Research
Data
Culture
Conversation

18 August 2023

Dear Dr Hatherly

We write on behalf of the Research Data Culture Conversation[1], an ongoing activity initiated five years ago by research data and infrastructure managers at Monash University, the University of Melbourne, the University of New South Wales, the University of Queensland and the University of Sydney.

In our submission to the Australian Universities Accord Panel we noted that:

> Institutions, where data ambitions and obligations meet budgets,
> need nationally coherent discipline-sensitive responses to be defined.

The Academy's Decadal Plan for Social Science Research Infrastructure 2023–32 is an example of how a discipline-sensitive response (for social sciences) could emerge. A permanent and enduring improvement response is required. Importantly, this response should consider the impact and alignment it has with the universities - who will inevitably play a significant role in the three areas you have identified:

1) Producing, discovering and accessing data;
2) Analysing data to generate new knowledge; and
3) Brokering high-value partnerships for innovation.

As part of articulating the emerging challenges of managing research data at scale, we have been measuring the growth in research data in the sector. This has initiated several further developments currently in progress.

Our main findings, which relate to all the questions you have listed under "Delivering Solutions", are summarised as appendices to this letter.

We would welcome a discussion on how the infrastructure that institutions will necessarily build and operate, could be arranged to better assist the disciplines of the social sciences.

Yours Faithfully

Ai-Lin Soo, Coordinator RDCC, (contact: ai_lin.soo@unsw.edu.au)
Luc Betbeder-Matibet, Chair RDCC
Rhys Francis, Facilitator RDCC

Appendix I
The first count of Australian Research Data at scale

Abstract for a Practice Paper, to appear at International Data Week, Salzberg, October 2023.
Ai-Lin Soo, Rhys Francis and Luc Betbeder.

The volume of data produced in the world has been estimated by IDC to be growing from 33 zettabytes in 2018 to an expected 175 zettabytes in 2025 [1]. These estimates are widely referenced such as in the European Data Strategy [2].

In that context we report here the practical experience arising from our effort to measure the actual volume and growth rate in Australian research data.

The Research Data Culture Conversation (RDCC) is a partnership of large Australian Research Universities addressing the question 'What is an effective research data culture' [3]. We have found that little is known. We do not know how much data there is, its various properties, where it is, what its total cost of ownership is, or the way in which those characteristics are changing over time.

The first ever practical estimation of the volume of Australian research data under the stewardship of the RDCC members was completed in 2021 [4]. The work was expanded in 2022 to include more Australian universities, Australia's national research infrastructures, its national science agency CSIRO and its medical research institutes [5]. The results suggest a volume of 300 Petabytes (PB) at the end of 2021 and that a growth rate 'doubling every three years' (similar to that projected by IDC) may be occurring in managed research data in Australia.

Because the majority of research expenditure occurs in a small number of institutions, we worked with sixty of the larger research participants and extrapolated to the entire sector using research intensity measures. In the case of universities, our sample set covered two thirds of the total research activity performed by Australian universities.

Two primary learnings arise.

