THE UNIVERSITY OF MELBOURNE

FACULTY OF BUSINESS & ECONOMICS

Australian Science, Technology and Research Assessment (ASTRA) Feasibility Study: Final Report Prepared for DIISRTE

Paul Jensen and Julia Lane

February 2013

MELBOURNE INSTITUTE®
of Applied Economic and Social Research

Australian Science, Technology and Research Assessment (ASTRA)

Feasibility Study

Final Report Prepared for DIISRTE

by

Paul Jensen and Julia Lane

February 2013

THE UNIVERSITY OF MELBOURNE

MELBOURNE INSTITUTE®
of Applied Economic and Social Research

Table of Contents

Executive Summary

Australia has traditionally had an international leadership role in documenting the results of research investments. But new technologies, and, more importantly, new scientific advances in understanding the research and innovation ecosystem, have created a new environment and a new opportunity for Australia to lead. The ASTRA Feasibility Study investigates the potential for doing just that. It examines the potential for using new technologies to describe the results of research investments for both *advocacy* and *management*, and to build a scientific community to understand how those research investments stimulate innovation in both Australian research and the Australian economy.

The underlying framework for the study is based on the view that research funding has an effect on knowledge by influencing the way in which research ideas are created and transmitted through research networks, which can be across campuses, disciplines, states and international boundaries. The eventual goal is to fully capture the activities of networks in order to measure the research, social, economic and workforce "products" that are generated as a result of the interactions. By describing these "outputs" – and explaining the mechanisms underpinning their creation – ASTRA will provide an important stepping stone in deepening our understanding of the impact of science and research investments on Australian (and indeed global) society. Notwithstanding these important long-term objectives, it is important to bear in mind that these are very complex issues and what is presented here is a Feasibility Study on an important first stage of the process.

The Feasibility Study – which was conducted between April and December 2012 – examined whether data on individual researchers at two universities (the Australian National University (ANU) and the University of Melbourne (UniMelb)) can be linked to other existing data on their activities such as their research output. If this is feasible, we can then link the researcher-output information back to researcher grants provided by Australian funding agencies such as the Australian Research Council (ARC) and the National Health & Medical Research Council (NH&MRC). Using this framework, our objective is to document the results of funded research grants across all domains of the Australian research landscape.

Our vision is that the data infrastructure which is at the heart of ASTRA will become an important part of the toolkit required to understand our returns to investment in science and research. It will achieve this by enabling systematic, rigorous evaluation of research policy implemented by governments. This evidence-based approach to policy making is common in other areas of the public policy space, but is new to the domain of research policy. This systematic evidence will perfectly complement the case-study evidence provided by the recent Excellence in Innovation for Australia (EIA) trial conducted by the ATN-Go8.

Given this, ASTRA is perfectly aligned with the National Research Investment Plan (NRIP) action item (#2.7) that "Investment should be subject to regular, rigorous and transparent monitoring and evaluation to assess efficiency and impact." Moreover, ASTRA's plan to link existing pieces of administrative data together – which are often held in different government agencies – in order to enable policy evaluation is consistent with NRIP's action item (#6) which states that "ARCom will provide advice on a whole-of-

government approach for opening access to the outputs and data from publicly funded research."

The ASTRA Feasibility Study concluded that:

- University data on grants can be repurposed to identify ***who is supported***. This includes not only chief investigators, but also individuals on research teams. These data can be used to describe both jobs and research links (Section 5).
- University data on grants can be repurposed to identify ***expenditures on vendors and sub-awards***. These data can be used to document the immediate economic impact of research funding (Section 5).
- The ANU and the University of Melbourne have ***sufficient data on grants*** to identify chief investigators and co-chief investigators, and to provide links to their research activities (Section 6).
- The ANU and University of Melbourne data have ***sufficient text information*** to describe ***what*** research is being done. The text information can be used to generate topic areas of research which could potentially be compared to other research activities internationally (Section 6).
- The data can also be used to describe ***where funded research*** is being done, both by the research institution and by collaborating institutions (Sections 5 and 6).

These data are also potentially robust enough to be visualized using both geographic and topical visualizations of the information. It is envisaged that this visualization – in the form of a new website called the ASTRA Portfolio Lens – will be an invaluable tool for governments, universities and industry who are interested in describing Australian research competencies and emerging areas of research capability (Section 8). Moreover, the data utilized in the ASTRA infrastructure can be automatically generated without burdening researchers, and can extend well beyond the period of research funding. Our ability to capture data without the need for additional reporting by researchers has been achieved in part by leveraging off the considerable investments made into the collection of data for the Excellence in Research for Australia (ERA) initiative.

Although the ASTRA project is only in its infancy, it has received very favorable responses from a broad range of stakeholders including State Government Departments, Deputy Vice Chancellors (Research) in leading universities, Chief Scientists and others (see Section 9). This indicates that there is substantial latent demand for a project like ASTRA. It is important to stress the international momentum on similar initiatives and Australia's potential leadership role: there is a program already in place in the US (which is called STAR METRICS), and similar projects have started (or are in the early stage of development) in France, the European Union, the Czech Republic, Spain and New Zealand.

For social scientists familiar with analytical frameworks and evidence-based public policy, ASTRA makes obvious sense. In particular, the quantity and quality of labour and social policy analyses in Australia has increased dramatically since the creation of the Household, Income and Labour Dynamics in Australia (HILDA) Survey back in 2001, which was funded FaHCSIA and managed by the Melbourne Institute. Thus, there are precedents for Commonwealth support of large-scale data infrastructure in the social sciences. Our belief is that ASTRA can have the same transformational effect of research policy in Australia.

1. Introduction

"How much should a nation spend on science? What kind of science? How much from private versus public sectors? Does demand for funding by potential science performers imply a shortage of funding or a surfeit of performers?.... A new "science of science policy" is emerging, and it may offer more compelling guidance for policy decisions and for more credible advocacy".

John H. Marburger III, "Wanted: Better Benchmarks" *Science*, May 20, 2005.

Jack Marburger – who was the Science Advisor to the US President George W. Bush from 2001 to 2009 – challenged the research community because his experience was that the existing state of measurement was inadequate to inform science and research policy. Jack identified the two key policy issues that have emerged internationally: evidence that provides compelling guidance for policy decisions (i.e. *management*) and permits more credible *advocacy*.

The need for evidence is common in all policy domains, and most policy fields have settled on well known indicators to inform policy. For example, labour policy relies heavily on the unemployment rate as a labor market indicator; economic policy on a plethora of measures, including the growth rate of GDP; and education policy makers rely on international mathematics and science scores as measured by the PISA test. These measures, while often imperfect, are widely used and the underlying theoretical justifications for their use are well known among policy makers. This, however, is not the case in research policy. There is a proliferation of metrics, often without a clear rationale for their choice other than that the data exist and metrics can be generated from them. In terms of *advocacy*, the stakeholders for whom they are generated do not understand what they mean, or what they are intended to convey. In terms of *management*, the existing metrics distort science: getting the right conceptual and empirical framework matters, lest resources and people get squandered because the incentives are wrong.

