Research Data–Preserve, Share, Reuse, Publish, or Perish

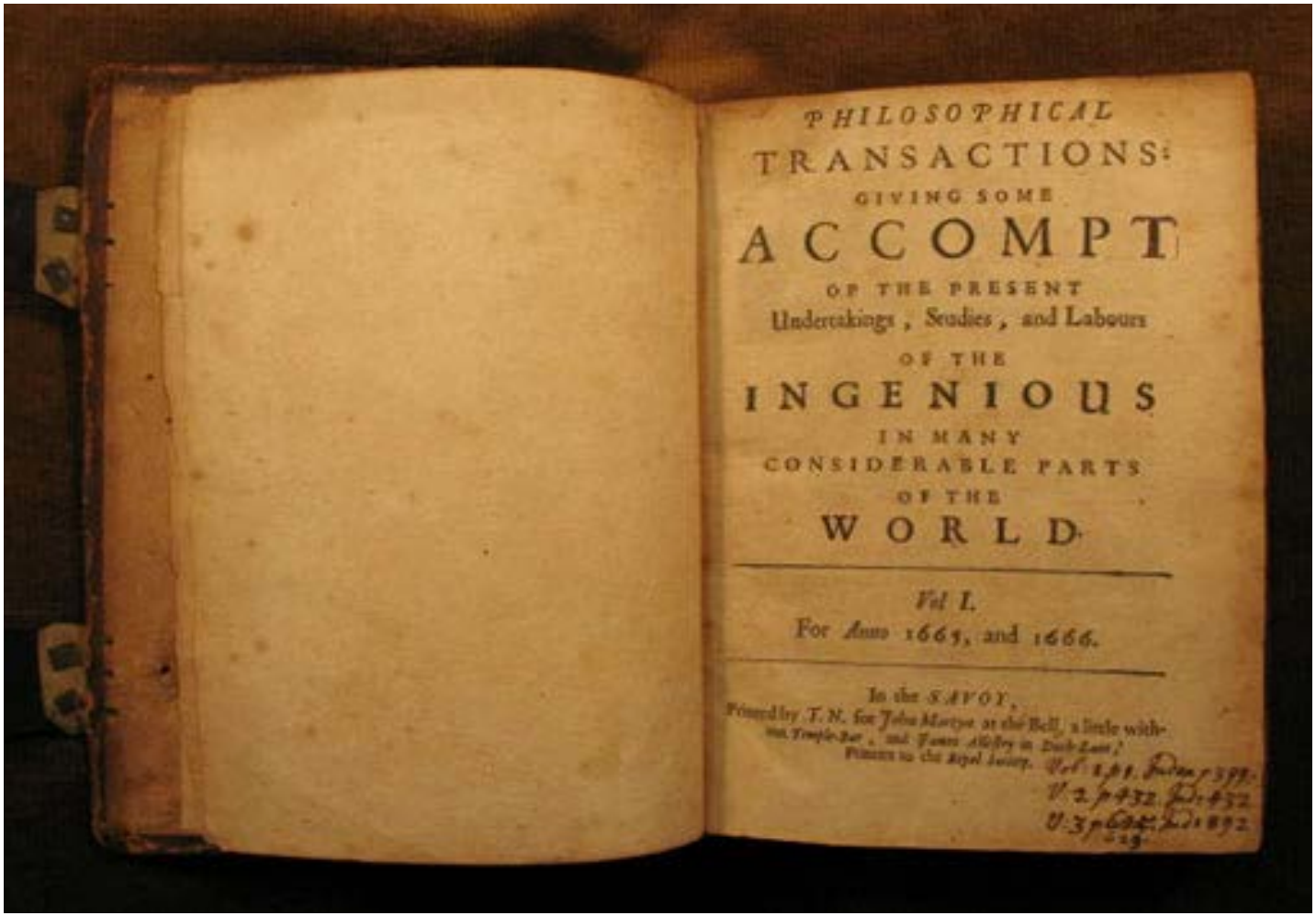
Mark Gahegan
Director, Centre for eResearch
24 Symonds St

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND

Centre for eResearch
The University of Auckland

Outline

- The need
- The approach
- The progress to date



PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOMPT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.

Vol I.
For Anno 1665, and 1666.

In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane, Printers to the Royal Society.

Vol 1 pp 1-399.
V 2 p 432 Dec 432
V 3 p 624 Dec 1892

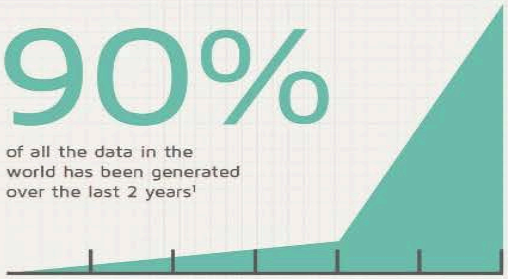

© Jon Cornforth | www.cornforthimages.com

Love your data – practise safe science

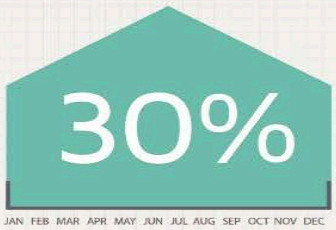
Data output is growing rapidly

90%

of all the data in the world has been generated over the last 2 years[1]



30%

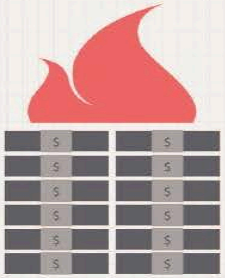


Scientific data output increases at an annual rate of 30%[2]

