ARDC Data Retention Project:

Building a coherent national storage investment model

eResearch Australasia 2023 October

PRESENTED BY

Max Wilkinson
max.wilkinson@ardc.edu.au

Australian Research Data Commons

Purpose

To provide Australian researchers with competitive advantage through data.

Mission

To accelerate research and innovation by driving excellence in the creation, analysis and retention of high-quality data assets.

NCRIS National Research Infrastructure for Australia
An Australian Government Initiative

ARDC is enabled by NCRIS

*Image — **Framestock - 245023951 / STOCK.ADOBE.COM***

DATA RETENTION

The Data Retention Project partnered with the research sector to increase the impact of investment in data storage infrastructure for significant data collections

Brief

New investment strategy for data storage

Modulate capital investment in storage infrastructure

- Generally driven by technology/scale
- Often separated from data management overheads

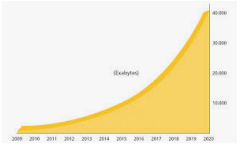
To become more inclusive and equitable

Current Market

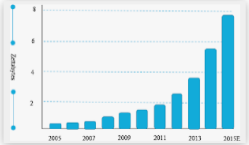
1. **Scale**
    - What data are where
2. **Capability**
    - to manage metadata
3. **Appetite**
    - Incentives for change

The Challenge (part 1 - why)

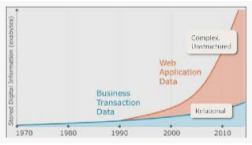
Significant Data Growth



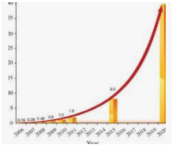
The exponential data growth estimated ...
researchgate.net

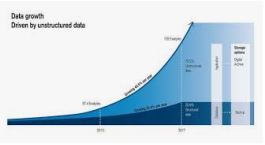


rapid growth rate of data in Zettabytes ...
researchgate.net

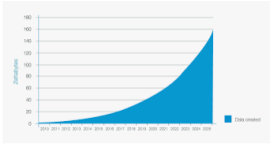


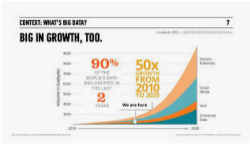
Data growth and expansion (IDC, 2009 ...
researchgate.net


Global growth trend of data volume ...
researchgate.net

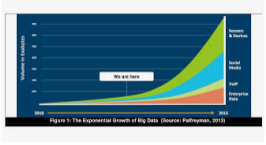


Everything a Data Scientist Should Know ...
kdnuggets.com

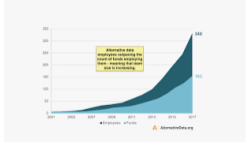

Forecast of exponential growth of ...
reddit.com

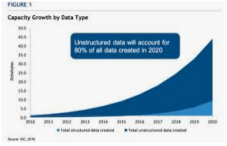


Rise of the Data Warehouse | Avora
avora.com

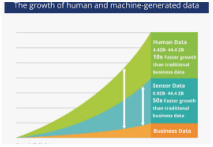


Data Analytics: Concepts, Technologies ...
semanticscholar.org

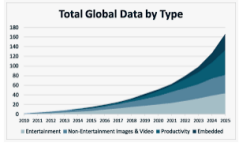

Buy-side Alternative Data Employee ...
alternativedata.org



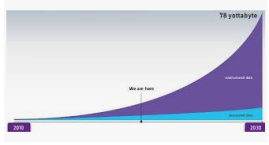
Industry Verticals Tackle Unstructured Data
kevinljackson.blogspot.com



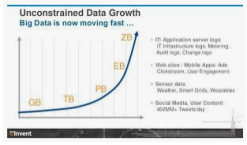
IoT, Big Data and AI – the New ...
business2community.com



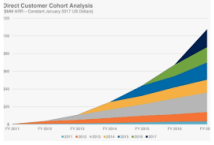
The Data Deluge - Drowning in Data ...
uncommonlogic.com

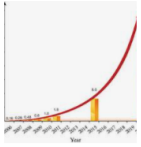


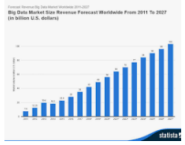
Introduction to BIG DATA: What is ...
bigdatapath.wordpress.com

