



The Australian National Data Service

David Groenewegen
Deputy Director

What is ANDS?

- An initiative of the Australian Government being conducted as part of the National Collaborative Research Infrastructure Strategy (NCRIS)
- A collaboration between Monash University, the Australian National University and CSIRO
- Ran an establishment project in 2008, fully operational since January 2009

NCRIS

- Announced in the 2004-2005 budget, NCRIS is providing \$542 million from 2004-05 to 2010-2011 to develop and fund national research infrastructure projects.
- Through NCRIS the Australian Government is implementing a strategic and collaborative approach to investment in world-class research facilities, networks and infrastructure that are accessible to researchers and meet their long-term needs.
- NCRIS has departed from common practice in its approach to developing project proposals for research infrastructure.
- Rather than seeking proposals, the NCRIS Committee commissioned independent external facilitators to develop strategic, national investment plans for priority capabilities identified in the NCRIS Roadmap. In developing the investment plans, facilitators worked with researchers, research managers, research funders and users, to define the infrastructure requirements and the collaborative arrangements for managing the operation and accessibility to facilities and equipment. While it is not a competitive grants program, the projects must represent excellence in their fields.

NCRIS Investments

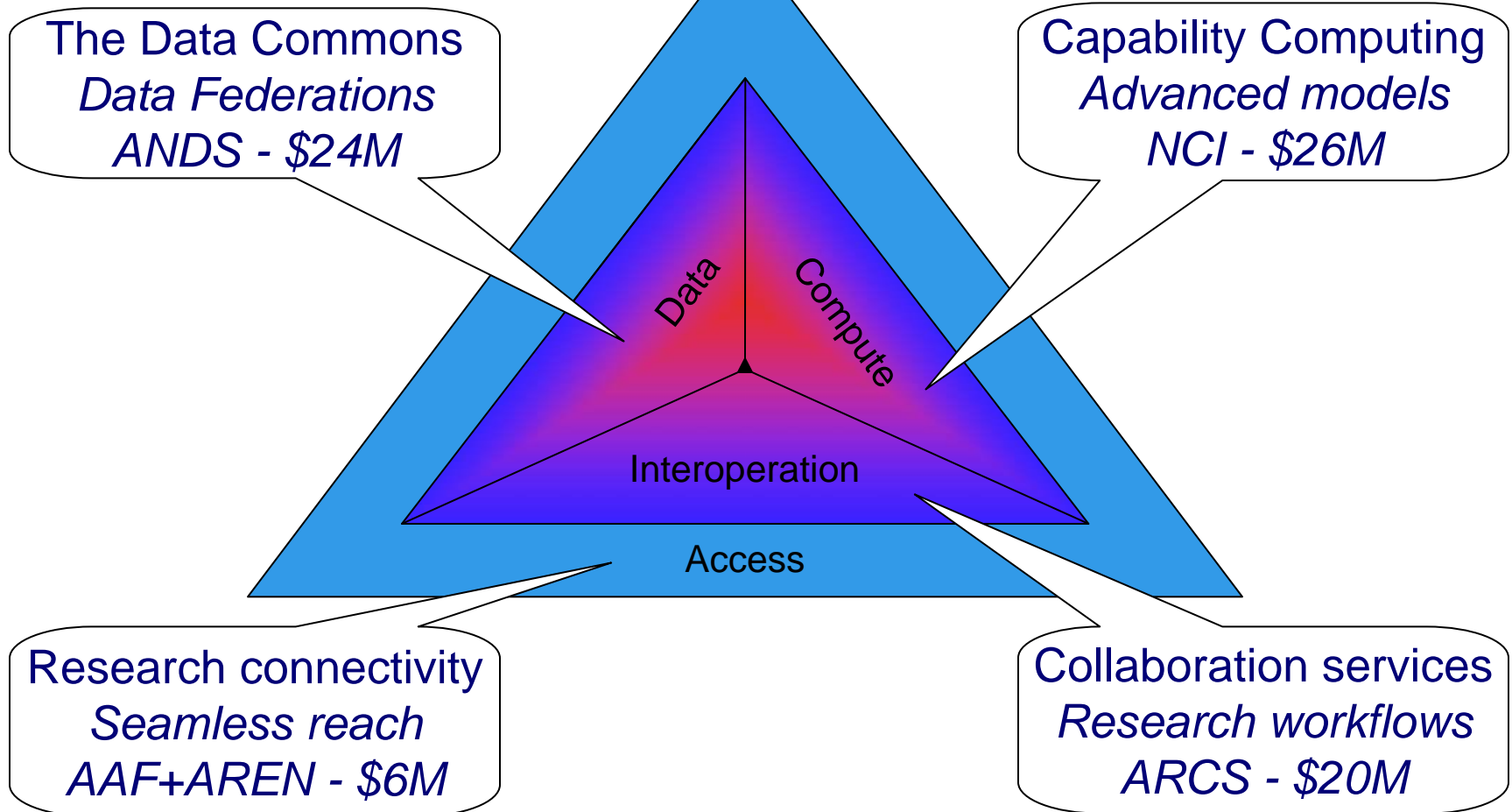