Despite significant investment, data is not being managed effectively

$1.5 TRILLION

is the current estimated total global spend on R&D, which could be at risk[3]



80% lost

In one study, the odds of sourcing datasets declined by 17% each year, with 80% of datasets over 20 years old not available[4]

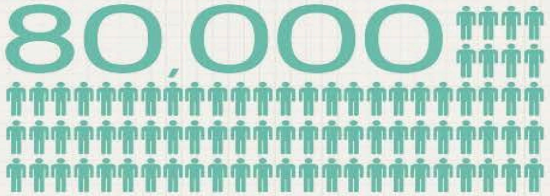
Much of the data remains unverifiable

54%

of the resources used across 238 published studies could not be identified, making verification impossible[5]

Time and money is wasted, impacting on science and society

80,000



Since 2000, over 80,000 patients have taken part in clinical trials based on research that was later retracted because of error or fraud[6]

x5

The number of retractions due to error has grown over fivefold since 1990[7]

Funders now require data management and sharing policies

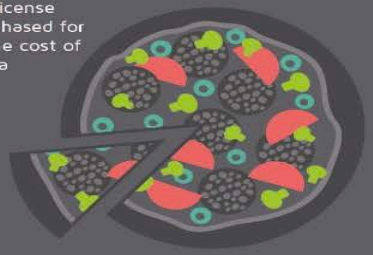
34

countries have signed up to the "Declaration on Access to Research Data from Public Funding"[8]

Key funding bodies such as the NIH, MRC and Wellcome Trust now request data management plans be part of applications[9,10]

Fortunately, digital tools like Projects & figshare make it easy to store, index, search, share, cite and backup data

A Projects license can be purchased for less than the cost of a large pizza



Over 1 million items of research data have been uploaded to **figshare** for safe storage

If you love science, then protect your data. Practise safe science

Projects figshare

Part of the Digital Science family

projects.ac | figshare.com

1. SINTEF (2013, May 22). Big Data, for better or worse: 90% of world's data generated over last two years. http://bit.ly/1IDXQ3Z. 2. Pryor, G. (2012). Why manage research data? In G. Pryor (Ed.), Managing research data (pp. 1-16). 3. 2013 Global R & D Funding Forecast, Advantage Business Media, http://bit.ly/UyETJN. 4. Vine T.H. et al. (2013), The availability of research data declines rapidly with article age. Current Biology (24)1: 94-97. 5. Vasilevsky N.A. et al. (2013), On the reproducibility of science: unique identification of research resources in the biomedical literature. PeerJ 1:e148. 6. Problems with scientific research: How science goes wrong (Oct 19th 2013). The Economist. 7. Steen R.G. et al. (2013), Why has the number of scientific retractions increased? PLoS ONE 8(7): e68397. 8. OECD members and partners: http://bit.ly/SQEyu0. 9. The Digital Curation Centre (DCC), Overview of funders' data policies: http://bit.ly/1LvdR7. 10. NIH Data Sharing Policy and Implementation Guidance: http://1.usa.gov/1cNN0NO

What do we need?

- Data storage services that are reliable
- And that are quick to provision
- Clarity over what we are entitled to
- Services that integrate well into research workflows and practices

What do others need?

- Open data
- Discoverable data
- Well documented or self-describing data
- Reliable data (or at least the quality is assessed)
- Data that can automatically configure itself to our needs

Turning the question around

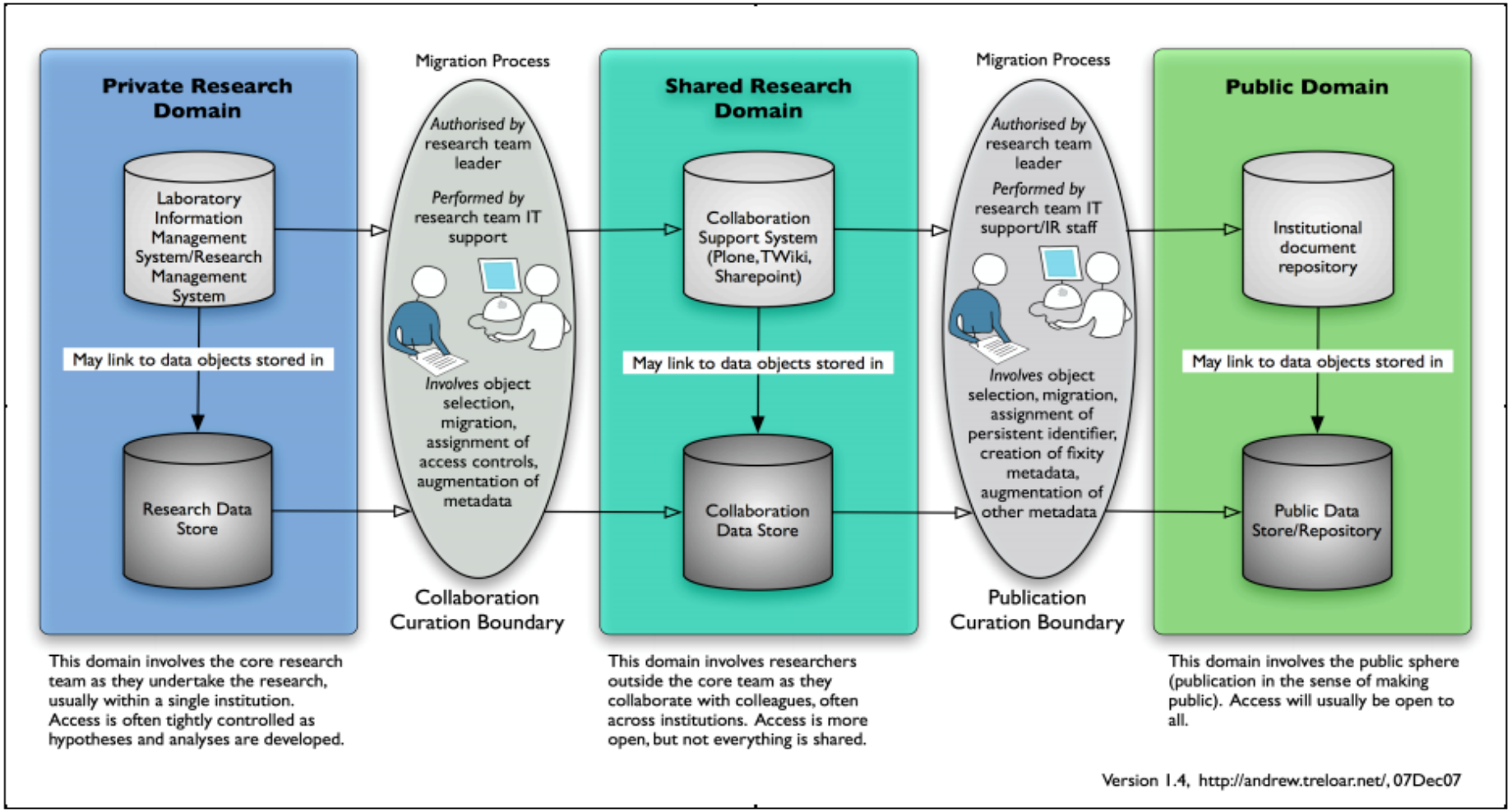
Not: *How do I want to describe the data?* But rather: *What does a future data consumer need to know?*