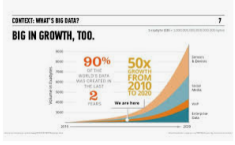


big data growth - Google 搜尋 | Big ...
pinterest.com

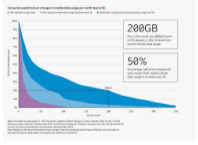

MongoDB: Riding The Data Wave (NASDAQ ...
seekingalpha.com


Global growth trend of data...
researchgate.net

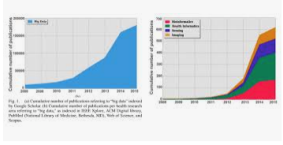

10 Charts That Will Change Your ...
forbes.com

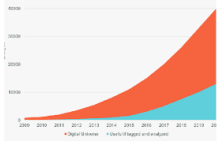


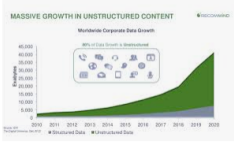
Ensure Business Growth via Big Data ...
promptcloud.com



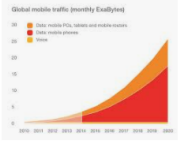
How Much Will 5G Data Usage Increase ...
spectrummattersindeed.blogspot.com


Healthcare Big Data Analytics
healthitanalytics.com

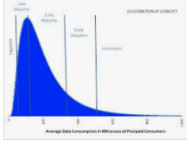

Data growth between 2009 and 2020 ...
researchgate.net



The Massive Growth in Unstructured Data ...
researchgate.net

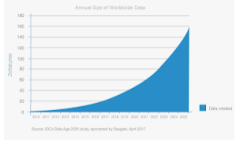


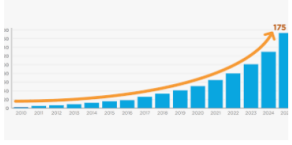
Global Mobile Data Traffic 2010–202...
whatsthebigdata.com


Mobile Data Growth ... The Perfect Stor...
techneconomyblog.com

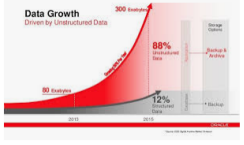
Related searches

- unstructured data growth
- structured vs unstructured data
- infographic data growth