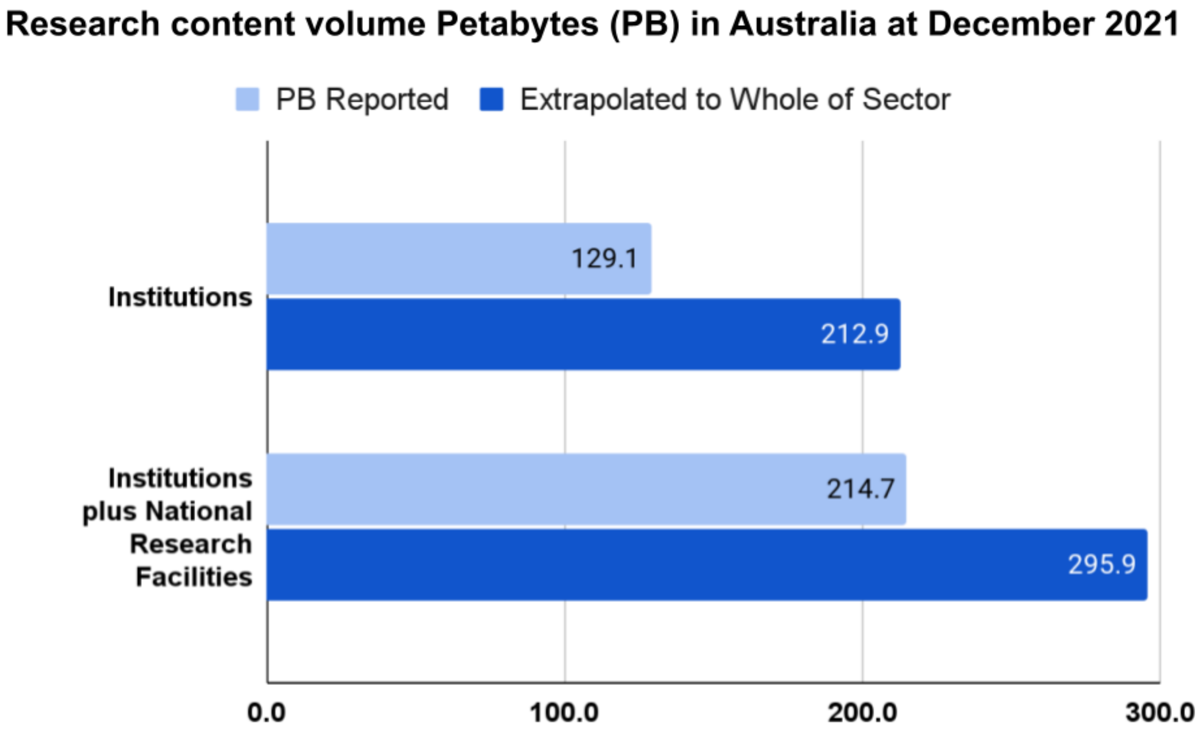
1) The work was made more difficult than anticipated due to the absence of robust internal reporting on the state of the 'research data asset' held within each institution.
2) It proved impossible to measure the characteristics of data and instead we measured proxies in the form of characteristics of the systems supporting the data.

For example, instead of measuring the volume of the unique research data held for future access, our original aim, we found we could only measure the first copy of the total digital corpus under institutional management. Naturally not all of such content is quality research data and some will be copied and counted multiple times.

Similarly, it proved impossible to know the extent of the data from an institution that is openly accessible but it was possible to measure the volume of digital content held in services institutions operate to support open access. We found a total of 86.3 PB of research data held in Australia was openly discoverable, mostly astronomy data and mirrors of overseas reference data and that, by excluding those two categories, institutional open access services contributed 6.6 PB.

For sensitive data, we were able to measure the first copy of the total content held in services intended to meet sensitive data requirements. This produced a volume of 76.5 PB dominated by the medical research institutes. Of course not all sensitive data is in appropriate services and not all the content in sensitive qualified services is in fact sensitive data.

[1] https://www.researchdataculture.org/



We set out to characterise the Research Data Asset created by research in Australia, and managed for future access, producing the estimates outlined above. However, a significant gap exists between the intent and what was realised. An application of this work in Aotearoa New Zealand confirmed the interest in an ability to characterise national research data assets and the challenges involved.

Work to more accurately characterise the two national data assets is now underway. Our first step involves establishing a small set of characteristics of research data to be measured. Given the desire to support Indigenous Data Sovereignty in Aotearoa, and Te Tiriti [6] requirements around active protection of taonga, this will include the properties of Indigenous data. Then a more robust calibration of the Research Data Asset in Australia and in New Zealand will be made at the end of 2023 and again at the end of 2024. The results will then be available to report at IDW 2025.