\$542M* over the five years: 2007-2011

- | | |
|--|--|
| <ul style="list-style-type: none"> ■ Evolving bio-molecular platforms and informatics ■ Integrated biological systems ■ Characterisation ■ Fabrication ■ Biotechnology products ■ Optical and radio astronomy ■ Integrated marine capability ■ Structure and evolution of the Australian continent | <ul style="list-style-type: none"> ■ Networked biosecurity framework ■ Population health and clinical data linkage ■ Terrestrial ecosystem research network |
|--|--|

+ Platforms for Collaboration (allocated \$82 M ~EU45M)

***Note: scaled to EU or US economies this is analogous to 1B USD per annum*

Platforms for Collaboration: Major Investments 2007-2011



The increasing importance of data

- With more data online, more can be done
- The cost of data acquisition is going down – dramatically
- Possible to get new answers from old data
- Increasing focus on problems across disciplinary boundaries e.g. climate, biosecurity

Research Data Intensity

- Square Kilometer Array will generate an exabyte per day
- Large Hadron Collider will generate a petabyte a month
- A current generation gene sequencer can generate a terabyte per day
- Sensors will routinely be deployed to generate enormous and varied data streams for all disciplines
- More data is being captured now that cannot be ever captured again
- More data was created last year than can be stored – IDC -

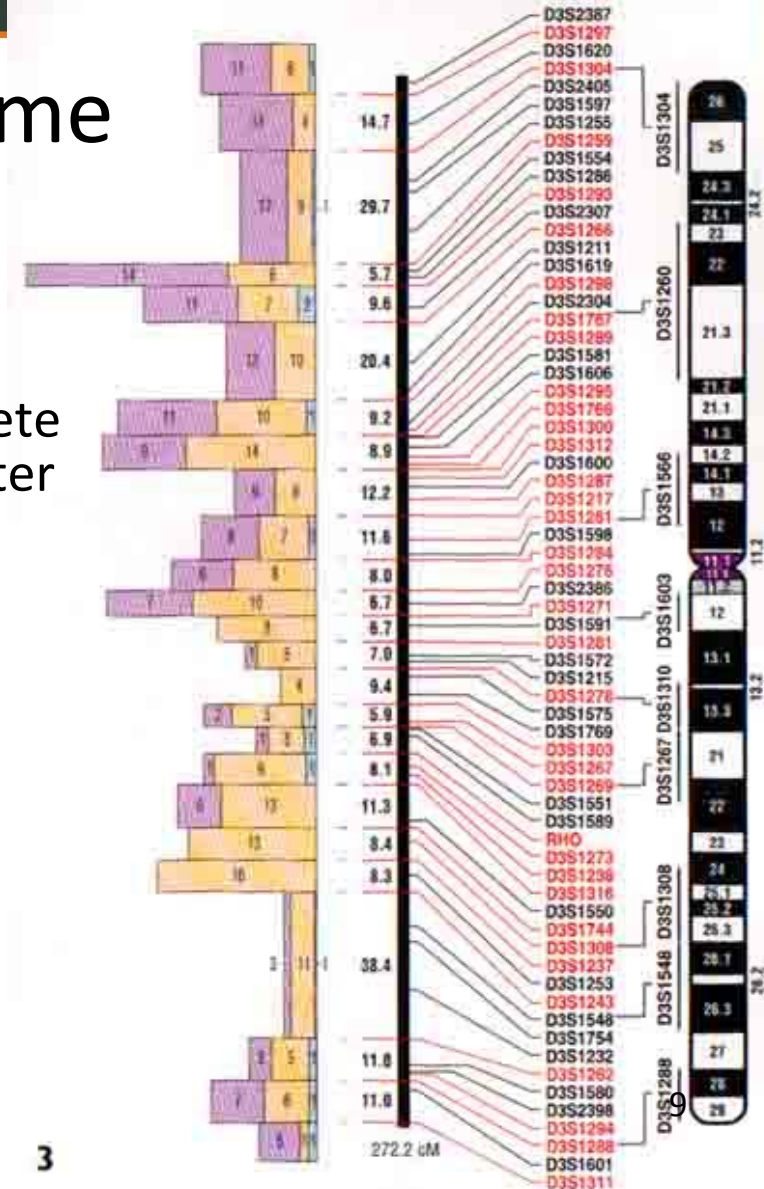
<http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf>