The emerging scientific approach is to recognize that research is fundamentally about the creation, transmission and adoption of ideas, *not* about counting documents. A number of countries have begun to build communities of practitioners and researchers who are focused on creating a more scientific basis for research policy. For example, the United States established both a White House Inter-agency Group of the major science agencies and the National Science Foundation established a research program called the Science of Science and Innovation Policy (SciSIP). Japan and Norway have now begun similar programs. The results have led to a US roadmap as well as an EU-US roadmap, hard evidence about a variety of different investments and the results of different program structures, and a wealth of additional insights from Science of Science and Innovation Policy researchers.

The overarching goal of the ASTRA program is to ensure that Australia is engaged in the global SciSIP movement, and possibly takes a leadership role. This Feasibility Study assesses how to build, within Australia, a consistent empirical framework and a community of practice that can work with the international SciSIP community. A robust SciSIP program enables policymakers to both better *advocate* for, and *manage* research, by:

1. Using a consistent, internationally recognized, evidence-based platform to assess the results of research investments;
2. Buildomg a theoretical and empirically-based understanding of the dynamic interaction between science, research and innovation; and
3. Developing a framework to describe the potential future outcomes of investments in research.

The Feasibility Study closely followed the tools and methods used to develop the STAR METRICS (Science and Technolgoy for America's Reinvestment: Measuring the Effect of Research on Innovation Competitiveness and Science) data platform in the United States. Leveraging off the substantial investments made in STAR METRICS has greatly enhanced – and lowered the costs of – the Feasibility Study. In particular, the ASTRA Feasibility Study assessed the viability of:

1. Building a tool – called the ASTRA Portfolio Lens – which presents a visualization of the number and topic of all government research grants using existing university data. This will be based on the STAR METRICS Portfolio Explorer;
2. Applying 'topic modeling' techniques – which utilises natural language processing – to describe who is doing what research in the universities, thereby mapping the contours of the Australian research community;
3. Constructing a workforce composition report – which describes the number, location and industry of jobs directly supported by government research grants – based on existing HR and Finance data at the two research institutions involved in the Feasibility Study (the University of Melbourne and ANU);
4. Setting up the data platform required to host ASTRA, including the need for (and cost of) a dedicated data server; and
5. Assessing the confidentiality and privacy considerations posed by ASTRA.

The Feasibility Study was complemented by extensive outreach to universities, industry partners and professional associations across Australia to garner input on the project, as well as by extensive engagement with the international community to ensure that ASTRA is consistent with the emerging international systems.

The Feasibility Study has benefited enormously from the contributions made by staff at the University of Melbourne (Liz Sonenberg, Simon Porter, Melissa Makin, Tim Baldwin, Jey Han Lau, Terry Nolan, Revathi Pendyala, Rodney Wheeler and John (JC) Campbell), the Australian National University (Margaret Harding, Paul Wong, Mark Matthews, Paul Harris, Irwan Krisna, Roslyn Markezic, Lorraine Piper and Ben Lees) and Geronimo Creative Services (Emma Stewart). We are also grateful for the input and advice from Roland Scollay (Chair of Scientific Advisory Council, Bio21) and the contributions made by the US team including Ron Lai, Dave Newman, Peter Rosenthal and Sonya Lai.

2. ASTRA's Theory of Change and Guiding Principles

The core theory of change upon which ASTRA is built is that research is done by researchers, and hence metrics should be based on describing human interactions and the results of those interactions, rather than counts of patents and publications. A sensible organizing framework has been provided by Ian Foster, and is illustrated in Figure 1. This identifies individual researchers (or the research community consisting of networks of researchers) as the 'engine' that generates ideas. Here, the theory of change

is that there is a link between funding and the way in which those networks assemble. Then, in turn, there is a link between research networks and the way in which those ideas are created and transmitted — and hence generate scientific, social, economic and workforce 'products'. These causal links are often so long and tenuous that tying results to individual grants, rather than to the networks that are supported by those grants is, quite simply, destined to fail.

The goal of ASTRA is to provide policy makers with a better understanding of the process of research and, hence, the capacity to provide better support for it. Indeed, the SciSIP community has been developing a body of knowledge about how to think about and identify those links, rather than just saying, as the cartoon would have it, 'that a miracle occurred'.

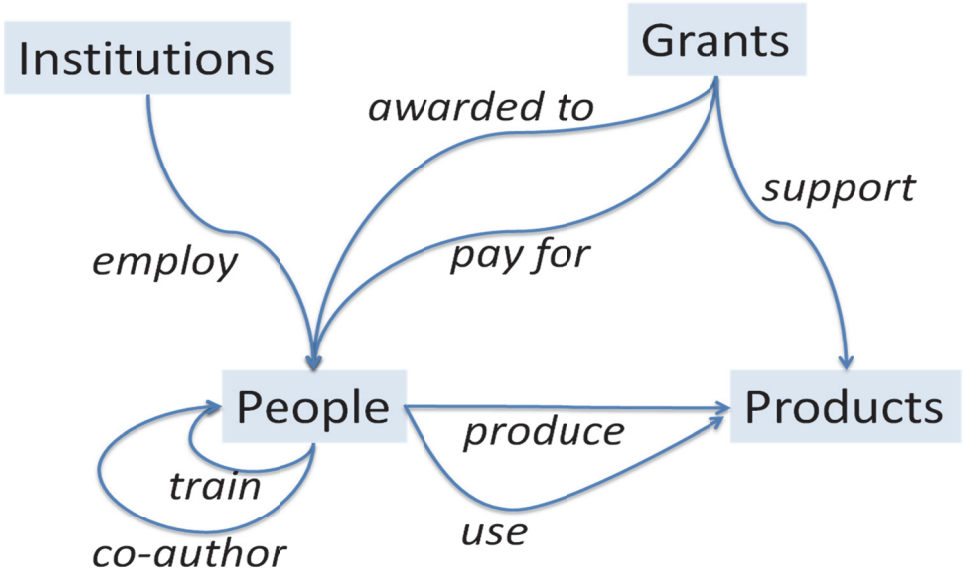


Figure 1: The Theory of Change

The ASTRA approach is thus fundamentally different from the bibliometrics which have come to dominate much of research measurement. The main goal is to build a flexible data infrastructure which can be used to describe the research process; research networks; and the role of funding in stimulating the creation, transmission and adoption of ideas through those networks.

Bibiliometric approaches have a great deal of value in many contexts and certainly contribute substantially to our understanding of the corpus of publications. However, their use to describe the performance of science is limited — indeed, they were not designed to do so. A recent article summarized a decade of study and asserted, "Evaluators often rely on numerically-based shortcuts drawn from the closely related fields (Hood and Wilson, 2001) of bibliometrics and scientometrics — in particular, Thompson Scientific's Journal Impact Factor (JIF). However, despite the popularity of this measure, it is slow (Brody and Harnad, 2005); narrow (Anderson, 2009); secretive and irreproducible (Rosner, et al., 2007); open to gaming (Falagas and Alexiou, 2008); and based on journals, not the articles they contain" (Priem and Hemminger 2010).