So, as well as asking:

> *How do we share our data?*

…we should also be asking:

> *What kinds of data descriptions are demonstrably useful to facilitate data reuse?*



Andrew Treloar, Australian National Data Service (ANDS)

Storage Solutions

- **Backup Service**: taking a snapshot of the current stare, just in case…
- **File Sharing Service**: (like *DropBox*) for collaborative work on ‘active data’
- **Data Publishing Service**: (like *figshare*) to publish & discover data, capture metadata and track impact
- **Archive Service**: planned, long-term data preservation for important resources
- **Fast Data Transfer** for accessing remote science equipment

Data Lifecycle Services

- **Creating Data Management Plans (DMPs)**: good practice, and increasingly required by funders
- **Data Ethics Advice**: privacy, encryption, access considerations, disposal
- **Database Design**: for ease of data management and preservation
- **Data Publication advice**: including metadata, ownership and licensing

Data Management Plans (DMPs)

- They describe how (and when) you will take care of, and share, your data
- They show funding agencies they can trust you in turning their money into data
- They help IT support groups understand your needs
- They allow the institution to know what we have, and when we can delete it.

EXPLAIN IT

- **contextualise your material and data**
  Describe the circumstances prevailing at the time of your research and the parameters within which you were working.
- **describe your research process**
  Help people understand your material and data in the future by explaining why you used a particular methodology, or how you analysed your data.
- **explain acronyms and jargon**
  Don't assume the reader will understand specialist terms - remember they may be reading your material in several years' time.
- **provide information (sometimes called metadata) about each file**
  This will help a preservation service to index your material and people to find it. Some of this might be generated automatically by the digital equipment you use.