Research Data as Research Product

- The Human Genome project is known for its data
- Currently linguistics is concentrating on data capture rather than analysis as languages expire
- Research data that is used collaboratively can be central to research
- Hubble telescope data is an output

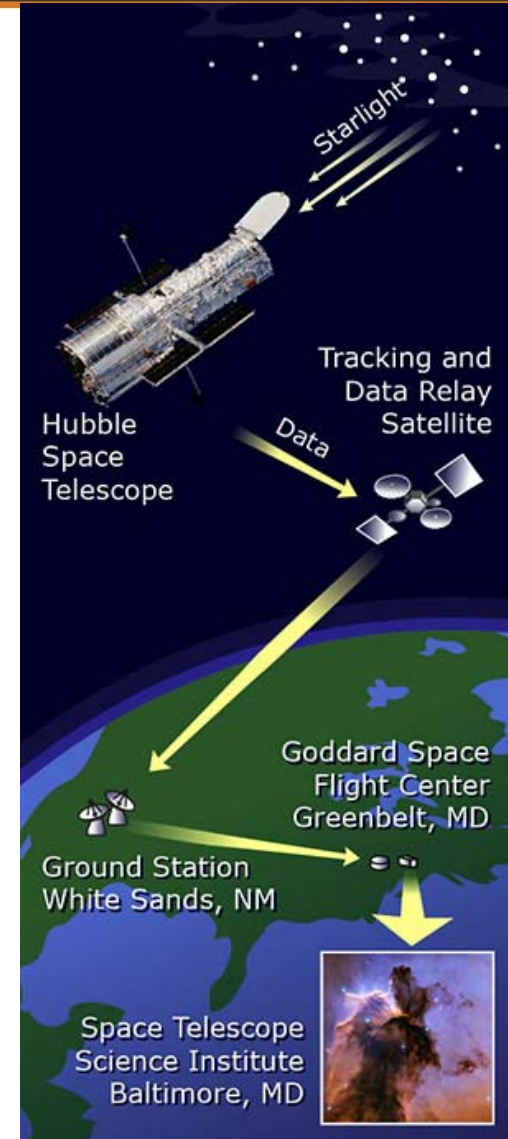
Mapping the Human Genome

- Took a large team of scientists 10 years to map the 30,000 genes that describe the human body
- In 2007, Craig Venter, published his complete DNA sequence, unveiling the six-billion-letter genome of a single individual for the first time
- The work required a large team using new instruments to produce a large dataset – indeed 2 competing large teams!
- No single lab could have completed this project with available technology in a reasonable time



The Hubble Telescope

- The Hubble telescope launched in 1990
- Increasing focus on cross-disciplinary science
- Observations are proposed, and if accepted, data is collected and made available to the proposers – who then write a research paper
- Each year around 1,000 proposals are reviewed and approximately 200 are selected, for a total of 20,000 individual observations
- The data is stored at the Space Telescope Science Institute
- There are more research papers written by “second use” of the research data, than by the use initially proposed



Excellence

Sharing Detailed Research Data Is Associated with Increased Citation Rate

- 48% of 85 cancer microarray clinical trial publications with publicly available microarray data received 85% of the aggregate citations

Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308

To extend the data driven research world

More researchers re-using data more often

To achieve this:

- Lower the costs and raise the benefits
- Data is seen as a first class research output
- Data is seen as an output of first class research
- Partnerships across the sector

Who cares about data?