ASTRA is designed to develop more open and transparent measures of research activity and progress by capturing information on who is supported by research funding; what research is done, with whom and where; as well as the results. We discuss possible indicators that could be developed at the end of this report.

In order to guide the creation and development of ASTRA, we have formulated a series of guiding principles which outline the approach we plan to adopt. These principles are important

ASTRA is based on a scientific framework:

- Researchers do research: ASTRA is based on researchers being at the centre of the framework.
- Researchers understand science best: ASTRA engages the research community in identifying outputs and outcomes.
- Researchers can measure impact: Use statistical and econometric framework to quantify impact.

ASTRA uses modern technologies:

- Automated data collection: we rely on 21st century tools and techniques to collate and link existing data sources.
- Automated text analysis: rather than relying on manual classification schemes, ASTRA uses computational linguistics to characterize the subject of the scientific research.
- User-developed tools: The data infrastructure will be open source, so that users can developed their own tools and analyses.

ASTRA is part of an international activity:

- Common international pressures: ASTRA will build on the international interest and expertise in building evidence based policy and evaluation of research.
- Global research community: ASTRA will work to place Australian research in an international context.
- Global science of science policy community: ASTRA will work with the international community developing common data, models and tools.

3. National and International Context

In 2011–12 Australia invested almost $10bn on science, innovation and research. Yet, as has become increasingly apparent, there is little scientific basis for the way in which these investment decisions are being made or the *ex post* evaluation of investments once made. To make these sorts of decisions in a rigorous, analytical way requires a deeper understanding of the direct (and indirect) effects of research investments on all aspects of Australian (and global) society. This is not an easy task, and nor ASTRA being undertaken in a vacuum: there are a broad range of related projects going on, both domestically and internationally, that are relevant to ASTRA's development. In the

following sections we provide some national and international context to the development of ASTRA.

National

The Commonwealth Department of Innovation, Industry, Science, Research and Tertiary Education (DIISRTE) has commissioned a number of projects on the theme of 'research impact', of which ASTRA is just one. The others include:

1. The Excellence in Innovation for Australia (EIA) trial is a joint initiative of the Australian Technology Network of Universities (ATN) and the Group of Eight (Go8). The aims of the trial are to identify and demonstrate the contribution that high quality research has made to the economic, social, cultural and environmental benefit of society; and investigate the means by which these benefits may best be recognised, portrayed and assessed by institutions and government. Twelve universities are participating in the trial: five from the ATN (Curtin University; RMIT University; University of South Australia; University of Technology, Sydney; and Queensland University of Technology); four from the Go8 (The University of Melbourne, The University of New South Wales, The University of Queensland; and The University of Western Australia); Charles Darwin University; the University of Newcastle; and the University of Tasmania.
2. The Patent Analytics Pilot Study is developing and testing a variety of patent metrics for assessing the impact of university research. The metrics will be based on the mining, analysis, interpretation and visualisation of patenting activity by selected Australian universities and will include indicators of their collaboration performance with industry. The study is being undertaken by IP Australia's National Patents Analytics Hub. The Pilot Study is intended to continue the work undertaken by IP Australia in their 2009 university benchmarking study and to extend it for the purposes of impact assessment.

In addition, there are a range of other ongoing projects and reports which resonate with the framework underpinning ASTRA. For example, the National Research Investment Plan (NRIP) – which was developed by the Australian Research Committee (ARCom), under the auspices of the Chief Scientist of Australia, in response to the need for greater coordination of Australia's research investment – considers a broad range of strategic investment issues from human capital to infrastructure and collaboration. Following the release of a Discussion Paper in July 2012, the Government released the research investment plan on 28 November 2012, in which a series of 12 Action Items were announced.

The data infrastructure which is at the heart of ASTRA is designed to enable evaluation of research policy which is perfectly aligned with the NRIP action item (#2.7) that "Investment should be subject to regular, rigorous and transparent monitoring and evaluation to assess efficiency and impact." Moreover, ASTRA's plan to link existing pieces of administrative data together – which are often held in different government agencies – in order to enable policy evaluation is consistent with NRIP's action item (#6) which states that "ARCom will provide advice on a whole-of-government approach for opening access to the outputs and data from publicly funded research."

International

The United States is at the forefront of developments in the science of science and innovation policy (SciSIP). The recent SciSIP Principal Investigators' Conference – which was organised by the National Academy of Sciences and held in Washington DC (September 20–21, 2012) – illustrates the scale and quality of the SciSIP program in the US. There were around 50 presentations from leading scholars around the United States, including academics from MIT, Harvard, Boston and Berkeley. The two-day conference facilitated scholarly exchanges between SciSIP principal investigators – that is, those investigators funded by the NSF's SciSIP program – and was the largest gathering of SciSIP principal investigators since the program's inception (in 2006). Topics addressed included advances in the emerging SciSIP field, including models, frameworks, tools, and datasets comprising the evidentiary basis for science and innovation policy. Presentations focused on: implementing science policy (including the politics of science policy); scientific discovery processes; human capital; organizations, institutions, and networks; innovation; data extraction and measurement; mapping science; and assessment and program evaluation.

The international effort to produce interoperable data platforms that capture high quality, and continuously updated data on research spending, is continuing at pace. Related projects include the STAR METRICS program in the United States, the HELIOS project in France, the STAR METRICS program in Japan, the CAELIS project in the Czech Republic, the Science of Science policy program in Norway, the DESCRIBE program in the United Kingdom and nascent potential programs in Spain and New Zealand. There are also parallel activities in research management information systems; notably the European Union funded EuroCRIS project, the Canadian CASRAI project, and researcher networking activities like the VIVO project in the United States, the international ORCID effort, and the Japanese ResearchMap at the National Informatics Institute.

4. The U.S. Experience with STAR METRICS

The STAR METRICS program has very similar goals to ASTRA. The first goal (Level 1) is to document the levels and trends in the scientific workforce supported by federal funding. The second goal (Level 2) is to develop an open automated data infrastructure and tools that will enable the documentation and analysis of a subset of inputs, outputs, and outcomes resulting from federal investments in science. This section provides background information on the experience with STAR METRICS, in terms of current operational structure, linkage with universities and hosting the data server.

Current Operational Structure

There are two components to the STAR METRICS program. The federal component is run by a committee of six federal agencies (OSTP, NIH, NSF, DOE, USDA, EPA) with a separate loose consortium of about 100 research institutions which is organized by the Federal Demonstration Partnership (http://thefdp.org) with interest and support from the American Association of Universities, the Association of Public LandGrant Universities and the American Association of Medical Colleges. See Figure 2 for the Federal Structure of STAr METRICS and Figure 3 for a breakdown (by State) of the participating institutions.