STORE IT SAFELY

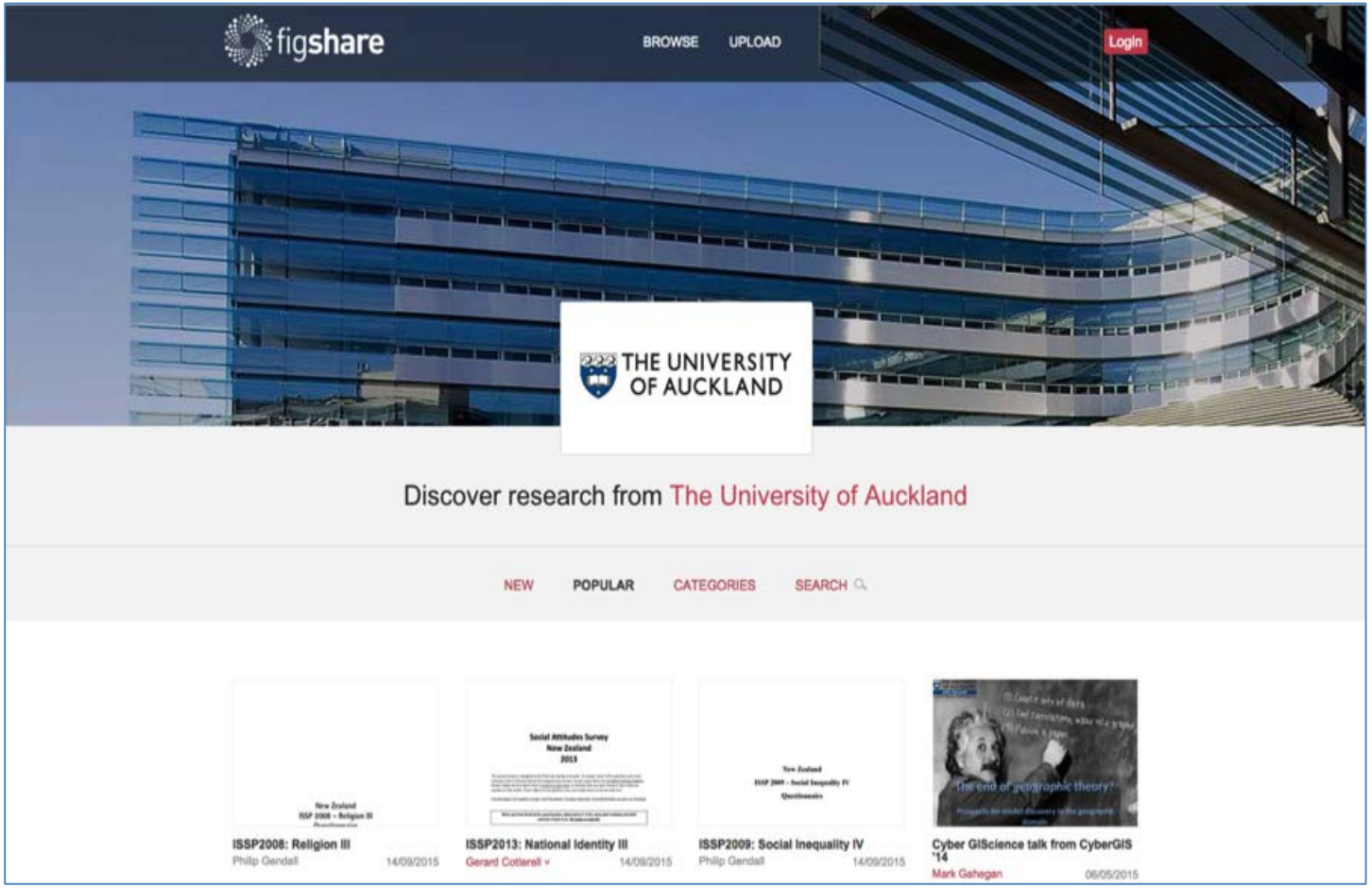
- **make multiple copies**
  Use different types of storage media and store copies in different locations.
- **use open file formats where possible**
  Choosing non-proprietary formats means that files are more likely to be readable in the future. Your library or preservation service should be able to advise you on suitable formats.
- **control who can access your files**
  Take particular care about how you handle and store sensitive information.
- **decide when to delete digital material and data**
  Be selective about what you keep so that it is easier to find relevant and useful information.

SHARE IT

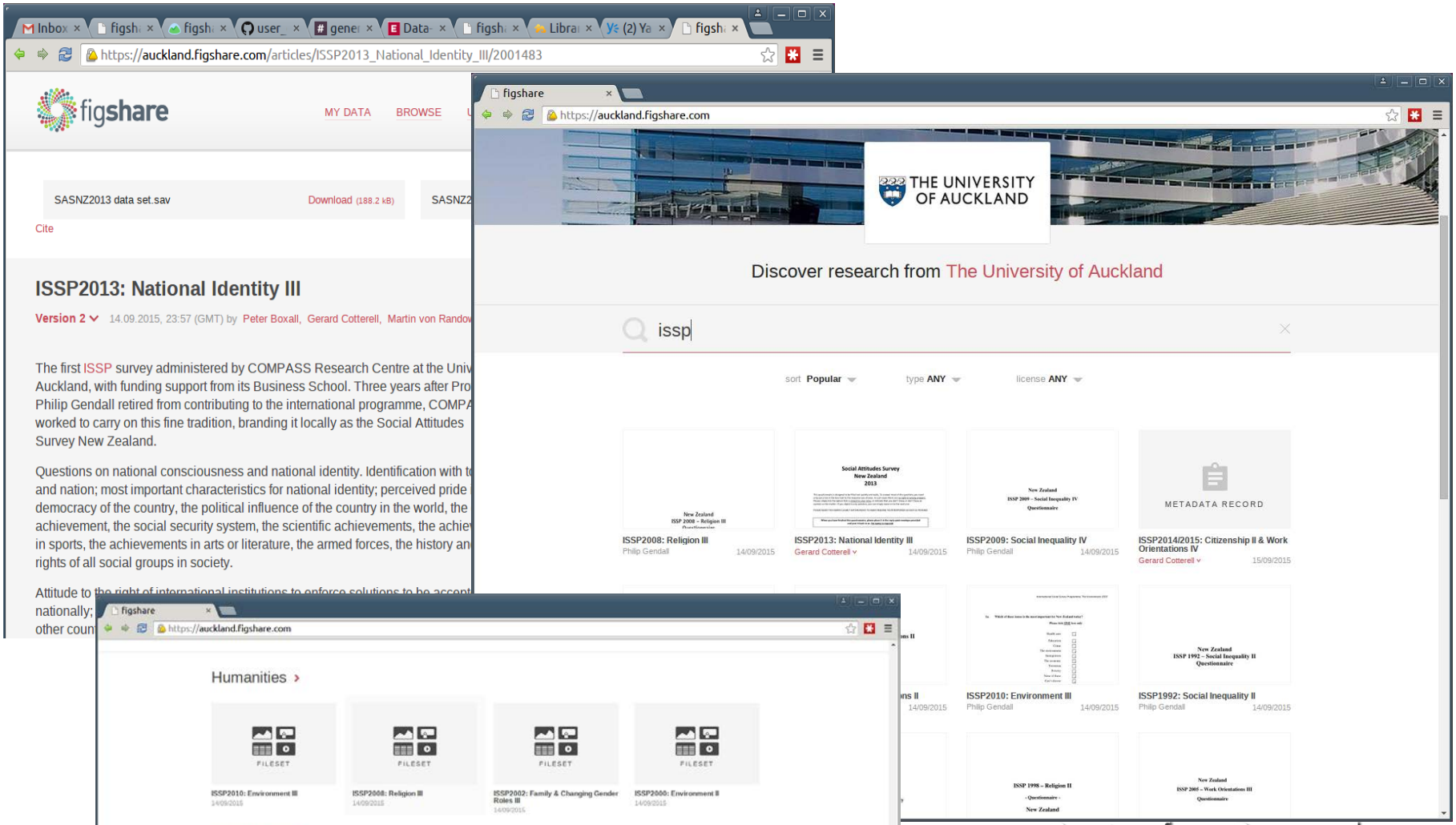
- **to gain more impact**
  Other researchers - in your field or in different disciplines - may want to make use of your material, now and in the future.
- **to enhance your reputation**
  Making research available allows you to demonstrate research excellence, increases your citations and can lead to collaborations.
- **to increase the chance of funding**
  Most funding agencies respond positively to you making your material and data available to others.
- **use repositories and data centres for archiving your material**
  Consider making your research openly available. Choose a repository with controlled access if this is more appropriate for your research.
- **redact or embargo when necessary**
  Your material can still have value when personal or confidential information is removed, and most preservation services will embargo your material while you wait for publications or patents.

