Submission
2016 National Research Infrastructure Roadmap
Capability Issues Paper

Name	**Robert C. Williamson**[1]
Title/role	**Professor of Computer Science / Chief Scientist**
Organisation	**ANU / DATA61**

Questions

Question 1: Are there other capability areas that should be considered?

Question 2: Are these governance characteristics appropriate and are there other factors that should be considered for optimal governance for national research infrastructure.

The principles seem fine. But what is missing is any suggestion that evidence will be examined regarding the best ways to govern such initiatives. It may be worthwhile trying to learn from the problems that plague other sorts of infrastructure[2].

Question 3: Should national research infrastructure investment assist with access to international facilities?

Question 4: What are the conditions or scenarios where access to international facilities should be prioritised over developing national facilities?

Question 5: Should research workforce skills be considered a research infrastructure issue?

If you define "infrastructure" appropriately as that which other things depend upon and build upon (your current definition is very poor), then yes, of course.

Question 6: How can national research infrastructure assist in training and skills development?

Question 7: What responsibility should research institutions have in supporting the development of infrastructure ready researchers and technical specialists?

Question 8: What principles should be applied for access to national research infrastructure, and are there situations when these should not apply?

---

[1] Although I am chief scientist for DATA61 (a CSIRO business unit), this submission is on my own behalf wearing my ANU hat only (I speak here for neither institution, although the submission is closely aligned with the Data61 science vision that I can speak for). I am aware that CSIRO is making a single (institutional) submission to which I have contributed.

[2] See

- Bent Flyvbjerg, Survival of the unfittest: why the worst infrastructure gets built – and what we can do about it, *Oxford Review of Economic Policy* 25(3), 344-367 (2009)
- David Ribes and Thomas A. Finholt, The Long Now of Technology Infrastructure: Articulating Tensions in Development, *Journal of the Association for Information Systems* 10 (special issue), 375-398 (May 2009)
- Ola Henfridsson and Bendik Bygstad. The Generative Mechanisms of Digital Infrastructure Evolution. *MIS Quarterly* 37.3, 907-931, (2013)

The assertion made in 3.5 "Broad accessibility … maximises the value of the Government's investment in these facilities" does not stand up to logical analysis. Consider two classes of user (A & B). Class A generates $2 to the economy for every $1 of infrastructure used; Class B generates $10. If one wanted to "maximise the value…" one would disallow class A, and only allow class B.

More importantly, the principles need to reflect the empirical reality of how infrastructures are used and evolved; see comments at the end of this submission.

Question 9: What should the criteria and funding arrangements for defunding or decommissioning look like?

Question 10: What financing models should the Government consider to support investment in national research infrastructure?

Question 11: When should capabilities be expected to address standard and accreditation requirements?

This is complex – standards can either help or hinder[3]. Too early adoption leads to lock-in. Too few standards lead to lack of sharing. Central policy should focus on the end goal (interoperability[4]) and encourage many small standards that are lightweight and interoperable (rather than a single massive standard, which would like be ignored[5]), and particularly encourage the development of standard "gateways" (akin to gateway technologies[6]).

Question 12: Are there international or global models that represent best practice for national research infrastructure that could be considered?

I do not believe there are great models. There is lots of good scholarship on the question of infrastructure in general[7], and information infrastructure in particular[8] that it would be a big

---

[3] See

- Ole Hanseth, Eric Monteiro and Morten Hatling, Developing Information Infrastructure: The Tension Between Standardization and Flexibility, *Science Technology Human Values*; 21, 407-426 (1996)
- Section 7.7 of Robert Williamson *et al*., *Technology and Australia's Future*, ACOLA, (2015)

[4] John G. Palfrey and Urs Gasser, *Interop: The Promise and Perils of Highly Interconnected Systems,* Basic Books, (2012)

[5] See the discussion of the lack of use of ecological metadata standards on p 112 of Christine L. Borgman, *Big Data, Little Data, No Data: Scholarship in the Networked World*, MIT Press (2015)

[6] Paul A. David and Julie Ann Bunn, The Economics of Gateway Technologies and Network Evolution: Lessons from Electricity Supply History, *Information Economics and Policy* 3, 165-202 (1988)

[7] Most famously: Susan Leigh Star, The Ethnography of Infrastructure, *The American Behavioural Scientist,* 43(3), 377-391 (1999)