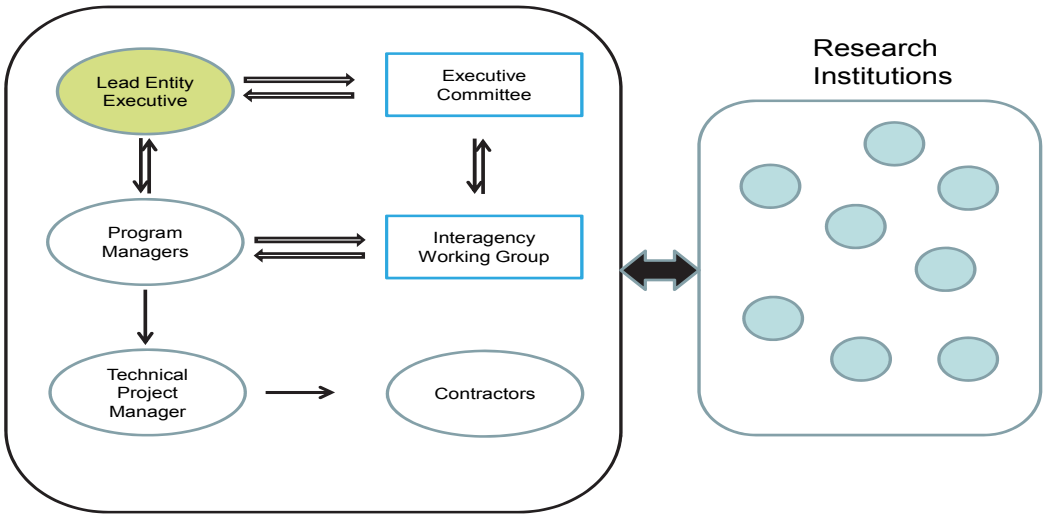


Figure 2: The Federal Structure (source: NIH presentation at FDP, January 28, 2013)

The plans for future work on the federal agency side[1] include implementing a statistical study to guide enrollment, evaluating data quality, evaluating job estimate calculations, developing a data policy and soliciting participation of minority serving institutions. The future steps include collecting grant data from the agencies, integrating and standardizing it and then loading it into a pilot web-based data system. About 76,000 projects are currently loaded in the database.

The future work on the university side[2] includes developing better benchmarking of occupational classifications. Another important focus is bringing Vice Presidents for Research, funding agencies, and the research community to build a more scientific basis for studying research and the research workforce. One example of that is a workshop being organized by Barbara McFadden Allen, the director of the Committee on Institutional Cooperation (http://www.cic.net/Home.aspx), Bruce Weinberg of Ohio State University, and Julia Lane.

[1] Source: George Chacko (STAR METRICS program manager) presentation at the FDP meeting, January 28, 2013.
[2] Source: Marietta Harrison and Jim English (FDP STAR METRICS representatives) presentation at the FDP meeting, January 28, 2013.

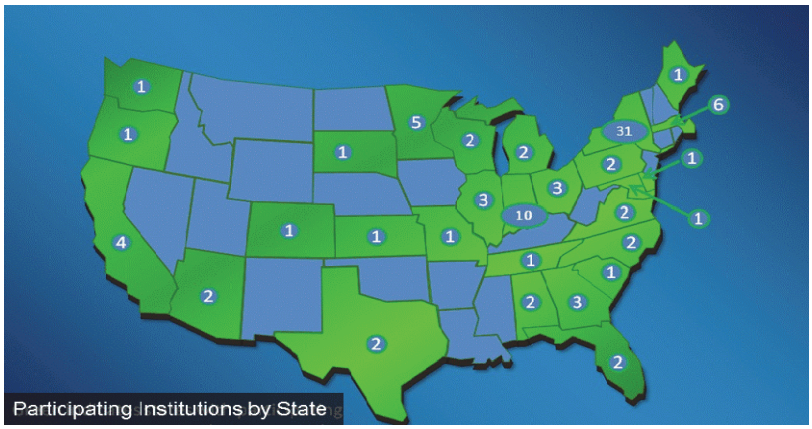


Figure 3: Participating Research Organizations (source: NIH presentation at FDP, January 28, 2013)

The goal of the meeting is to further inform the regional, national and international debate by developing open and transparent measures and methods to study the research enterprise and develop measures that can be used and trusted by policy makers. In order to develop that agenda, the workshop will brings together the Vice Presidents for Research of CIC institutions – who have been trying to develop practical measures – and a number of academic researchers in those institutions who have been developing data, models and tools in the academic arena. It is based on building out the STAR METRICS data that is being used by many of the CIC institutions[3].

Lessons Learned from Hosting Tools on Data Servers

In general, there are a number of overarching issues that need to be addressed: security, ownership, administration, cost and sustainability. In particular, the following sets of issues need to be addressed:

Security: Will data be secure, kept confidential, kept private, and be safe from hackers. Who has access to data? How is list of users having access controlled? What training do these users undergo? What legal issues exist around access to private data? Can users transfer confidential files onto own computers and/or public domain? For example, the processing of NSF proposal data required the installation of a server internal to NSF so that topic models could be built. A stand-alone Linux desktop/server was purchased and placed in the NSF machine room / data center. This server was ad-hoc situated in the machine room (i.e. not racked nor officially maintained). This server was connected, behind NSF firewall, to receive proposal data, as designed.

Ownership: Who owns the server? If the server located in-house, is server part of IT infrastructure being maintained? Is the server compliant with existing IT infrastructure?

[3] More information about this CIC initiative is provided in the beta-version website for the meeting at http://econ.ohio-state.edu/CIC. Username=starmetrics. Password: iC3mc26#.

Administration: Who can administer the server? If the server is located in-house, can contractors/non-employees have access? What happens when the server goes down? Is remote access possible? How is remote-access security guaranteed (e.g. VPN)?

Cost: Who is paying for the server? Who is paying for maintenance of the server?

*Sustainabilit*y: Who is responsible for keeping hardware and software operating?

There are also some operational lessons learned about public use of servers. There were two types of external servers used for STAR METRICS. The first was to provide for the upload of STAR METRICS Level I data[4] at a secure NIH site. Developing this site to be fully operational, and compliant with federal government rules and protocols cost about $750,000 and took almost one year.

The second was to host prototype tools for public comment and input. Two types of tools were hosted at the Networking and Information Technology Research and Development (NITRD) program's site: the R&D Dashboard, at the request of the White House CTO (Aneesh Chopra) and CIO (Vivek Kundra) and the NSF Portfolio Explorer. Each of these cost about $250,000 and took about three months. This approach was certainly useful in terms of getting valuable input, but both sites proved to be too popular and there were substantial stability issues with the server. For example, a graduate student at George Washington was attempting to scrape all the data from the NSF Portfolio Explorer for her dissertation which created new problems in server stability. As a platform increases in popularity, it requires more intensive infrastructure to deliver data. It is clear that a prototype should exist for a short amount of time before taken down to focus on a more production like system. A process should be well documented/tested so this requires balancing of execution vs. thorough process; ad hoc solutions are never advised especially with agencies that require a more thorough process.

In sum, there are really two sets of options.

Locate server in-house. This has a number of advantages: it is high security (based on in-house representatives), there is clear ownership of data and easier transitions for direct data exchange mechanisms. However, the burden of administration is placed on the server providers so training must happen on that side to prevent issues. Coordination must be taken on both sides.