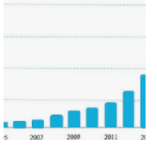

Orchestrating Enterprise Data with Data ...
virtustream.com

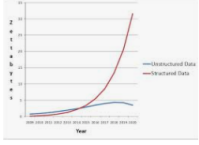


Big data overview | AP CSP (article ...
khanacademy.org



data growth driven by unstructured ...
pinterest.com

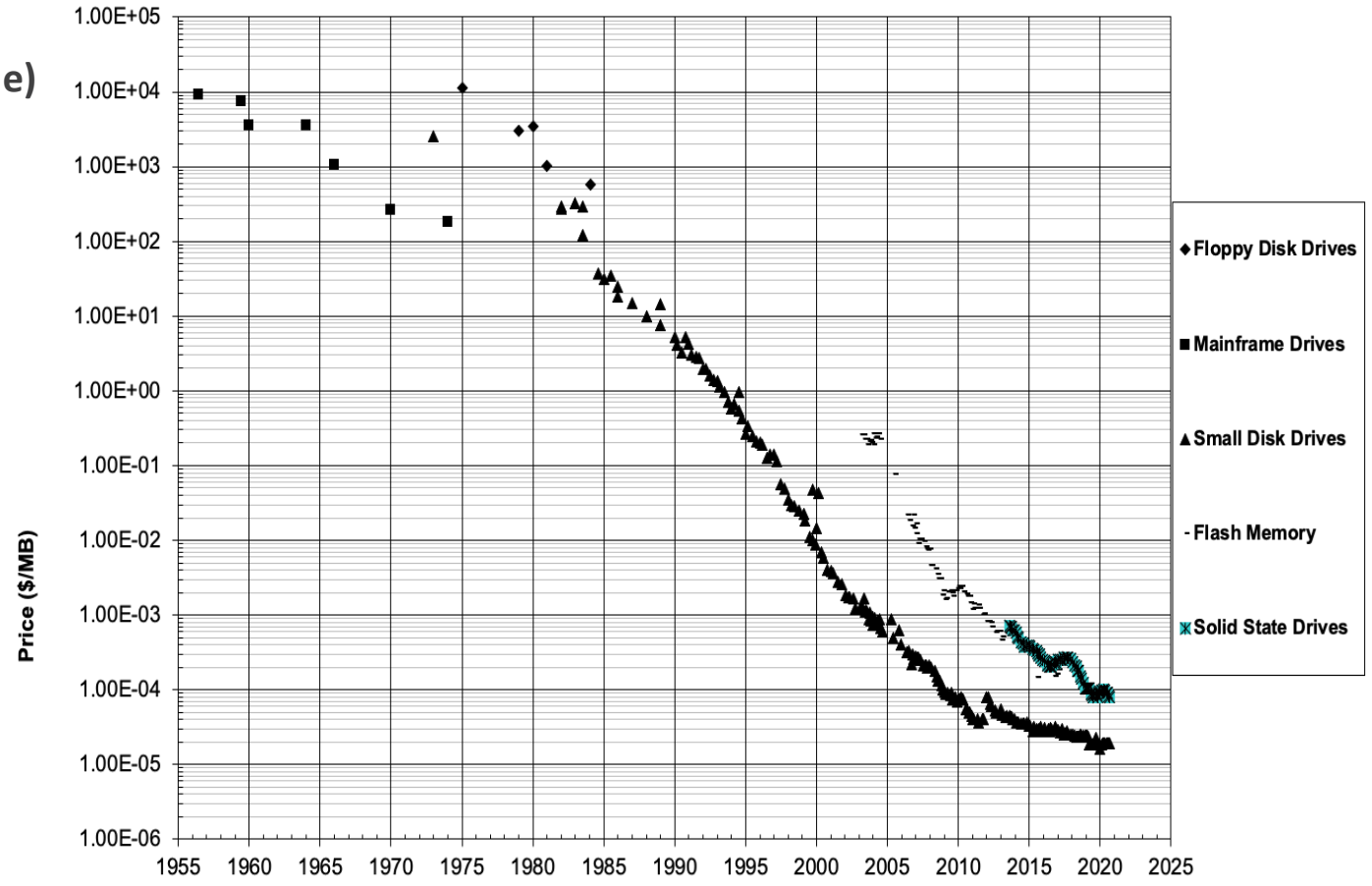

rapid growth rate of data in Z...
researchgate.net



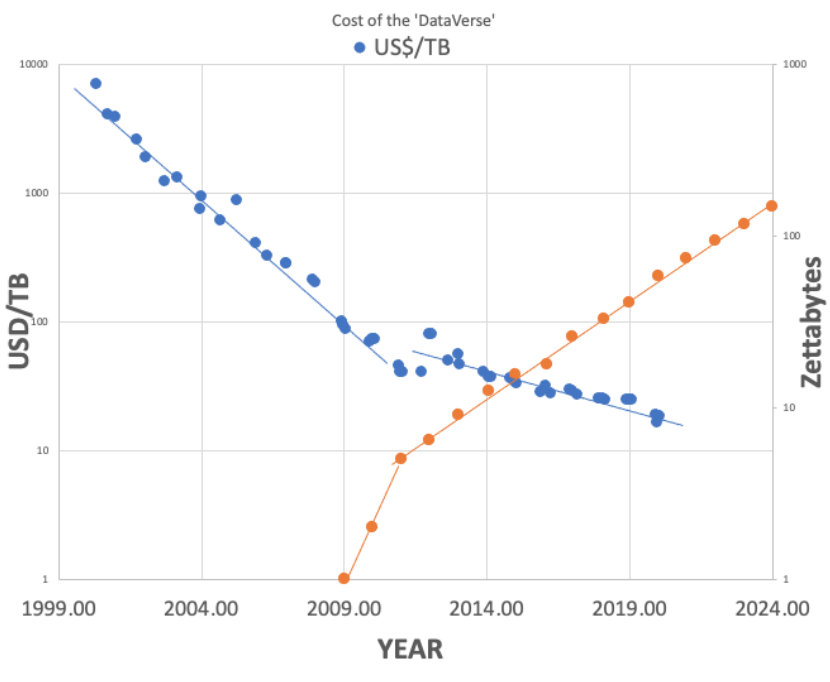
structured vs unstructured data growth ...
tomkendig.wordpress.com