[8] See for example:

- Susan Leigh Star and Karen Ruhleder. Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information systems research* 7.1, 111-134, (1996)
- Paul N. Edwards *et al. Knowledge Infrastructures: Intellectual Frameworks and Research Challenges*, Report of a workshop sponsored by the National Science Foundation and the Sloan Foundation University of Michigan School of Information, (25-28 May 2012)
- Paul N. Edwards, Steven J. Jackson, Geoffery C. Bowker and Cory Knobel, *Understanding Infrastructure: Dynamics, Tensions, and Design*, Report of a Workshop on "History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures" (January 2007)
- Paul N. Edwards, *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming*, MIT Press, (2010) (especially chapter 11, 'Data Wars')

mistake to ignore. It would be worthwhile investing in empirical (ethnographic) research to better understand what design decisions in building research data infrastructure actually lead to wide uptake[9] to avoid a repetition of current programs whose "dirty little secret" is that "not much sharing may be taking place"[10]. In particular, there is a fundamental tension between "control and drift" in large information infrastructures, which if denied causes grief, but if embraced can assure vibrancy[11].

Question 13: In considering whole of life investment including decommissioning or defunding for national research infrastructure are there examples domestic or international that should be examined?

Question 14: Are there alternative financing options, including international models that the Government could consider to support investment in national research infrastructure?

Health and Medical Sciences

Question 15: Are the identified emerging directions and research infrastructure capabilities for Health and Medical Sciences right? Are there any missing or additional needed?

Question 16: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Question 17: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Health and Medical Sciences capability area?

- Geoffery C. Bowker, Karen Baker, Florence Millerand and David Ribes, Toward Information infrastructure studies: ways of knowing in a networked environment, in J. Hunsinger et al. (eds.), *International Handbook of Internet Research*, Springer Science+Business Media B.V. (2010)
- Christine L. Borgman, *Scholarship in the Digital Age: Information, Infrastructure, and the Internet*, MIT Press (2007)

[9] Along the lines of work such as:

- Ann S. Zimmerman, New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data, *Science, Technology and Human Values* 33(5), 631-652 (2008)
- Christine L. Borgman, *Big Data, Little Data, No Data: Scholarship in the Networked World*, MIT Press (2015)
- Christine L. Borgman, Peter T. Darch, Ashley E. Sands, Irene V. Pasquetto, Milena S. Golshan, Jillian C. Wallis, and Sharon Traweek, Knowledge infrastructures in science: data, diversity, and digital libraries. *International Journal on Digital Libraries*, *16*(3-4), 207-227, (2015)
- Victoria Stodden, *The scientific method in practice: reproducibility in the computational sciences*, MIT Sloan School Working paper 4773-10, (February 2010)
- Rob Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures and their Consequences,* Sage (2014)

[10] Christine L. Borgman, The conundrum of sharing research data, *Journal of the American Society for Information Science and Technology* 63(6), 1059-1078 (2012)

[11] See:

- Claudio U. Ciborra et al., *From Control to Drift: The Dynamics of Corporate Information Infrastructures*, Oxford University press, (2000)
- Antonio Cordella, Information Infrastructure: An Actor-Network Perspective, *Social Influences on Information and Communication Technology Innovations* 20, (2012)

The claim (section 5.1.5) that "all of the emerging capability directions would benefit from the implementation of an Australia-wide research quality management system infrastructure…" is merely stated with no proof. It could just as easily produce a net dis-benefit.

Environment and Natural Resource Management

Question 18: Are the identified emerging directions and research infrastructure capabilities for Environment and Natural Resource Management right? Are there any missing or additional needed?

Question 19: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Question 20: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Environment and Natural Resource Management capability area?

Advanced Physics, Chemistry, Mathematics and Materials

Question 21: Are the identified emerging directions and research infrastructure capabilities for Advanced Physics, Chemistry, Mathematics and Materials right? Are there any missing or additional needed?

Question 22: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Question 23: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Advanced Physics, Chemistry, Mathematics and Materials capability area?

Understanding Cultures and Communities

Question 24: Are the identified emerging directions and research infrastructure capabilities for Understanding Cultures and Communities right? Are there any missing or additional needed?

Question 25: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

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?

National Security

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

Question 28: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Question 29: Is there anything else that needs to be included or considered in the 2016 Roadmap for the National Security capability area?