There are also disadvantages. The server generally would have a higher cost and require a longer time to set up a physical machine. The platform sustainability depends on both coordination and learning.

External hosting of server (e.g. Amazon Web Services). The advantage is that the cost would generally be lower and it is faster to set up an external machine. In addition, the sustainability of the platform can be managed by the internal team, which is experienced and can solve problems more rapidly. The negative side is that this approach might offer lower security; this can change depending on the sort of information that is provided[5]. In addition, data coordination is likely to be more of a challenge: the administration burden is placed on the external server maintainer, and both sides need to coordinate activities.

[4] https://starmetrics.nih.gov

[5] For example, using cloud services, access can be restricted to specific IP addresses.

5. Who is Supported by Research Funding?

The following are the key findings of the workforce reports:

- University data on grants can be repurposed to identify ***who is supported***. This includes not only chief investigators but also individuals on research teams. These data can be used to describe both jobs and research links at the project level.
- University data on grants can be repurposed to identify ***expenditures on vendors and sub-awards***. These data can be used to document some of the immediate economic consequences of research funding.

Two workforce reports[6] were generated: one for the University of Melbourne; and the other for the Australian National University. In each case, these data were generated at relatively low cost (see Appendix A8) and were of reasonably high quality[7].

The systems used are similar to those used at other Australian institutions: A comprehensive list of Finance and HR systems for the sector is included in Appendix A11. This was a survey conducted by Higher Ed Services in 2008 (http://www.hes.edu.au/). There may have been some changes since 2008, but in all likelihood most of the systems listed are still currently in use by the corresponding universities.

There is no in principle or technical problem with scaling ASTRA project up at the national (and even international) level – every university in Australia uses similar systems and relational database technologies in their business (HR, Finance, Research etc) – these are industry standards with common query and programming languages and interfaces. But the internal business practices and procedures in data collection and system usages can vary from university to university. This is true even though accounting standards may be the same for the whole sector. So for ASTRA to be scalable to the national scale, it is critical that partnership with local experts from within each participating institution must be established to ensure knowledge of the local condition and practice is captured in the work.

University of Melbourne

The University of Melbourne uses an industry standard Human Resources and Finance system (Oracle) across the whole university. In accordance with audit requirements, grants are managed within separate account codes, with salary for supported staff expensed directly against the account.

The University of Melbourne workforce report provides information about the research workforce directly supported by expenditures associated with all research grants (aside from block research grants) made to the University of Melbourne for calendar year 2011. In other words, the workforce reports are compiled using data from all competitive grant

[6] The complete workforce reports are provided separately. A simple summary is provided here.

[7] There are a number of examples where the tracking of externally funded research expenditure can be an issue. These include the following examples: Bulk billing by a vendor (e.g. office consumables, or water or electricity), a single line amount can be recorded on the bill – determining how much is charged against a particular research/grant project can be difficult. Credit card payment/cash/reimbursement of research expenditure don't usually have detailed vendor information recorded – as shown in the data supplied some expenditure are recorded against "vendors" with university IDs as reimbursement of funds to individuals within the university.

schemes – such as ARC Linkage, Discovery and Linkage Infrastructure grants plus the NH&MRC Project and Program grants – and other research grants awarded by agencies such as the Cancer Council, Dairy Australia and the National Heart Foundation. The breakdown of the $200m in research funds is: ARC Linkage (6%), ARC Discovery (8%), ARC Other (4%), NH&MRC Program (17%), NH&MRC Project (20%), NH&MRC Other (6%) and All Other (39%).

The workforce information includes data about the occupational distribution of the workforce, the number of individuals supported and the number of full-time equivalent employees. Estimates are also provided for the total number of jobs related to expenditures by those institutions at collaborating institutions, and of vendors who provide support to those grants – for example supplying research services. The estimates are for jobs directly generated by research awards at the University of Melbourne as a result of awards active in 2011. They do not include estimates of multiplier effects.

The key findings include:

- University of Melbourne researchers were supported by 1,792 active awards. In addition to the Chief Investigators (whose salaries are not included), these grants supported:
    - 2,699 individuals or 1,044 FTE positions;
    - Over 100 postdoctoral fellows or research assistants.
- Almost $200 million in external expenditures were generated from the awards. These expenditures:
    - were spent on 1,378 external vendors;
    - generated 1,235 jobs throughout Australia;
    - were spread throughout Australia, primarily in Victoria and New South Wales.

The Australian National University

ANU data were well structured within standard relational database (Oracle DB) with appropriate data model underpinning the data. The ANU uses PeopleSoft for both Finance and HR (version 9.1). An internal challenge was tracking internal journal transfers between different ledgers (R recurrent vs S sponsored research) associated with academics – this is done as a matter of administrative convenience, but it can mask salary cost (and FTE) of research expenditure. The ANU has a high success rate in securing external (e.g. ARC) funding and high level of continuing academic appointment, the use of journal transfers between R and S ledger can thus be common. This is likely to mean that head count and FTE figures based on data supplied under estimate the true figures.

The Australian National University workforce report provides information about the research workforce directly supported by expenditures associated with ARC and NH&MRC grants made to the Australian National University for the first quarter of 2012. The workforce information includes data about the occupational distribution of the workforce, the number of individuals supported and the number of full-time equivalent employees. Estimates are also provided of the total number of jobs related to expenditures by those institutions at collaborating institutions, and of vendors who provide support to those grants. The estimates are for jobs directly generated by research

grants at the Australian National University as a result of grants active in the first quarter of 2012. They do not include estimates of multiplier effects.

The key findings include:

- Australian National University researchers were supported by 637 active grants. In addition to the Chief Investigators (whose salaries are not included), the grants supported:
    - 1,065 individuals or 726 FTE positions;
    - Over 260 postdoctoral fellows or research assistants.
- Over $18 million in external expenditures were generated from the grants. These expenditures:
    - were spent on 624 external vendors;
    - supported 758 jobs throughout Australia in that quarter;
    - were spread throughout Australia, primarily in the Australian Capital Territory, Victoria and New South Wales.

6. Research Collaborations and Outputs

The key finding from our analysis of the data is that both the ANU and the University of Melbourne have sufficient data on grants to develop an analytical data platform that links funding to researchers and their research activities. This approach reduces the need for manual reporting: the links between funding and the associated research can be generated through scientists and their scientific collaborations. In other words, ANU and the University of Melbourne have sufficient data on grants to identify chief investigators and co-chief investigators, and provide links to their research activities.

In particular:

1. The grant data are comprehensive: when the university data were compared to official benchmarks,[8] the correspondence ranged between 95 and 100 per cent.
2. There is sufficient text data on grants to topic model the project descriptions.
3. It is possible to uniquely identify chief investigators and co-chief investigators.
4. It is possible to provide links to researchers' publication activities.
5. There is sufficient data to topic model publications.