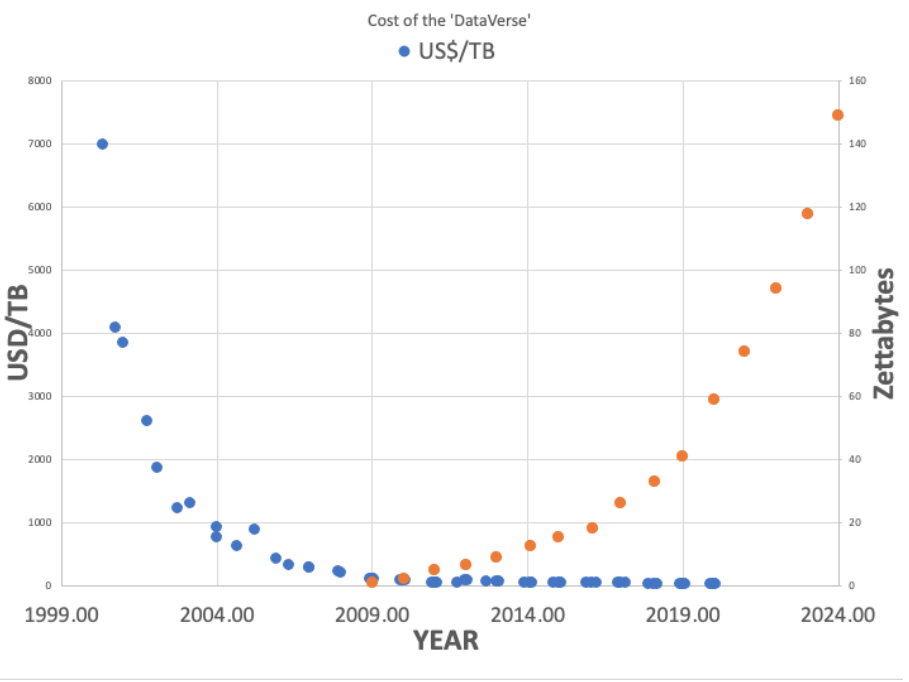
The Challenge (part 2-issue)