Underpinning Research Infrastructure

Question 30: Are the identified emerging directions and research infrastructure capabilities for Underpinning Research Infrastructure right? Are there any missing or additional needed?

The discussion of High Performance Computing in the issues paper is silent on the rapid commoditisation of HPC. Ed Lazowska's analysis[12] that much science that previously requested, and got, their own specialised HPC infrastructure, is better served by the commercial cloud is extremely pertinent in this regard. This could save substantial sums of money!

Question 31: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Question 32: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Underpinning Research Infrastructure capability area?

Data for Research and Discoverability

Question 33 Are the identified emerging directions and research infrastructure capabilities for Data for Research and Discoverability right? Are there any missing or additional needed?

Question 34: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Question 35: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Data for Research and Discoverability capability area?

The stated goal of the exercise (the first bullet point of section 2.1) combined with

- the statement on page 7 "data for research and discoverability have been identified as pervasive across the research system";
- "as you might expect, data is everywhere…" (page 12);
- the table on page 55 where data infrastructure underpins every national research priority;
- and the fact that sections 5, 6 and 8 listed the infrastructure needed for dealing with data *first* in their list of emerging directions;

suggests strongly that infrastructure for data is indeed paramount. Data infrastructure can be viewed as infrastructure for (other) research infrastructure. Precisely because it can be used so widely, it is reasonable to expect the payoff for a $ invested here will be greater than a $ invested in any other area. This needs calling out since, because of its pervasiveness, digital infrastructure can be take for granted even more than other types – "all these funded projects have a data component, so we do not need to separately fund an initiative in developing ground breaking data infrastructure for scientific discovery" (paraphrase of a comment by the leader of a well-known research

[12] Ed Lazowska, *A Plea for Greater Attention to Data-Intensive Discovery, Greater Investment in Intellectual and Software Infrastructure, and Greater Use of the Commercial Cloud*, Remarks to the CSTB Committee on Future Directions for NSF Advanced Computing Infrastructure to Support US Science in 2017-2020 http://lazowska.cs.washington.edu/CSTB.pdf (December 2014)

organisation). This problem of the "unseen labour" (the lack of understanding of how much work is involved) behind digital infrastructure extends beyond research data infrastructure[13]. The funding of this infrastructure is as a consequence particularly problematic[14] - it is useful to *everyone*, but *nobody* wants to pay. For that very reason, this would seem a most suitable area for direct central government funding.

Improved research data infrastructure can lift the game of all *research done in Australia.* It is not merely a matter of making an archive of collected data, but it can be underpinning infrastructure for all the work that researchers do.

Infrastructure (as shared resources) is not merely data centres and large computers. As Ed Lazowska (op cit.) argued, investment in software can be far more valuable than building physical infrastructure.

Section 10.3.5 (geospatial data) of the issues paper says very little. It would be valuable for the committee to consider the successful example of nationalmap, whose *federated* model with a focus on no heavyweight infrastructure in the traditional sense (there is literally no back-end to nationalmap – it is "merely" some code that is executed in the user's browser) shows the rapid and large impact possible by an agile software driven approach to infrastructure. This approach will underpin the design and construction of the new data.gov.au through the NISA funding provided to DATA61.

Re the digitisation of physical specimens, I make the obvious point that the investment should be in the infrastructure that *enables* the widespread digitisation of specimens, rather than funding the digitisation itself. That is, invest in the robotics, analytics and data management to help scientists do the requisite digitisation.

It might be helpful to distinguish between "data" and "capta": Rather than being *given* to us ("data" comes from the Latin *dare* meaning "to give"), it is necessary to *take* the data – to actively select and gather it, and then, of course, to *do something* with it. It is useful to distinguish data from "capta" [15] (from the Latin *capere* meaning "to take, seize, obtain, get, enjoy or reap"[16]). This terminology signals that data collection is an active process, not passive. Most of the points in the paper about

---

[13] Nadia Eghbal, *Roads and Bridges: The unseen labor behind our digital infrastructure*, Ford Foundation, https://www.fordfoundation.org/library/reports-and-studies/roads-and-bridges-the-unseen-labor-behind-our-digital-infrastructure/ (2016)

[14] See *pp*57*ff* of Rob Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures and their Consequences,* Sage (2014)