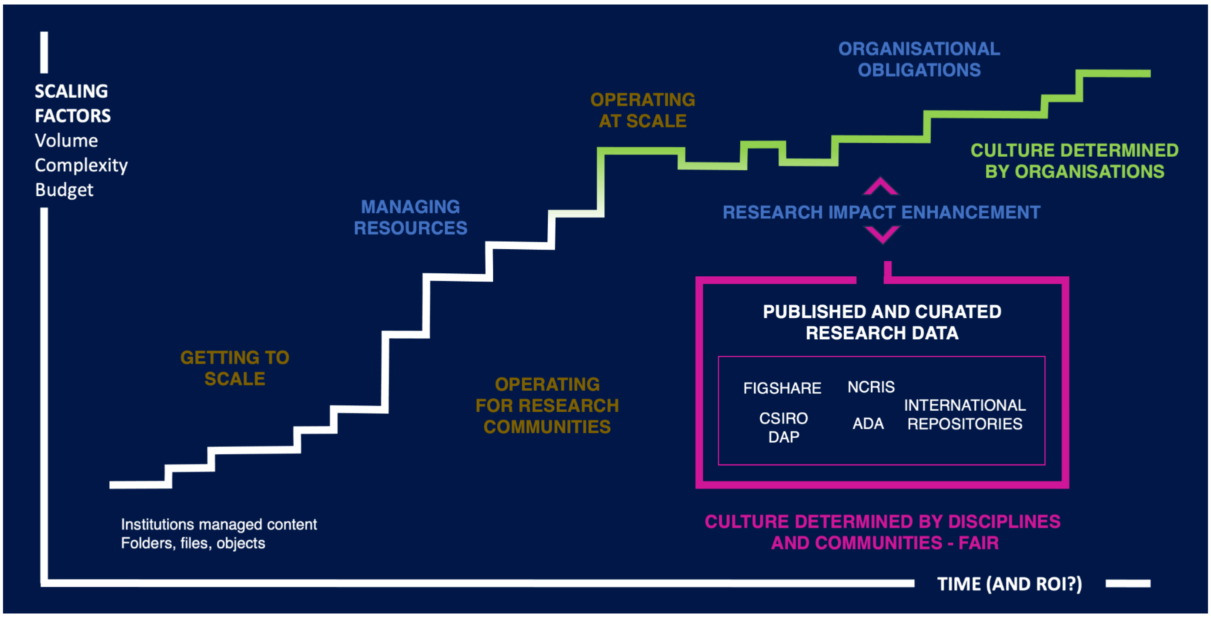
1: IDC 2018
2: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020DC0066
3: https://doi.org/10.26180/20235570
4: For information on the RDCC see https://www.researchdataculture.org/
5: https://doi.org/10.26180/22776320
6: For information on Te Tiriti see https://www.waitangitribunal.govt.nz/

[1] https://www.researchdataculture.org/

Appendix II
The Australian Research Data Culture Conversation

Abstract for a Practice Paper, to appear at International Data Week, Salzberg, October 2023.
Ai-Lin Soo, Rhys Francis and Luc Betbeder.

We have measured the research data volume under management in Australian research institutions [1]. Doing so was made significantly more challenging by the absence of any alignment on definitions for our national research data assets, components of which are retained in each institution. Further, creating this 'Macro View' of the research data assets being retained for future access revealed that institutions were addressing a different challenge. Namely, to permanently retain the unlimited expansion of their uncurated digital corpus within which resides (actual and FAIR) research data.


SCALING FACTORS
Volume
Complexity
Budget
GETTING TO SCALE
MANAGING RESOURCES
OPERATING AT SCALE
ORGANISATIONAL OBLIGATIONS
CULTURE DETERMINED BY ORGANISATIONS
RESEARCH IMPACT ENHANCEMENT
OPERATING FOR RESEARCH COMMUNITIES
PUBLISHED AND CURATED RESEARCH DATA
FIGSHARE
NCRIS
CSIRO DAP
ADA
INTERNATIONAL REPOSITORIES
Institutions managed content
Folders, files, objects
CULTURE DETERMINED BY DISCIPLINES AND COMMUNITIES - FAIR
TIME (AND ROI?)

The most important concept is that institutions are a primary location where data ambitions and obligations meet budgets. To support these data ambitions, institutions need nationally coherent discipline-sensitive responses to be defined for the curation and publishing of research data.

Six years ago, Monash University, one of Australia's largest universities undertook an activity to assess their historical research data growth. The results showed that a 75% compound annual growth rate (CAGR) was experienced by central research storage infrastructure between 2009 to 2015, followed by a period of 40% CAGR between 2017 to 2020 [2]. Two key insights emerged. First, while the reduction is significant, and despite it being closer to the technology improvement curve, an ongoing CAGR of 40% presents a growing curation load which is difficult to sustain under constrained financial conditions. The analysis also revealed that the decrease in CAGR correlated with the introduction of new data management policies that provided an intervention point on researcher's behaviour.

As a result, five of Australia's largest universities (Monash University, The University of Melbourne, The University of New South Wales, The University of Queensland, The University of Sydney) assembled under the banner of the Research Data Culture Conversation (RDCC). The topic was 'what is an effective research data culture for our institutions' and what practices can we develop and share in response to the 'data deluge' we measured [3]. It was understood that technology based solutions may not address key underlying practices of research and hence would not have the same impact as interventions that addressed beliefs, values and

[1] https://www.researchdataculture.org/

practices of research and researchers - ‘the research data culture’.

To further understand an emerging culture that reflected research practices, the RDCC held many national meetings predominantly with Australian universities who also ‘felt the pain’.

Findings in the initial years of conversations were as follows [4].