http://www.lib.cam.ac.uk/dataman/pages/preservation.html

figshare

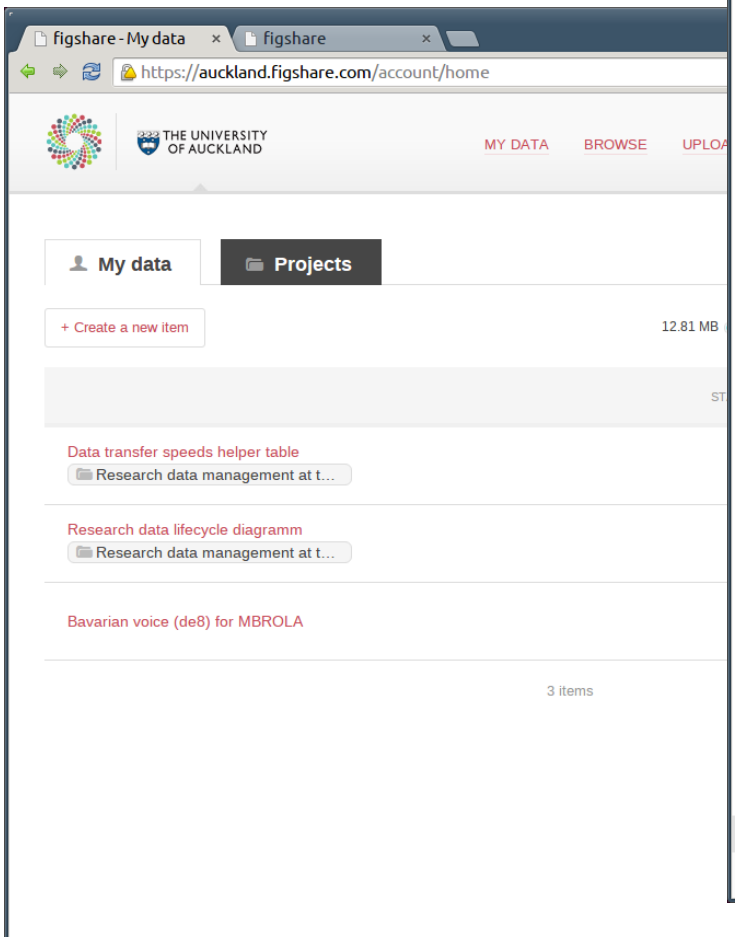

figshare
BROWSE
UPLOAD
Login
THE UNIVERSITY OF AUCKLAND
Discover research from The University of Auckland
NEW
POPULAR
CATEGORIES
SEARCH
ISSP2008: Religion III
Philip Gendall
14/09/2015
ISSP2013: National Identity III
Gerard Cotterell
14/09/2015
ISSP2009: Social Inequality IV
Philip Gendall
14/09/2015
Cyber GIScience talk from CyberGIS '14
Mark Gahegan
06/05/2015