Flatlining Storage costs



https://jcmit.net/index.htm

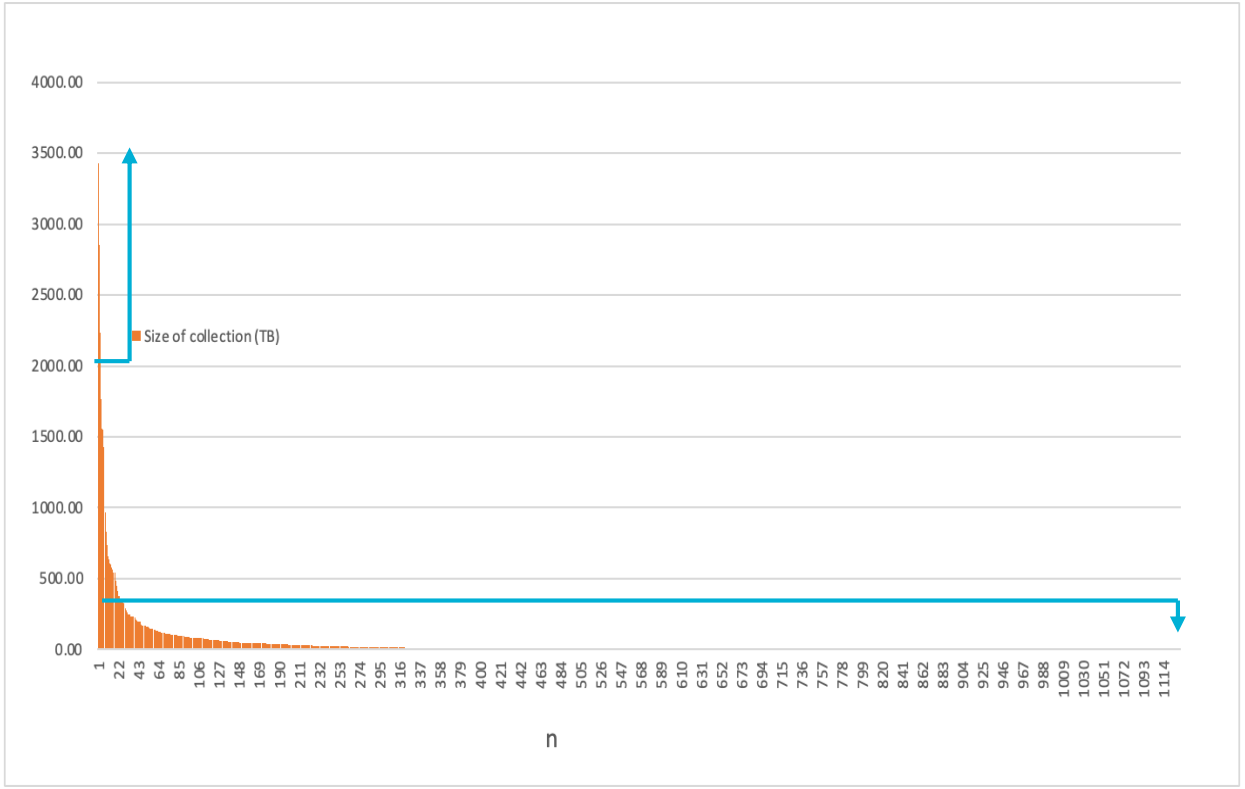




The Challenge (part 3 - how)

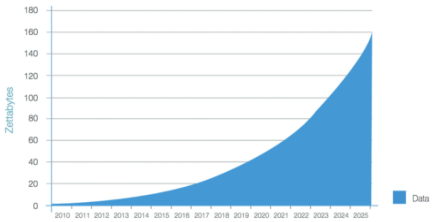
Burden Distribution

Range TB	n	%
0-400	1089	96
>2000	3	0.27

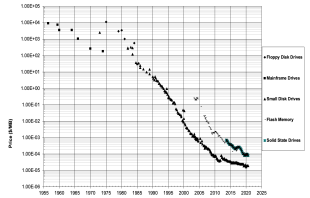


The ~~two~~ three curves

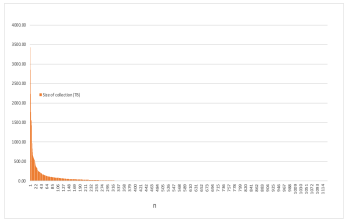
Rapid Accumulation



Flattening capacity costs



Scale distribution



A National view via Metadata

Investment for change

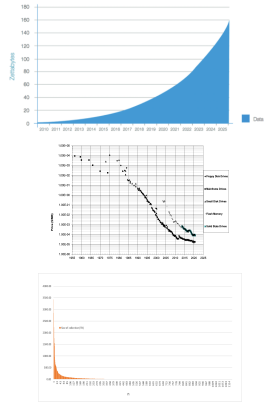
Foundational metadata specification to measure and assess.

Business intelligence – Size, source, ownership

Value proposition – citation, grant, strategic intent

Sector dynamic –where do data persist and how do they get there

Project Approach



SSD

>m/d<

>data<

DataCite
FIND, ACCESS, AND REUSE DATA

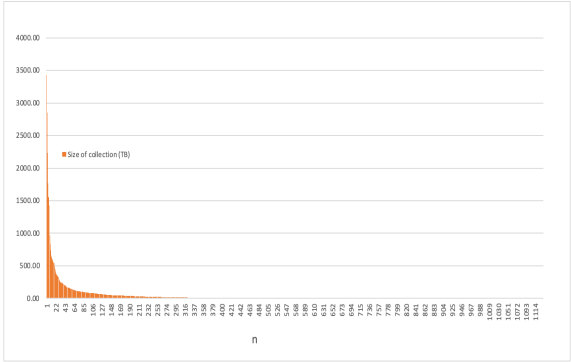
>doi<

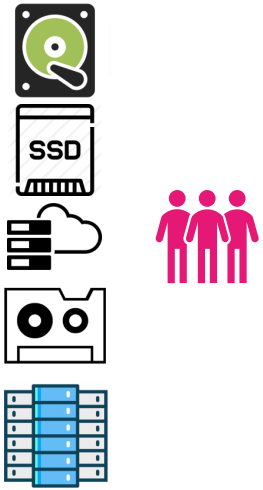
Australian Research Data Commons

NCRIS National Research Infrastructure for Australia
An Australian Government Initiative

ARDC is enabled by NCRIS

Business Intel: Story of divided concerns





Short tail - scale is important.