1. *A strong conversation has formed around Sharing, Preservation and Reuse however, a strong conversation in Sensitivity, Resourcing and End-of-Life is absent.*
   Articulated as the “Yin and Yang” of research data, the RDCC highlighted that many institutional and national investments had been made in progressing Sharing, Preservation and Reuse but less had been made in progressing Sensitivity, Resourcing and End-of-Life decisions, which are also important.

2. *Data lifecycles only exist if data is treated differently at different points in time.*
   The RDCC notes that in order to enable effective interventions, there needs to be a better understanding of the decision points that occur in practice which change the characteristics and interactions with research data.

3. *RDMPs must drive machine-actionable decision making over data life times.*
   RDMP’s as they are currently implemented are not used to understand, make or predict decisions about the future management of research data. As data volumes rise and the effort available to curate research data remains static, a new form of RDMP is needed, that is machine-actionable and more closely tied to university services.

A fourth and more critical observation was uncovered during our latest ‘Macro View’. This latest survey now includes more of Australia’s universities, along with its medical research organisations, national research infrastructure providers and the CSIRO[1].

4. *“Research institutional” (Green space) services and infrastructure are not being designed, managed or resourced to be able to deliver on “Research Community” (Pink space) objectives.*

Understanding the scale, the research ecosystem and characteristics of digital research content under management is essential to deliver effective research data infrastructure design for organisations. A critical problem uncovered during the reporting of the latest Macro View, is represented conceptually as “ Green and Pink Space”. For the RDCC, the Green and Pink space helps to illustrate this key distinction between the objectives and obligations of the (mostly uncurated) digital corpus managed in institutional systems and those (more FAIR) research data managed for research communities. Our observation is that the “Green” (research institutional space) infrastructure is not being designed, managed or resourced to be able to contribute to, interface with or deliver to “Pink” (research community space) objectives.

Our experience is that:

1. A continued conflation between the two is detrimental to the achievement of Pink space outcomes.
2. Institutions, which are a primary location where data ambitions and obligations meet budgets, need nationally coherent discipline-sensitive responses to be defined.

[1] https://doi.org/10.26180/22776320
[2] https://www.researchdataculture.org/macro-view
[3] https://doi.org/10.26180/20235570
[4] https://doi.org/10.5281/zenodo.3887399

[1] https://www.researchdataculture.org/

Appendix III
Submission to Australian Universities Accord Panel

Dear Professor O'Kane,

We write on behalf of the Research Data Culture Conversation[1], an ongoing activity initiated five years ago by research data and infrastructure managers at Monash University, the University of Melbourne, the University of New South Wales, the University of Queensland and the University of Sydney. As part of articulating the emerging challenges of managing research data at scale we have measured the growth in data in the sector.

We believe the Accord can set in motion activities to address the challenges of research data growth, now and over the coming decades, leading to a shared agenda, aligned planning and investment by governments, universities and other research bodies. Improved data management does lead to knowledge creation and collaboration. However, the current unmanaged research data growth creates a long term resource challenge for universities and Governments, including the Government's NCRIS program.

Last year we invited a wide range of research sector institutions to answer the following question:

> "What volume of unique data is being intentionally managed by your institution for the purpose of future access"

The work took approximately six months. Notwithstanding that sixty seven institutions, including universities, medical research institutions, CSIRO and national research facilities, were involved in meetings and discussion, not one was able to answer that question. However, we were able to collect information on the total corpus of digital research content produced through research project activity that is being managed for future access.

As of December 2021 we estimate that the total content managed is at least three hundred petabytes, and we estimate it to be growing at about twenty five percent per year and doubling every three years.

The creation of this corpus was funded by research schemes but at the end of the schemes, and after operational and legal retention requirements are met, our Universities believe they are obliged to continue to retain it, in case the corpus might be valuable. *The result is an unbounded unfunded unending liability.*

We can also report that our survey respondents were unable to respond with the fraction of their holdings that is in fact valuable, what fraction is sensitive, or what part of the whole is original or copies. The observation we make is that while research data is known to be a valuable asset, we are currently unable to report very much at all about it and its properties.

Given that research data is a key component of our national and global stock of knowledge, a far better understanding of it is needed. This understanding belongs in the Accord between universities and the Australian Government because of the role research data plays in the value creation of universities, a role that is expected to continue and grow. Further, the stewardship of a culture and practice that enables the cost efficient value of research data to be realised, for Australia, requires a systemic response.

Yours Faithfully

Ai-Lin Soo, Coordinator RDCC, (contact: ai_lin.soo@unsw.edu.au)
Luc Betbeder-Matibet, Chair RDCC
Rhys Francis, Facilitator RDCC

[1] https://www.researchdataculture.org/