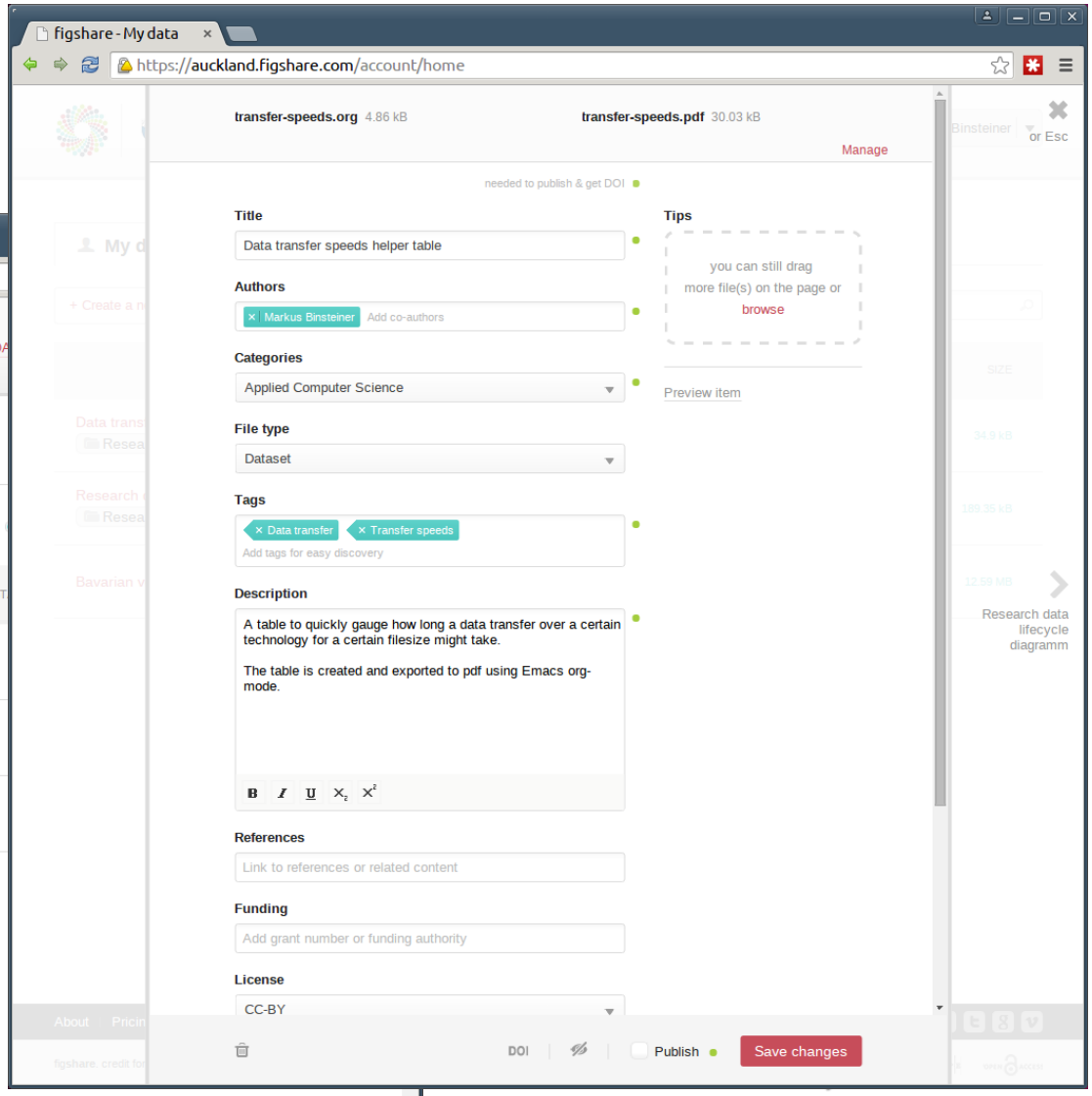
Discover data



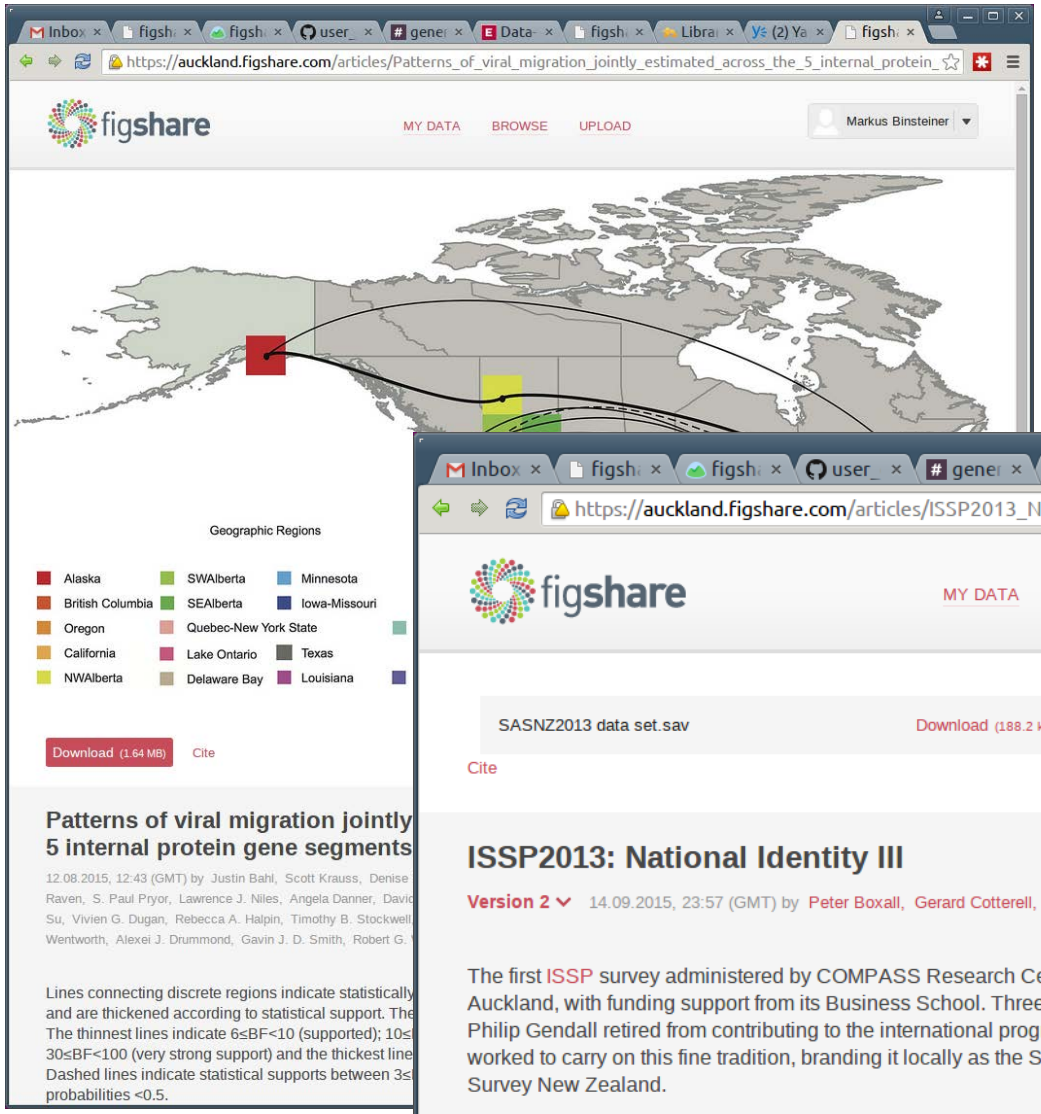
Centre for eResearch
The University of Auckland