- The Australian Government - \$72M for ANDS***
- The Cutler report – our innovation system
- The institutions
- The disciplines??

- ...and researchers should care..

Why might an institution care about ANDS?

- Research has become more data intensive → data management
- Data is increasingly a research output, rather than a research by-product → data infrastructure
- Excellence in research is correlated with size of effort and with data outputs → data preservation
- Effective response to the AVCC Code for the Responsible Conduct of Research may be best done collectively

Australian Code for the Responsible Conduct of Research

- It describes the responsibilities of institutions and researchers in the management of research data and primary materials
- Institutions are to retain research data, provide secure data storage, identify ownership, and ensure security and confidentiality of research data
- Researchers are to retain research data and primary materials, manage storage of research data and primary materials, maintain confidentiality of research data and primary materials

http://www.nhmrc.gov.au/publications/synopses/_files/r39.pdf

ANDS has conducted fora on Institutional responses to the data aspects of the code - <http://ands.org.au/news/codeforums.html>

Why should researchers care – and why not?

- The Code requires them to
- The role of data citation
- The changing nature of research

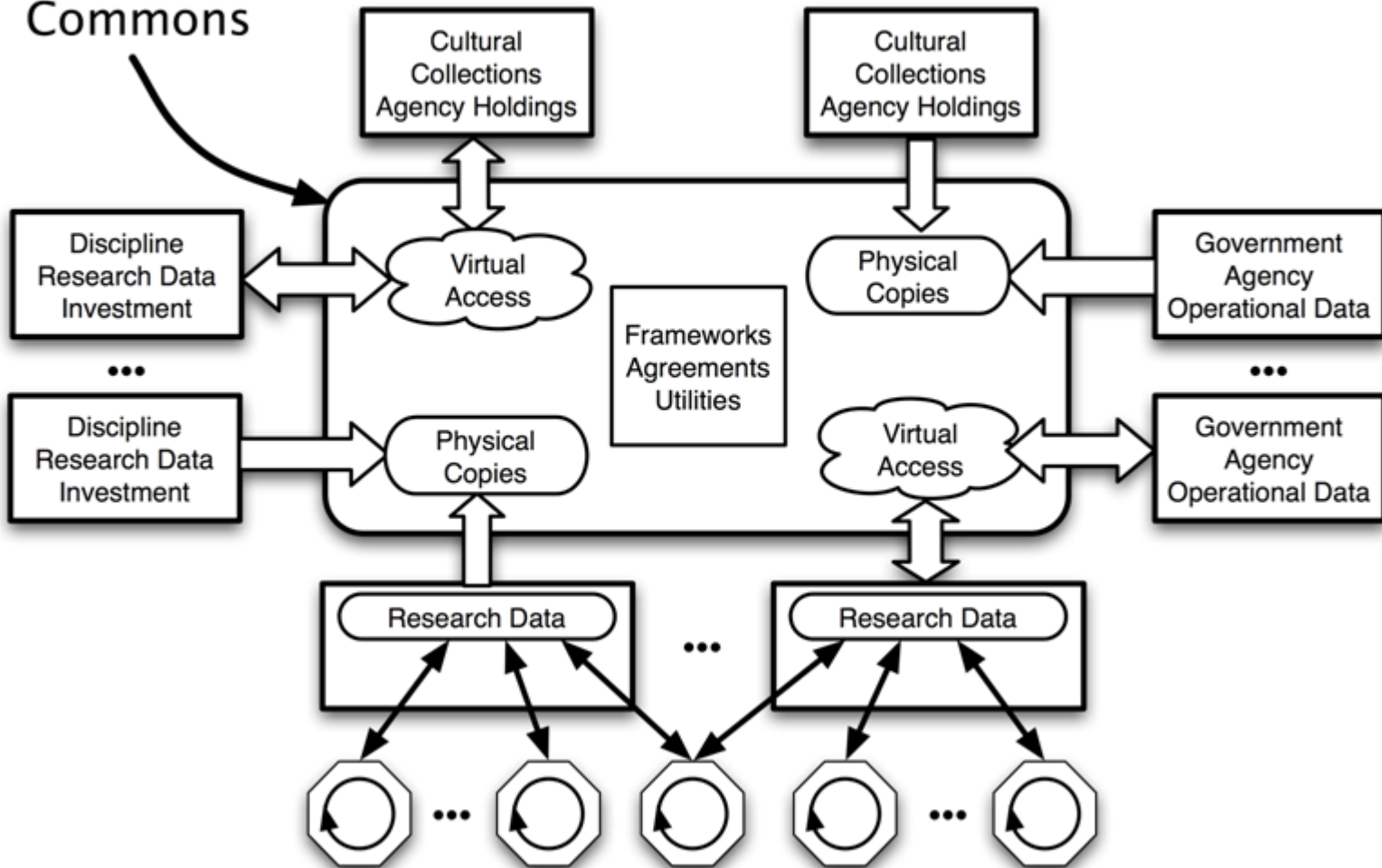
.. but the costs often outweigh the benefits, so ANDS needs to help change the equation