Sufficiency – traditional capital investment

Long Tail - automation is important.

Efficiency – investment in curation burden

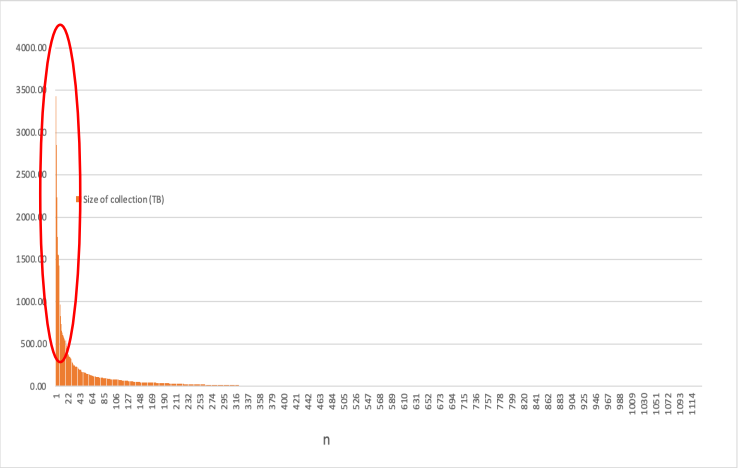
Middle ground - flexibility is important

***x*Efficiency + *y*Sufficiency**

Project Outcomes: in numbers

PHASE 1		
Collection Registers	686	
Register capacity (TB)	34,210	
Collecitons validated	186	**27%**
Validated capacity(TB)	26,855	**79%**
PHASE 2		
Collection Registers	576	
Regsiter capacity(TB)	14,338	
Collections validated	358	**62%**
Validated capacity(TB)	13,731	**96%**
Total Collections	1262	
Total n validated	544	**43%**
Total Capacity(TB)	48,548	
Total Capacity validated	40,586	**84%**

Short tail



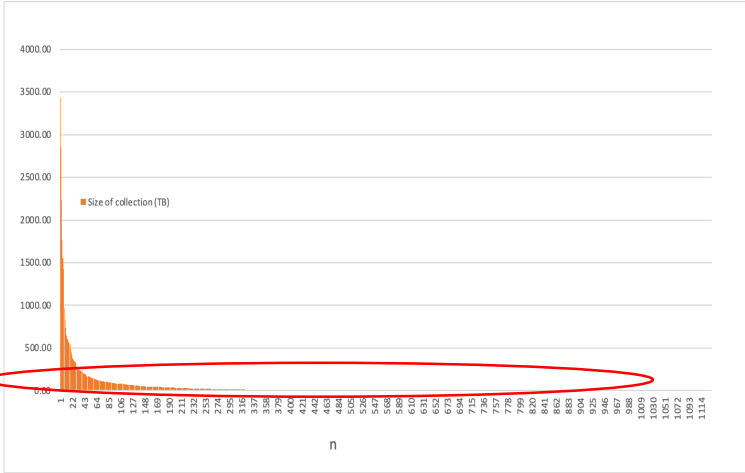
Size TB	n	%
0	77	6.80%
200	1012	89.40%
400	22	1.94%
600	7	0.62%
800	5	0.44%
1000	2	0.18%
1200	0	0.00%
1400	0	0.00%
1600	3	0.27%
>1800	4	0.35%

Scale = 22PB in 21 Data Collections

$/TB

Thresholds of incentive = when operational overheads outweigh curatorial overheads

Long tail



Size TB	n	%
0	77	6.80%
200	1012	89.40%
400	22	1.94%
600	7	0.62%
800	5	0.44%
1000	2	0.18%
1200	0	0.00%
1400	0	0.00%
1600	3	0.27%
>1800	4	0.35%

Complexity = 18PB in 1,111 Data Collections

$/collection

Thresholds of incentive = when curatorial overheads outweigh operational overheads

Known Biases

- **Likely a significant under representation of long tail**
- **Skills distribution**
- **Accuracy in metadata**

