come to know, taking advantage of newer features not available in the FAIMS software, and securing preservation of the data in the long-term in the hands of a much better resourced staff focussed solely on digital preservation. This enables the FAIMS team to focus on enhancing digital data collection services.

6. Choice for archaeologists

As we have gained experience with data archiving and publication, it has become clear that different datasets need different homes. Other repositories and data publication services have also matured and improved in the time since we began developing the FAIMS Repository. In future, FAIMS will act as a broker for various data services. While we expect tDAR to remain an excellent choice for many archaeologists, mukurtu.org (for example) might work better for culturally sensitive datasets requiring sophisticated access controls, whereas opencontext.org might be better suited to highly structured digital datasets with no access restrictions.

Repository Migration: At a glance

Benefits

- Storage of data in an internationally recognised repository.
- Continuity of basic software functionality and core terminology (i.e. Resource Definitions).
- Automatic application of all tDAR upgrades at no additional cost to the project or users.
- Access to better support from the professional service team that developed the software.
- Potential for international comparison of datasets.

Drawbacks

- There may be delays in corresponding with Digital Antiquity (based in Arizona) on account of the time difference (response times are still likely to be shorter than the FAIMS team can currently guarantee).
- US pricing: Australian customers will be subject to currency fluctuations.
- Loss of some minor features and customised terminology introduced into the FAIMS Repository (but we're working with tDAR to extend their terminology).

Non-issues

- Transferring from open-source to open-source.
- One-off fees are charged at ingest not download (no change; fees were planned for the FAIMS Repo beginning in 2016).

What does this mean for Users?

For users who contributed records to FAIMS:

- A new account will be created for you (this was necessary to transfer the records) but you need to log in and agree to the terms and conditions (which you may note are very similar to the FAIMS Repo).
- Records created in FAIMS will now appear at tdar.org (NB no draft records will be migrated); repo.faims.org will redirect to tdar.org.
- Project and resource numbers have changed, but you can still search for the 'old' numbers (see FAQs below).
- For users who registered with FAIMS but did not contribute:
- You will need to create a new account with tDAR if you do not already have one.

FAQs

Where did my record go?

The FAIMS Repository has been taken down as a precautionary measure prior to the migration and following an unsuccessful security update that may have left the site vulnerable to attack.

What will it cost?

The transfers of records is being funded by Macquarie University and there will be no cost to Users. New records deposited into tDAR will be subject to Digital Antiquity's one-off ingest fees. See tDAR Pricing for details.

Will the FAIMS team deposit records for me into tDAR?

No. The FAIMS Repository was designed as a self-service facility and the FAIMS team has never had the capacity to undertake data ingest. Digital Antiquity, however, do provide fee-for-service assistance to Users. See tDAR Pricing for details. Once the transfer to tDAR is complete, the FAIMS team will have no direct role in the management of Repository services.

What is the FAIMS Collection in tDAR?

The 'FAIMS Collection' includes all records and resources in tDAR that have been imported from, or linked to, FAIMS. Initially this collection will include all records transferred from the FAIMS Repository, but in future may include resources uploaded by Australian researchers or researchers working in Australia, who wish to make their data discoverable as part of the FAIMS collection, thus providing a way to designate Australian (or Australian-produced) data in tDAR.