To enable the change:

- Shared desire to change
- Professional services – research data analysts, research data carers, professional programmers
- Change partners such as ARCS, ANDS, eResearch orgs.
- Changed status of research data

So how do we plan to do this?

Research Data Commons



ANDS has been structured as four co-ordinated inter-related service delivery programs:

- *Developing Frameworks* – the frameworks that will enable research data producing institutions to capture, manage and share research data;
- *Providing Utilities* –services that reduce the cost of capture and ease the task of discovery;
- *Seeding the Commons* – improving local data capture and populating the data commons; and
- *Building Capabilities* - improving Australia’s capability to manage its research data.

Plus service development for specific disciplines funded through NeAT projects in partnership with ARCS

Frameworks

- Influencing relevant national policies
- Building common understanding of data management issues and solutions across government, research funding agencies, and research intensive organizations
- Encouraging move in favour of relevant default sharing practices

Utilities

- Building and delivering national technical services to support the data commons
- Examples:
 - Discovery
 - Persistent identifier
 - Collections registry
- Mostly outsourced delivery

Seeding the Commons

- Improving and standardising institutionally supported repositories
- Supporting Australian research data being routinely deposited into stable and sustainable data and preservation environments
- Working with existing document and data repositories; big focus on content recruitment
- Will need to provide targeted assistance for things ANDS cares about most

Capability Development

- Assisting researchers to align their data management practices with the needs and outputs of ANDS
- Assisting communities to do things that increase quantity and quality of data available to the data commons
- Learning from existing good practices
- Building culturally-relevant default sharing practices