The conceptual framework described in Figure 1 informs the development of the ASTRA data schema (Figure 2). At the core is the individual researcher. In principle, all research activities, such as publications, patents, research collaborations, and student training can be linked through scientists' activities. For the Feasibility Study, we focus on an abbreviated set of data elements, namely linkage and meta-data on scientists, grants and their publications. We further restrict the analysis to grants from the Australian Research Council (ARC) and National Health and Medical Research Council (NH&MRC) from the period January 1, 2007 to June 30, 2012 to maintain data integrity (as bookkeeping methods have changed over the years).

[8] We compare these figures to the official NH&MRC and ARC Australian Government funding sources: NH&MRC (http://www.NH&MRC.gov.au/grants/outcomes-funding-rounds); ARC Discovery (http://arc.gov.au/ncgp/dp/dp_outcomes.htm); and ARC Linkage (http://arc.gov.au/ncgp/lp/lp_outcomes.htm).

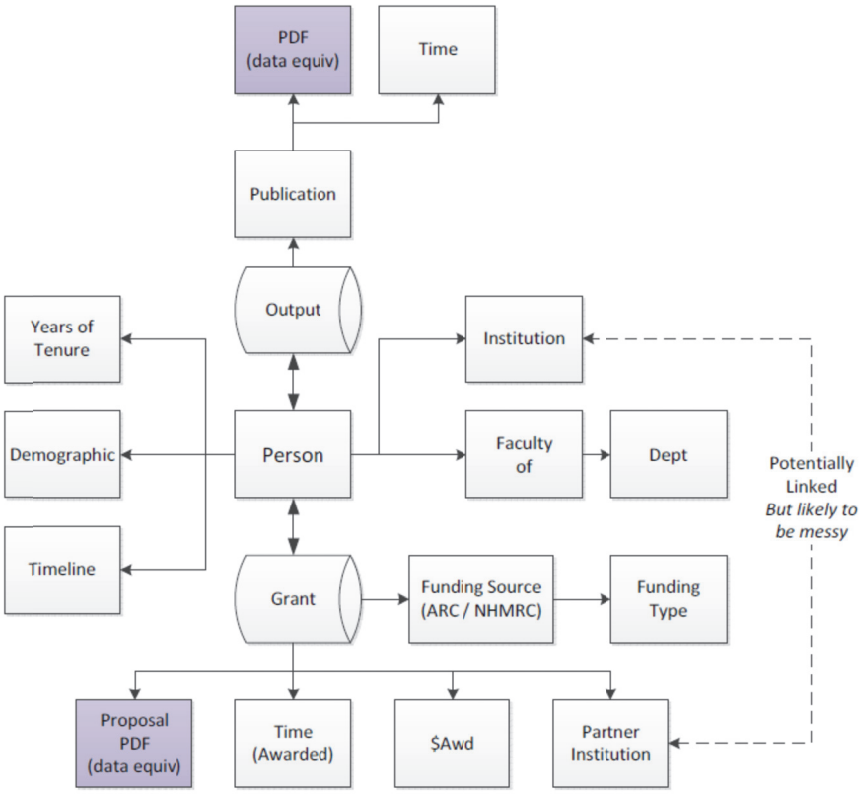


Figure 4: ASTRA Data Schema

Each institution provided data covering the period January 1, 2007 to June 30, 2012, for ARC Linkage, ARC Discovery and NHMRC Project Grants. However, the samples differ slightly, as detailed in Table 1.

Table 1: Sample Coverage

	Australia National University	University of Melbourne
Grants	* HERDC-eligible * does not have status of: "draft", "submitted", "awarded", or "offer of award" * non-externally led * "cash" funds payment type	* awarded project grants * no contract grants * non-externally led
Publications		* internally marked HoD certified * ERA 2010 or 2012

Australian National University

Research Grant Data

We began by analyzing ANU HERDC (Higher Education Research Data Collection) eligible grants for the period 2007 to 2011. Table 2 provides detail of the number of grants and the data within the grants over the period in question. The grant totals are consistent with published totals. Almost all have text data that can be analyzed; almost all have information about award amounts.

Table 2: ARC and NH&MRC Research Grants, ANU							
Year	Number of Grants			Dollars Awarded		Chief Investigators	
	ARC	NH&MRC	Total	Mean (thousands)	Total (millions)	Avg # CIs	Avg # CIs, Internal
2011	105	21	126	$406.88	$49.23	2.635	1.667
2010	113	20	133	$403.40	$52.85	3.023	1.805
2009	114	13	127	$375.11	$47.26	2.669	1.709
2008	94	22	116	$397.26	$45.29	2.543	1.724
2007	125	15	140	$383.99	$53.76	2.257	1.586
2007–11	**551**	**91**	**642**	**$393.02**	**$248.39**	**2.623**	**1.696**

The first column, **Year**, shows the data by year. The last row, **2007–11**, shows the information across the columns for the complete subset explicitly defined in the *Data Assumptions* section.[9] The **Number of Grants**, reported in columns 2–4, shows the counts for unique grants. **ARC** (col. 2) shows the figures for all Australian Research Council Linkage and Discovery grants. **NH&MRC** (col. 3) shows the figures for all National Health and Medical Research Council Project Grants. The **Total** column (4) shows all the grants received in the subset (therefore it is just the sum of **ARC** and **NH&MRC** columns).

Dollars Awarded, reported in columns 5–6, shows information about the grant dollar amounts. **Mean (thousands)** (col. 5) shows the average dollar amount awarded for each grant in thousands. **Total (millions)** (col. 6) signifies the total amount of these awards in millions.[10] **Chief Investigators** (cols 7–8) shows figures on chief investigator (CI) data. Column 7, **Avg # CIs**, and column 8, **Avg # CIs, Internal**, both show the average number of chief investigators per grant. The only difference is **Avg # CIs, Internal** only includes people internal to ANU, whereas **Avg # CIs** contains internal, external, and those not defined as either.

Chief Investigator Data

Since the focus of the ASTRA project depends on being able to capture data on individuals, it was critical to be able to tease out information on chief investigators and co-chief investigators.

Table 3 documents the number of unique investigators by year,[11] and is the basis for finding links between a person's activities and collaborations. The first three columns provide the basis for generating information about internal collaborations at the project level. **Primary CIs, Internal** (col. 2) shows the number of unique CIs who are internal to

[9] This will be a consistent template across all the data tables. Since the funding scheme with the grant year is stated, we use those figures instead of start date for ANU's data analysis. Usually the start date and grant year correspond, but not always.

[10] Please note, because a few dollar amounts are missing, it is not accurate to multiply the total number of grants (col. 4) by the average value of the awards (col. 5) to reach the total amount (col. 6).

[11] Note that this is different from the number of chief investigators on each grant, since individual investigators can work on multiple grants in the same year

ANU and are the primary person on a grant. **Co-CIs, Internal** (col. 3) reflects the number of people internal to ANU who are not primary CIs on a grant. **All CIs, Internal** (col. 4) shows the number of unique internal CIs, including both categories. This column is not the summation of the previous two, as someone could be a primary CI on a particular grant and a co-CI on another (within the same year). **All CIs, External** (col. 5) is the basis for documenting information about external collaboration at the project level; it shows how many external CIs are part of the grants. **All CIs** (col. 6) shows the total number of unique CIs. It includes all the internal (col. 4) and external CIs (col. 5), as well as those not marked as "internal" or "external" such as students.