[15] This distinction is quite old, but rarely used. See
Rob Kitchin, *The Data Revolution: Big data, open data, data infrastructures and their consequences,* Sage, Los Angeles (2014); this explains some of the history of the word.
Christopher Chippindale, Capta and data: on the true nature of archaeological information, *American Antiquity* 65(4), 605-612 (2000)
Bettina Berendt, *Big Capta, Bad Science? On two recent books on "Big Data" and its revolutionary potential*, Department of Computer Science, KU Leuven, https://people.cs.kuleuven.be/~bettina.berendt/Reviews/BigData.pdf (March 2015)

[16] Entry for *captus* in (*A Latin Dictionary. Founded on Andrews' edition of Freund's Latin dictionary. revised, enlarged, and in great part rewritten by. Charlton T. Lewis, Ph.D. and. Charles Short, LL.D.* Oxford. Clarendon Press. (1879))

"data" are really about "capta". The distinction is not pointless pedantry – it makes it clear that you have to choose what capta to collect.

One other point: much infrastructure for data is built with software. Software's incorporeality means it can be changed *much* faster than any hardware. Thus any governance structures suitable for managing hardware infrastructure are unlikely to be suitable for governing the evolution of research data infrastructure.

Finally, and perhaps most importantly, I believe that the time is ripe for a more ambitious and more readily executed (and more widely impactful) vision of data infrastructure for research. Rather than individual end user areas trying to develop this in an amateurish fashion, and rather than making the mistake of building an edifice and expecting people to come to it, something different is called for.

The vision below is being developed right now as a project lead by DATA61 (Project Mnemosyne[17]). The idea is to look at the entire lifecycle of data in the scientific enterprise – from its initial capture to ultimate use. This is certainly not a simple linear process, and is highly heterogeneous across different disciplines, and the requisite infrastructure needs to support this. In order to achieve the overarching goal, the following problems need to be deal with simultaneously:

- **Provenance, trust and reliability** – managing the provenance of both data, and the transformations that the data undergo (including the initial ingestion into pre-defined categories)
- **Management of legal rights** – this needs to be all done by machine by using formal methods of encoding legal rights, so that compositions of data (which are the norm) can have the appropriate legal rights enforced all the way through to use
- **Management of uncertainty –** the so-called age of big-data does not remove the problem of uncertainty; it just makes it worse. Being able to properly propagate uncertainty through the entire life-cycle is currently not possible (due to lack of interoperability)
- **Management of confidentiality –** While confidentiality is but one of many legal rights, it generates particular problems, particularly when data from multiple sources are to be combined. Sophisticated new methods offer the possibility of being able to systematically manage this.
- **Management of complex workflows**, including the ability to rewind what was already done, and the ability to more fluidly move between exploratory and production analytics, sharing workflows across boundaries, and having the workflow capture and maintain all of the other points listed here
- **Late binding ontologies** – data is typically captured for one purpose and organised for that purpose; but much value can accrue from using for other purposes. A system that allows one to rewind all the way back to the initial data categories (the data "ontology" or schema) will facilitate this more radical re-use.
- **Being able to cross organisational, jurisdictional, and technical boundaries –** the only way to build a big system is to build lots of small ones and have them talk to each other. Inevitably this leads to boundary problems. The whole data research infrastructure needs to

[17] Named after the goddess of memory and remembrance, who invented language and words, discovered the uses of the power of reason, and gave a designation to every object.

be conceived of, and designed, with this fluid boundary crossing in mind. It is antithetical to the typical commercial imperative of locking users into one fixed system.

- **Proper decoupling of technique from problem –** Much of the current data technology is still very technique oriented – users are invariably made aware of details of *how* things are being done. An approach that more rigidly separates problem from technique by forcing a declarative approach will also facilitate composition and delivery as a service.
- **The whole infrastructure being generative and user-centered**[18], and being able to be developed piece-by-piece in response to user demand
- **Avoidance of proprietary control –** Because all science depends upon data, the data infrastructure has a particularly strong requirement to remain open and avoid being controlled by any single entity.
- **Learn from lessons of the past**[19] - The notion of linking data, and more generally build cyberinfrastructure for science is far from new. But many of the efforts to date have not had the success original imagined because scientists not utilise the extant research infrastructure[20]. It is thus essential to deeply understand the factors affecting the uptake and use of research data infrastructure so that new infrastructure can be build that *is* used. The evidence suggests that the best approach to this is an agile software methodology where small components are developed on a fast clock cycle so that rapid feedback can be obtained from users.
- **Deliver as a service** – similar to the way much software development is now heading, it is envisaged that all this infrastructure is delivered as a layered[21] service (or micro-service)[22], rather than requiring complex software to be installed on scientist's computers. This service oriented model also opens the opportunity for rich market mechanisms which will likely lead to a rich and vibrant ecosystem.