ANDS engagements

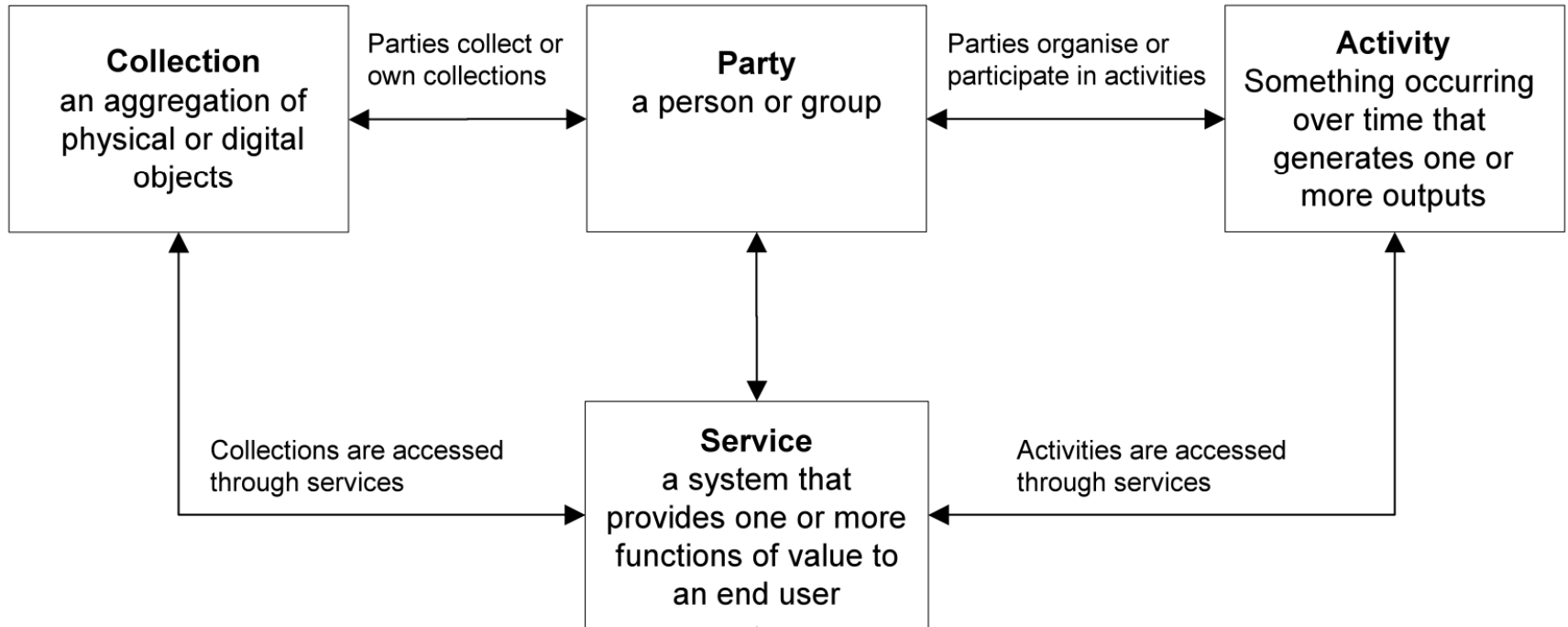
- Our role is to help researchers and institutions achieve their research data ambitions whilst creating the research data commons
- Our model is to partner at the institutional level and through local partners
- We don't have "best practice" to distribute, but sharing good practice
- It works best when the research office, the library and ITS work together to meet mutually desired outcomes

Early Deliverables

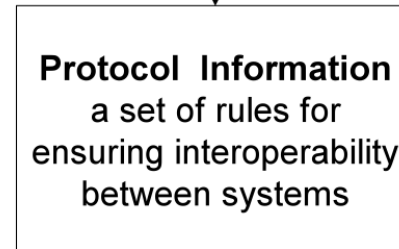
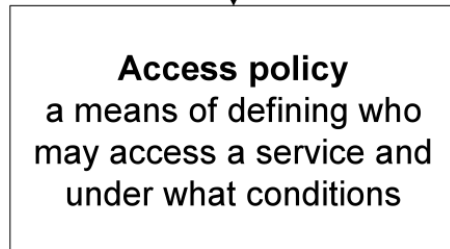
- A national collections registry service – *Register My Data* (<http://www.ands.org.au/services/register-my-data.html>)
- A national persistent identifier service – *Identify My Data* (<http://www.ands.org.au/services/identify-my-data.html>)
- A national discovery service – *Research Data Australia* – (<http://services.ands.org.au/pages/>)
- *Not a national storage and management facility, not provision of storage hardware*