Table 3: Number of Unique Chief Investigators, ANU

Year	Primary CIs, Internal	Co-CIs, Internal	All CIs, Internal	All CIs, External	All CIs
2011	117	84	191	116	307
2010	125	99	217	156	378
2009	121	86	197	117	316
2008	108	76	177	86	267
2007	133	79	207	86	295
2007–11	**444**	**347**	**676**	**524**	**1213**

Many different views of research can be generated with these individual level data in addition to linking them to research activities. For example, the link to grants can be used to show which researchers have generated the most grants; the link to other researchers can be used to show who is most connected with other researchers; the temporal information can be used to show which co chief investigators evolve into chief investigators; and the geographic information can be used to examine intramural and extramural collaboration patterns.

Table 4: Grant Information per Chief Investigator, ANU

Year	Avg # Grants per Unique CI			Avg $ per Grant per Unique CI (000)		
	Primary CIs, Internal	All CIs, Internal	All CIs	Primary CIs, Internal	All CIs, Internal	All CIs
2011	1.077	1.100	1.081	$392.45	$424.73	$429.03
2010	1.064	1.106	1.064	$391.50	$431.08	$436.77
2009	1.050	1.102	1.073	$370.76	$392.85	$405.23
2008	1.074	1.130	1.105	$394.74	$427.65	$431.12
2007	1.053	1.073	1.071	$382.40	$465.46	$455.32
2007–11	**1.446**	**1.611**	**1.078**	**$383.97**	**$428.90**	**$432.32**

Table 2, columns 7–8, from the previous section showed CIs in terms of grants. But the data also allows us to show grants in terms of CIs. This is showcased in Table 4 (later in the *publications* section, we will show another way using CI data as a link, thus, reinforcing its multidimensional nature).

Table 4 is broken down into TWO sections: **Avg # Grants per Unique CI** (cols 2–4) and **Avg $ per Grant per Unique CI (thousands)** (cols 5–7).

1. **Avg # Grants per Unique CI** gives the average number of grants.
2. **Avg $ per Grant per Unique CI (thousands)** gives the average amount of grant dollars in thousands.

Each section has grant information in terms of: a) **Primary CIs, Internal** (cols 2,5); b) **All CIs, Internal** (cols 3,6); and c) **All CIs** (cols 4,7).

a. **Primary CIs, Internal** shows the figures based on only those people internal to ANU who acted as the primary CI on a grant.
b. **All CIs, Internal** are those people internal to ANU, with no distinction made as to whether the person is a primary CI or co-CI.
c. **All CIs** reflect all chief investigators on the grants, including internal, external, and students.

The **Average number of Grants per Unique CI**'s in the last row (**2007–**11) is of particular interest, since it reflects the five-year concentration of grants across the same individual.

Publications Data

A major goal of the project is to establish the feasibility of providing a connection between grants to researchers (input) and subsequent publications (output). The data in Table 5 provide information about publications produced by CIs subsequent to having received a grant.[12]

Table 5: Number of Publications Subsequent to Grants, ANU

Pub Year	2010 Grants				2009 Grants				2008 Grants				2007 Grants			
	Jour Article	Book	Book Chapt	Conf Paper	Jour Article	Book	Book Chapt	Conf Paper	Jour Article	Book	Book Chapt	Conf Paper	Jour Article	Book	Book Chapt	Conf Paper
2011	683	10	64	174	615	9	61	138	490	5	60	70	605	5	66	147
2010					672	11	88	107	437	8	80	73	627	6	62	98
2009									501	11	71	98	656	7	70	126
2008													633	3	85	98
Summary	**683**	**10**	**64**	**174**	**1287**	**20**	**149**	**245**	**1428**	**24**	**211**	**241**	**2521**	**21**	**283**	**469**

[12] Note that no causal assumptions are made directly linking grants to publications. The ASTRA framework is structured to describe research done by a CI subsequent to receiving grants. This is consistent with how basic research is done: funding typically supports an entire research agenda, rather than more applied, or contract research, where funding is directly targeted at producing a specific result.

Table 5 provides a breakdown by year (**Pub Year**, col. 1) of the number of unique publications for ANU for each publication type: Book (A1), Book Chapter (B1), Journal Article (C1) and Conference Paper (F1).[13] The table has four sections and provides information about publications subsequent to the year grants are awarded, ranging from 2007–2010. Since these are counts of unique publications and each has a unique publication date, the **Summary** row (5) is the sum of all the rows above it. The first section is titled **2010 Grants** (cols 2–5). In **2011** (row 2) there were 683 journal articles (col. 2); 10 books (col. 3); 64 book chapters (col. 4); and 174 conference papers (col. 5) published from people who won grants in 2010. The **Summary** row shows all the publications by type for the relevant years from 2008–11. Since we are only comparing publications subsequent to grants, publications from 2011 are only compared to 2010 grants. Moving to the next section, the **Summary** row of **2009 Grants** (cols 6–9) includes publication figures for 2010–11.

Table 5 highlights some well-known patterns in publication types. Journal articles (**Jour Article,** cols 2, 6, 10, 14) make up the majority of publications. Then conference papers (**Conf Paper**, cols 5, 9, 13 and 17) and book chapters (**Book Chapter**, cols 4, 8, 12 and 16) follow. Books (cols 3, 7, 11, 15) make up the least.

Table 6 normalizes on individuals: it shows the average number of unique publications per unique grant (from unique individuals). These are the same figures from Table 5 but divided by the number of grants. For example, the first section, **2010 Grants** (cols 2–5), first row **2011** shows the breakdown by publication type: "5.84" in the **Jour Article** column (2) means that the mean number of journal articles produced in 2011 by researchers funded on a given grant in 2010 was 5.84 journal articles; the '1.25' in the **Book** column (3) has a similar interpretation. The **Summary** row (5) shows the figures for the relevant publication dates from 2008–11.

Table 6: Average Number of Publications per Unique Grant, ANU

Pub Year	2010 Grants				2009 Grants				2008 Grants				2007 Grants			
	Jour Article	Book	Book Chapt	Conf Paper	Jour Article	Book	Book Chapt	Conf Paper	Jour Article	Book	Book Chapt	Conf Paper	Jour Article	Book	Book Chapt	Conf Paper
2011	5.84	1.25	1.94	6.21	5.30	0.90	2.03	4.60	4.90	1.25	2.50	4.12	5.82	1.25	2.06	6.13
2010					5.84	1.10	2.15	3.82	4.41	1.14	2.16	3.65	6.46	1.00	1.63	4.90
2009									4.86	1.22	2.22	4.45	5.86	1.17	1.89	5.48
2008													5.70	1.00	2.07	3.92
Summary	**5.84**	**1.25**	**1.94**	**6.21**	**10.46**	**1.25**	**2.76**	**6.28**	**12.98**	**1.50**	**3.98**	**7.77**	**20.33**	**1.40**	**3.77**	**12.34**

While Table 6 provided grant level detail, Table 7 provides a breakdown (by year) by the people who generated the grants.