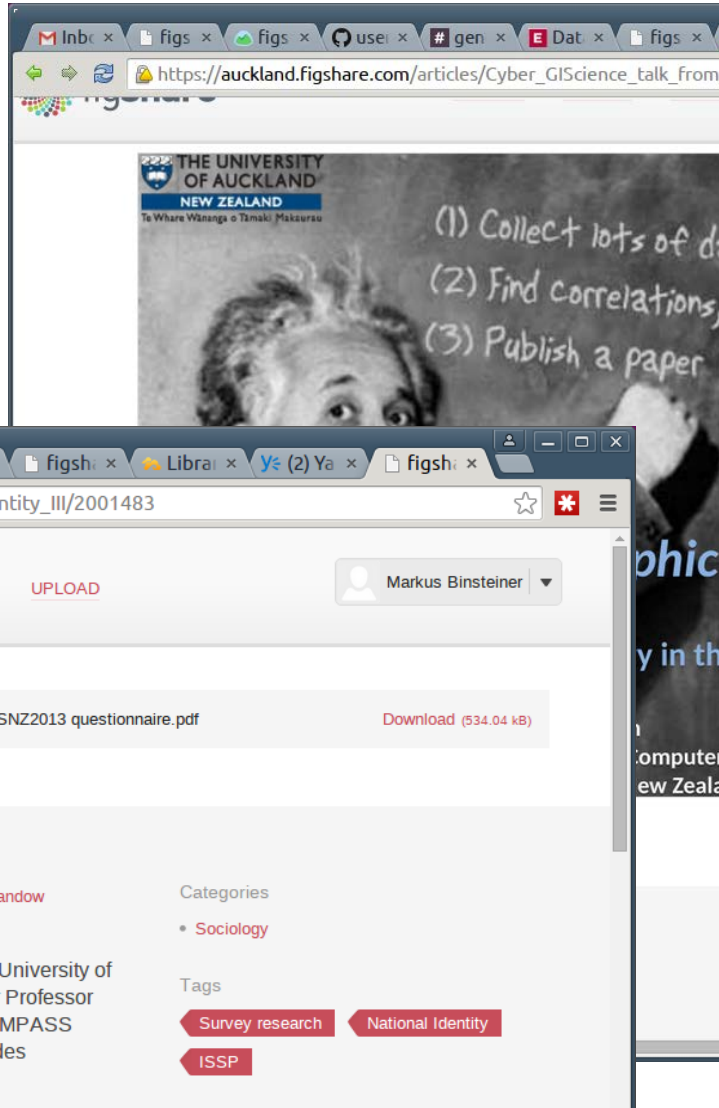
Manage data


figshare - My data
figshare
https://auckland.figshare.com/account/home
THE UNIVERSITY OF AUCKLAND
MY DATA
BROWSE
My data
Projects
+ Create a new item
12.81 MB
Data transfer speeds helper table
Research data management at t…
Research data lifecycle diagramm
Research data management at t…
Bavarian voice (de8) for MBROLA
3 items


figshare - My data
https://auckland.figshare.com/account/home
transfer-speeds.org 4.86 kB
transfer-speeds.pdf 30.03 kB
Manage
needed to publish & get DOI
Title
Data transfer speeds helper table
Authors
Markus Binsteiner
Add co-authors
Categories
Applied Computer Science
File type
Dataset
Tags
Data transfer
Transfer speeds
Add tags for easy discovery
Description
A table to quickly gauge how long a data transfer over a certain technology for a certain filesize might take.
The table is created and exported to pdf using Emacs org-mode.
References
Link to references or related content
Funding
Add grant number or funding authority
License
CC-BY
Tips
you can still drag more file(s) on the page or browse
Preview item
DOI
Publish
Save changes