Output along the way

Data Collections of National Significance
Version 1.0 Published Nov 2022. 10.5281/zenodo.7329325
Please cite as:
ARDC Ltd. (2022). Data Collections of National Significance.
https://doi.org/10.5281/zenodo.7329326

ARDC DataCite API Jupyter Notebook (on GitHub)

Version 0.1.0 Published Oct 202110.5281/zenodo.5574653
Please Cite as:
Liffers, Matthias. (2021). ARDC DataCite API Jupyter notebook (v0.1.0). Zenodo.
https://doi.org/10.5281/zenodo.5574653

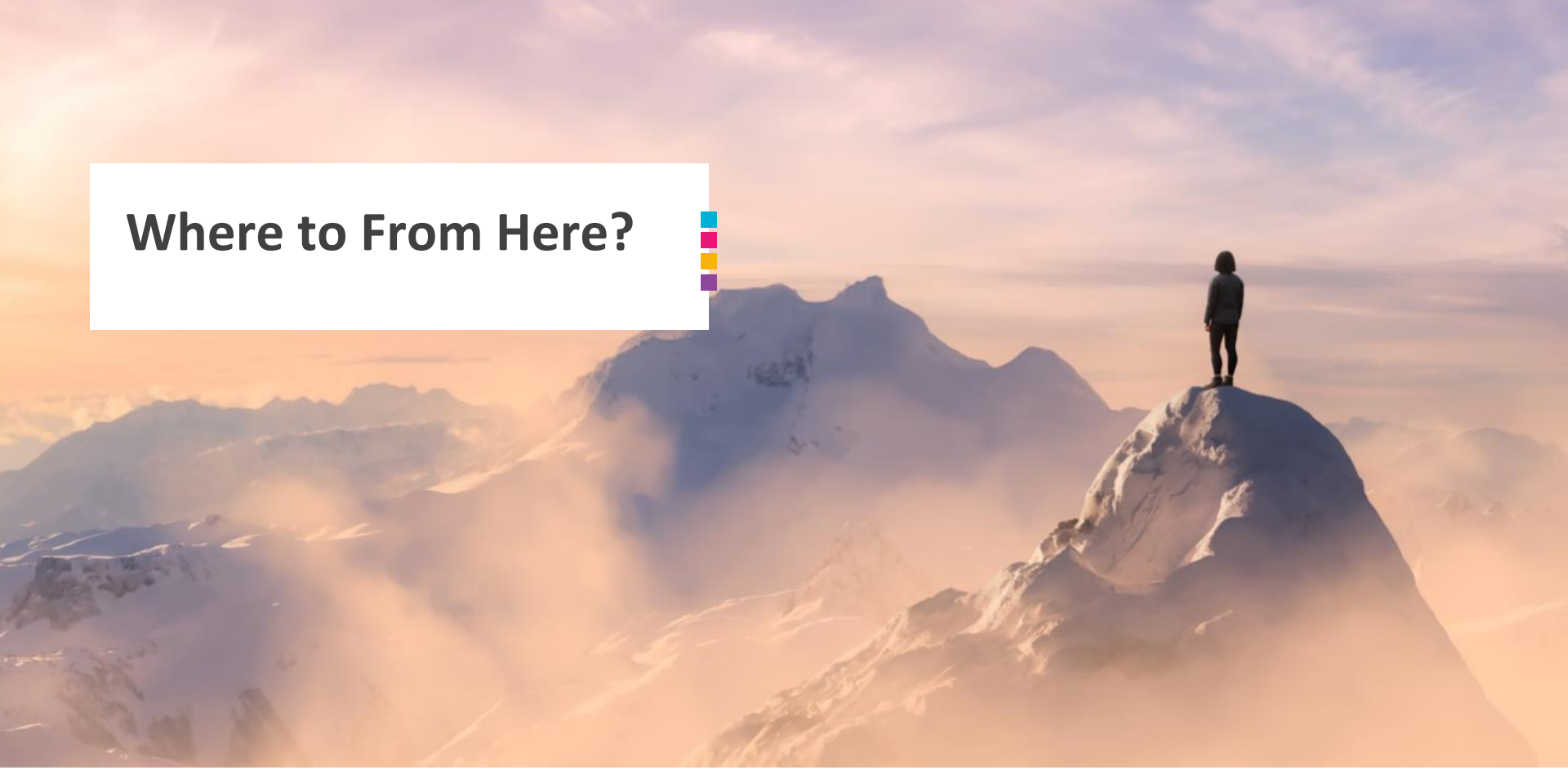
Briefing Note: Why Am I Being Asked About Metadata?
Version 1.0 Published December 2021. 10.5281/zenodo.5778322
Please Cite as:
Australian Research Data Commons. (2021). Why am I being asked for metadata about my research data?. Zenodo. https://doi.org/10.5281/zenodo.5778322