ISO2146

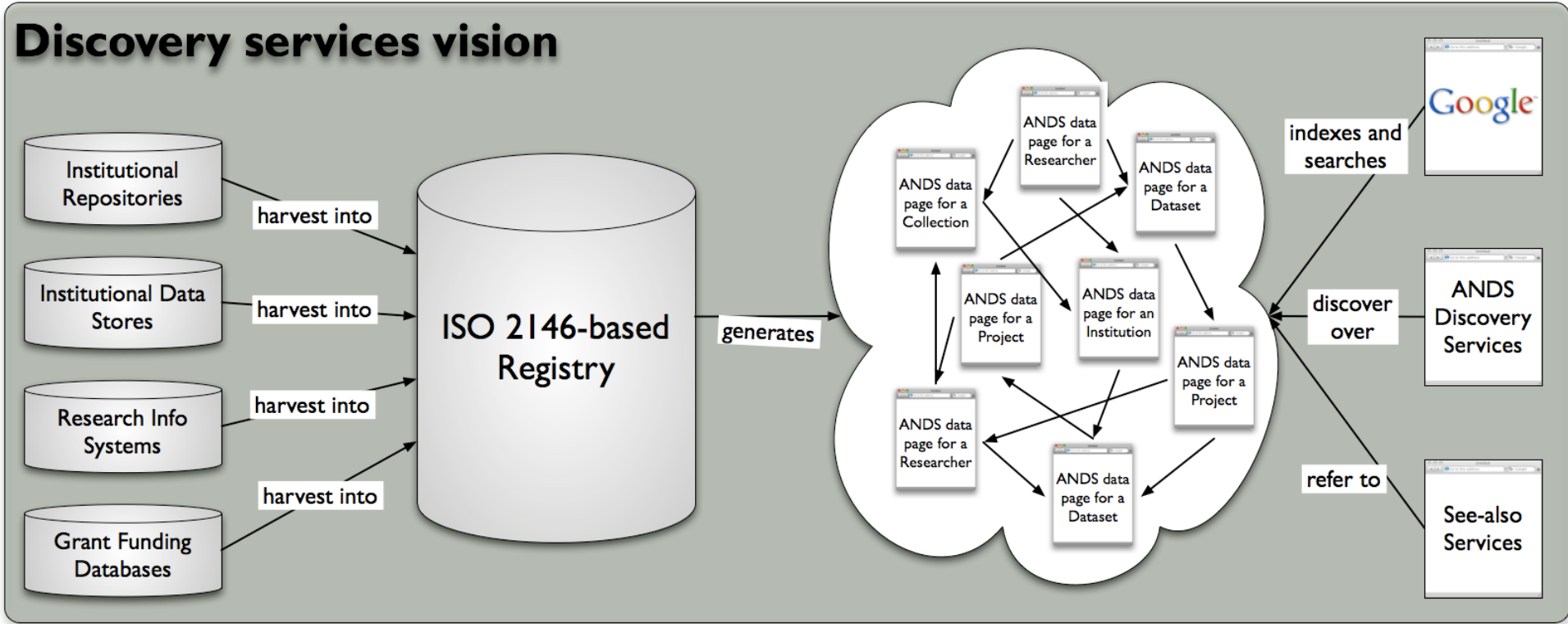
Registry Object



Services have access policies



Services may be delivered through protocols



Other focus areas

- Exemplar discoverable data spaces
 - Water?, Crystallography?, Law?, Bio-security?, Coast?
- Early advice on data capture and management practice
- Partnering with selected institutions
- Building shared practice in response to the Code
- Partner with institutional ITS and ARCS for local repositories and storage

What do researchers get?

- Locally managed data in a repository
 - Meets required practice, and enables personal re-use
- Persistently identified data
 - Enables data to be a first class research output
- Explorable data
 - Enables data to be explored in the context of the research group, the institute, the collection, the research project
 - A rich discovery environment
- ***ANDS has to help lower the cost of data provision and increase the benefits of sharing***

Education Investment Fund

The Education Investment Fund (EIF) was announced in the 2008-09 Budget and is a major component of the Government's Education Revolution. The role of the EIF is to build a modern, productive, internationally competitive Australian economy by supporting world-leading, strategically-focused infrastructure investments that will transform Australian tertiary education and research.

The EIF will provide funding for projects that create or develop significant infrastructure in higher education institutions, research institutions and vocational education and training providers, in order to:

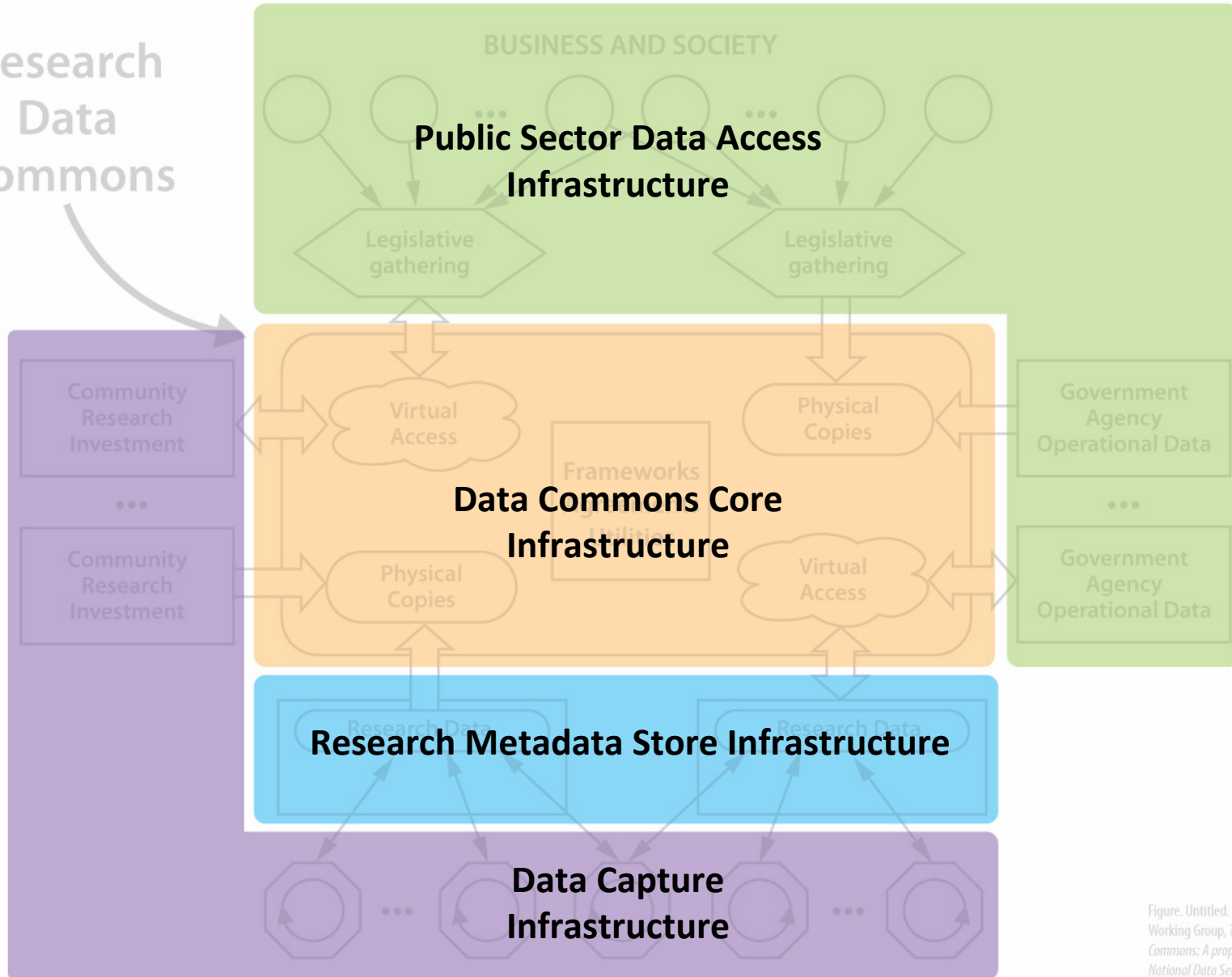
- transform Australia's knowledge generation and teaching capabilities
- boost participation in tertiary education
- position Australia to meet domestic skills needs now and into the future
- enhance Australia's innovation capacity
- invigorate the growth of Australia's research capabilities
- enhance Australia's international competitiveness in education and research.

ANDS Role in EIF

- The Australian Research Data Commons will support the discovery of, and access to, research data held in Australian universities, publicly funded research agencies and government organisations for the use of research. This investment will enable the construction of a range of ICT utilities to capitalise on and ensure greater use and re-use of existing data resources, as well as better management of new data generated in Australian research.
- Specifically the project will allow the
 - establishment of the core infrastructure of the Australian Research Data Commons (ARDC)
 - construction and installation of systems that will identify and describe significant data holdings for the commons
 - construction and installation of systems that will permit researchers to explore Australia's research data holdings, and
 - construction and installation of systems that will permit researchers to exploit the ARDC.
- Cost: \$AU48 million (EU28 Million) over 2 years

Data Commons Applications Infrastructure

Research
Data
Commons



How will we know if we have succeeded?

- An Australian Research Data Commons – a map across Australia's research data, and easy access down into the data
- Skills and tools that enable research data to be managed, shared and cited more easily
- Skills and tools that enable research data to be easily discovered
- **Researchers re-using research data more often**

Thank you for your attention

Questions?

Contact ANDS via:

<http://www.ands.org.au/contact.html>