Publish data





Long term Research Data Challenges

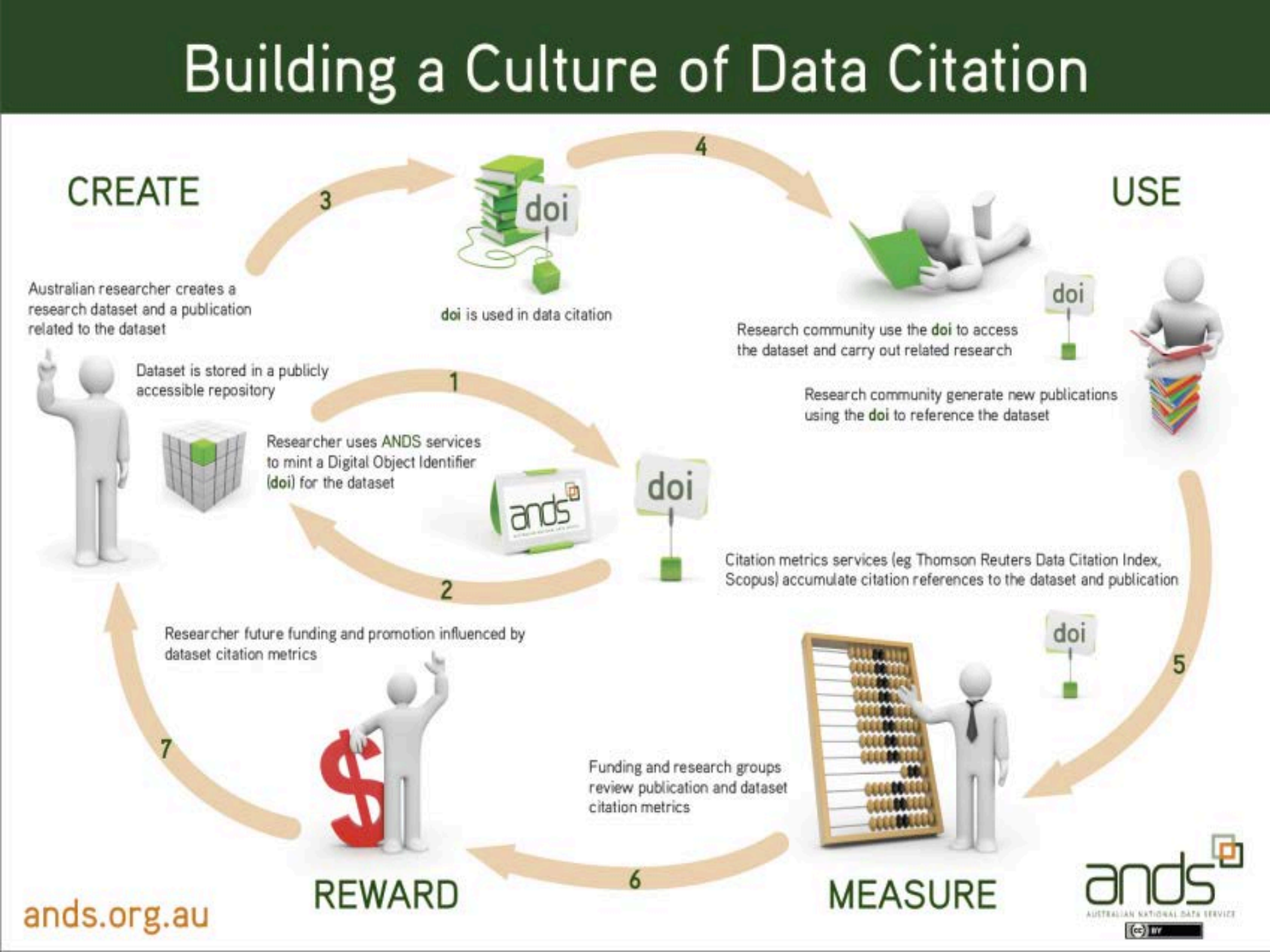
1. **Storing unprecedented volumes of data** (and accelerating)
   - Data production passed storage capacity in 2007
   - Cost differential is increasing, Rate of data production is increasing
2. **Describing what we have** in ways that are helpful to future users (and our future selves)
   - Metadata and Semantics for describing content (this tends to be producer-focused)
   - But also use-case metadata and emergent relationships (tends to be consumer-focused)
3. **Finding what we need**, in the context of our current task
   - semantically-enabled search engines that can use the above descriptions, (ideally from within analytical tools and workflows)
4. **Working out what we do not need to keep**
   - Because it will not be used again or offers no 'information gain'
   - Because it is easier to recreate than to store
5. **Governing data collections well**, within their communities of use
   - effective governance of data resources
   - quality control strategies, including peer review and rewarding excellence

Manifesto for Open Science

1. Remove restrictions on data re-use
2. Data and metadata should be persistent and linked
3. Build or learn strong descriptions of data to aid automated discovery and human comprehension
4. Expose the provenance and uncertainty where possible
5. Develop indicators of data quality
6. Encourage and support secondary use of data
   - Provide a contextual measure of Fitness For Purpose (FFP), to connect researchers with useful resources

Questions?

Building a Culture of Data Citation



ands.org.au

ands
AUSTRALIAN NATIONAL DATA SERVICE

1. Online submission of data for publication with basic metadata
*2. Editor verifies that the data is within the scope of the collection
3. Automated tools check data for obvious omissions and errors.
4. Online tools ingest and integrate data & generate tables of statistics
**5. Potential errors and omissions reported to data author and/or editors
6. Data author acts on this feedback
7. Automated checks verify that data set is complete and standardised.
***8. Data editor confirms that resubmitted data and metadata are correct
9. Independent peer review of data
10. Author responds to referees' comments
11. Editor makes a publishing decision based on quality standard achieved by da set, (including reject and revise and resubmit).

****12. Data and metadata are published online. The data has its own identity that tracks its use, or is integrated into the authoritative subject databases

*****13. Papers are published that consumed the data and any errors found have been corrected

Biodiversity data should be published, cited, and peer reviewed
Mark J. Costello1, , William K. Michener2, Mark Gahegan3, Zhi-Qiang Zhang4, Philip E. Bourne5

Storage devices: Cost vs Value vs Use

	$0.026/GB	$0.05/GB	$0.60/GB	>$1/GB	$0.12/GB/pa
2PB =	$52K	$100K	$1.2m	$2m	$240K/pa
IOPS	0	200	80K-100K	250K-2m	0

Near Future

3D 2.5” SSD form factor, 10TB, 20TB by 2017
$0.05/GB
Also 3D mSSD form factors aiming for 2TB
(laptop dimensions)

Centre for eResearch
The University of Auckland

Tim Chaffe, UoA Chief Architect: Valid as at 8/4/2015 with the proviso that this market is changing monthly