Australian Research Data Commons

NCRIS National Research Infrastructure for Australia
An Australian Government Initiative

ARDC is enabled by NCRIS

Where to From Here?



Australian Research Data Commons

NCRIS National Research Infrastructure for Australia
An Australian Government Initiative

ARDC is enabled by NCRIS

Projections

- **Likely large proportion missed in long tail**
  - More distributed and more heterogenous
  - Almost certainly as much capacity
- **Need to continue to build capability**
  - Tools and process change
- **ARDC Strategic Pillars**
  - Review metadata specification –PIDs and schemas
  - Extend eligibility criteria

Middle ground

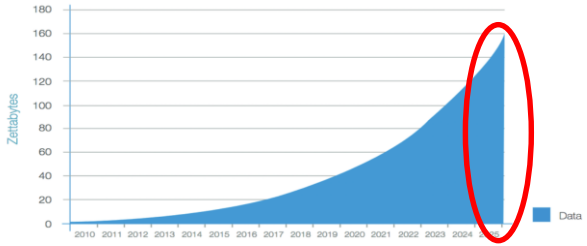
Subsidies that reflect actual market costs

>Complexity - $/collection
+
>Scale ($/capacity)

Investment caps based on real threshold data

>When operational and curatorial burdens are significant

TB=f(capacity)



Middle ground

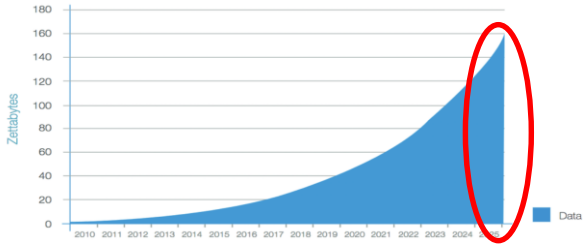
Subsidies that reflect actual market costs

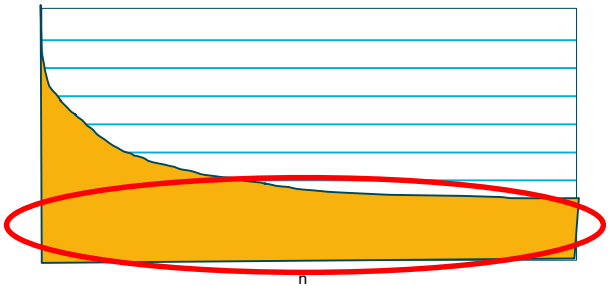
>Complexity - $/collection
+
>Scale ($/capacity)

Investment caps based on real threshold data

>When operational and curatorial burdens are significant

TB=f(value)





Further considerations

- **Access is not trivial**
- **Modifying traditional capital investment is a significant change**
- **Implies governance and ownership are understood (or at least declared)**
- **Reducing factors will vary across disciplines**
    - At least two further dimensions to ‘time’ and ‘value’
        - Lifecycle stage / FAIR maturity
        - Infrastructure system
    - adjacent efforts in
        - Characterising research data scale
        - Conceptual reference architectures for research storage capability

Proposed programme of Activity 2023-2028

1

Investment Model Evolution

- 5 year cycle with multiple accumulation points
- Extend Subsidy Parameters ($/TB + $/n)
- Continue to build community

2

Develop API and other support tools

- Enabling change

3

Support community around data storage infrastructure concerns.

>Scale

>Design

4

Version 2.0 Data Collections of National Significance

Subscribe to the **ARDC CONNECT** newsletter

THANK YOU

ardc.edu.au

contact@ardc.edu.au

+61 3 9902 0585

@ARDC_AU

Australian-Research-Data-Commons

About the ARDC

Australian Research Data Commons

NCRIS National Research Infrastructure for Australia

ARDC is enabled by NCRIS