A system that delivers all of the above does not exist. But I believe it can be built by exploiting the state of the art in technologies for data, from advanced machine learning, to provably correct software through to cryptographic / distributed ledgers and ethnographically grounded UX design methods. This requires substantial research to do, but every aspect above is within reach. Such

---

[18] Confer

- Liv Karen Johanessen, Deede Gammon and Gunnar Ellingsen, Users as Designers of Information Infrastructures and the Role of Generativity, *AIS Transactions on Human-Computer Interaction* 4(2), 72-91 (2012)
- Eric von Hippel, *Democratizing Innovation*, MIT Press (2005)
- Eric von Hippel and Ralph Katz, Shifting innovation to users via toolkits, *Management Science* 48(7), 821-833 (2002)

[19] Kitchin, op. cit.

[20] See:

- Christine L. Borgman, The conundrum of sharing research data, *Journal of the American Society for Information Science and Technology* 63, no. 6 1059-1078 (2012)
- Paul N. Edwards, Matthew S. Mayernik, Archer L. Batcheller, Geoffrey C. Bowker and Christine L. Borgman, Science Friction: Data, Metadata, and Collaboration, *Social Studies of Science* 41, 667-690, (2011)

[21] Re Layered architecture https://www.nginx.com/blog/time-to-move-to-a-four-tier-application-architecture/?utm_source=microservices-at-netflix-architectural-best-practices&utm_medium=blog

[22] Re microservices, see http://martinfowler.com/articles/microservices.html

research can neatly complement, and ultimately underpin the research infrastructure necessary for a substantially better handling of data in science in all of its aspects. A significant side-effect would be that much of the technology necessary for research data infrastructure will be able to be used for diverse (and commercialisable) other purposes without conflicting with the necessary openness and shared nature of infrastructure on which the scientific enterprise depends.

I believe that the tools described above should be conceived of as research infrastructure. I think there are lots of big opportunities to leverage existing assets (such as ANDS) in this regard[23].

Other comments

If you believe that there are issues not addressed in this Issues Paper or the associated questions, please provide your comments under this heading noting the overall 20 page limit of submissions.

The definition of infrastructure at the beginning of section 2.1 is inadequate. The first sentence implies that all parts of every research university in the country is included. A more typical definition would explicate the notion of sharing[24] and interdependency[25]. The second sentence makes no sense – what is meant by 'equally accessible' (which is not the same as sharing)? And even if you take a naïve view of what that means, it is hard to imagine *any* infrastructure equally accessible to *all* users. Figuring out what it is you are talking about would seem a worthwhile goal!

The second bullet point of section 2.1 (increase collaboration) is odd. Collaboration is a *means*, not an *end*. Promoting collaboration for its own sake has no logical basis. (Needless to say, I am not asserting collaboration is a bad thing; it is merely instrumental though.)

The third bullet point "...arising from lack of coordination" makes me worry that the means of coordination might be envisaged to be Soviet style central planning, rather than (the typically more agile and effective) market mechanisms. Very few federal interventions are approved by treasury without evidence of "market failure". I suggest an analysis of the best (as determined by empirical measurement) means of coordination of research infrastructures. I would be concerned if the default was going to be some coordination committee structure.

Finally, I do have a concern that there is no member of the working group with a background in data science and technology. While there is Robyn Owens (a capability expert with background in data analysis), given the centrality and importance of data (as clearly recognised in the issues paper), it seems odd to have such a paucity of expertise in what is arguably *the* most important area of infrastructure for modern science.

I hope these comments (and pointers to literature) are helpful. I am happy to elucidate if needed.

---

[23] This assertion is based on a discussion with Ron Sandland (Chair of ANDS steering committee).
[24] Brett M. Frischmann, *Infrastructure: The social value of shared resources,* Oxford University Press, (2012)
[25] See the definition in Susan Leigh Star, The Ethnography of Infrastructure, *The American Behavioural Scientist,* 43(3), 377-391 (1999)