[13] Please note journal article figures for Tables 4–7 include external articles with internal ANU authors.

Table 7: Number of Unique Internal Authors, ANU																
Pub Year	2010 Grants				2009 Grants				2008 Grants				2007 Grants			
	Jour Article	Book	Book Chapt	Conf Paper	Jour Article	Book	Book Chapt	Conf Paper	Jour Article	Book	Book Chapt	Conf Paper	Jour Article	Book	Book Chapt	Conf Paper
2011	176	7	48	40	162	8	35	36	140	4	33	19	135	4	32	27
2010					163	10	48	36	135	9	46	23	133	5	38	22
2009									147	10	41	28	142	7	35	27
2008													140	3	43	29
Summary	**176**	**7**	**48**	**40**	**180**	**16**	**66**	**49**	**166**	**20**	**72**	**37**	**163**	**15**	**80**	**43**

Table 7 has the same format as Tables 5–6: the publication years are shown in the first column (**Pub Year**) with a **Summary** row (5) reflecting figures for the range of all relevant years (publications subsequent to grant year). In the **2010 Grants** section (cols2–5), "176" in the upper left corner refers to 176 unique internal ANU authors who won grants in 2010 and published journal articles in 2011. The '7' in the column (3) next to it refers to the number of unique internal ANU authors who won grants in 2010 and published in 2011. As before there are more unique authors for journal articles than for any other type of publication, with book authors the least represented of all.[14]

Table 8 sheds light on the funding per publication type. It expands on Table 7 by taking the total of all the award amounts each internal author has received from grants and averages them by the number of unique internal authors. This information is broken down by grant year and publication year. The **Summary** row (5) reflects the relevant range of years from 2008–2011 for each grant year.[15]

Table 8: Average Amount ($000) Received by Internal Authors who Received Grants, ANU																
Pub Year	2010 Grants				2009 Grants				2008 Grants				2007 Grants			
	Jour Article	Book	Book Chapt	Conf Paper	Jour Article	Book	Book Chapt	Conf Paper	Jour Article	Book	Book Chapt	Conf Paper	Jour Article	Book	Book Chapt	Conf Paper
2011	507.8	602.5	373.8	468.2	456.6	386.8	435.5	429.7	569.1	304.9	845.3	555.4	463.8	199.4	409.5	439.8
2010					458.6	608.5	498.2	466.7	563.1	387.6	642.5	442.0	451.3	231.0	473.8	451.6
2009									542.1	371.9	730.6	583.5	458.5	303.5	400.1	444.4
2008													449.9	373.6	455.0	550.6
Summary	**507.8**	**602.5**	**373.8**	**468.2**	**450.1**	**475.5**	**468.1**	**458.6**	**554.0**	**356.4**	**620.1**	**547.6**	**441.1**	**291.0**	**457.8**	**482.8**

[14] It is worth noting that the **Summary** row cannot be summed across the rows as the same author may be reflected in multiple rows. For example, an author may have won at least one grant in 2007, and have published in both 2008 and 2010, and would be counted as a unique author in each of those years. Furthermore, that author may have released different types of publications such as a book and a journal article. Therefore, the rows cannot also be summed across publication types.

[15] Please note that for a grant with multiple people attached, each person was credited with the full dollar amount for Table 8; we did not divide grant amounts by the number of people weighted by their role, but this would be possible for further analysis.

University of Melbourne

For the University of Melbourne the data covered 2000–2012 (for grants) and 2007–2012 (for publications); for consistency purposes, we focused on the period 2007–2012. Please note, slightly differently from ANU which explicitly assigns the grants to a particular date, for UniMelb we use the year in start date as a proxy for grant year.[16]

Research Grants Data

In Table 9 research grants are divided into NH&MRC and ARC (Linkage and Discovery) grants. To synchronize the grants to only those types included in the benchmark, we consider awarded project grants only (therefore all non-awarded and contracts grants are removed). Also these figures do not include any non-lead or collaborative grants.

The columns broken down by **Year** (col. 1) are defined as follows:

Number of Grants

a. **ARC** (col. 2): The number of unique ARC Linkage and Discovery Grants.
b. **NH&MRC** (col. 3): The number of unique NH&MRC Grants.
c. **Total** (col. 4): The total number of grants in our sample; this is the sum of columns **ARC** and **NH&MRC** since we only consider those grant types.

Dollars Awarded

a. **Mean (thousands)** (col. 5): The average amount awarded per grant.
b. **Total (millions)** (col. 6): The total dollar amounts awarded for the grants.

Chief Investigators

a. **Avg # CIs** (col. 7): The average number of chief investigators per grant.
b. **Avg # CIs, Internal** (col. 8): The average number of chief investigators per grant who are internal to UniMelb.

Table 9: ARC and NH&MRC Research Grants, University of Melbourne

Year	Number of Grants			Dollars Awarded		Chief Investigators	
	ARC	NH&MRC	Total	Mean (thousands)	Total (millions)	Avg # CIs	Avg # CIs, Internal
2012	105	120	225	$446.09	$100.37	3.627	2.797
2011	140	99	239	$460.20	$109.99	3.686	2.485
2010	154	83	237	$421.32	$99.85	4.228	2.694
2009	153	93	246	$478.64	$117.75	3.996	2.476
2008	147	125	272	$402.03	$109.35	2.824	2.038
2007	146	89	235	$350.00	$82.25	3.060	2.332
2007–12	**845**	**609**	**1454**	**$426.11**	**$619.56**	**3.622**	**2.507**

[16] These tables follow the same templates as the ANU tables in the previous section. Please refer to those tables for further explanations.

Chief Investigator Data

Tables 10 and 11 show data on chief investigators.[17] Table 10 gives the counts of unique chief investigators, broken down by **Year** (col. 1) and CI types. **Primary CIs, Internal** (col. 2): Numbers of primary chief investigators internal to UniMelb. **Co-CIs, Internal** (col. 3): Number of co-chief investigators (no primary CIs) internal to UniMelb. **All CIs, Internal** (col. 4): Number for all CIs internal to UniMelb. These numbers do not necessarily sum up across columns 2 and 3 as there are some CIs defined otherwise.

Table 10: Number of Unique Chief Investigators, University of Melbourne

Year	Primary CIs, Internal	Co-CIs, Internal	All CIs, Internal
2012	195	240	406
2011	205	225	409
2010	220	205	405
2009	226	213	415
2008	241	191	411
2007	200	184	374
2007–12	**843**	**904**	**1423**

Table 11 provides information about the number of grants per unique internal chief investigator. The average number of grants per unique CI (**Avg # Grants per Unique CI**, cols 2–3) and the average dollar amount of those grants (**Avg $ per Grant per Unique CI (thousands)**, cols 4–5) are explored across both primary internal CIs (**Primary CIs, Internal**, cols 2,4) and all internal CIs (**All CIs, Internal**, cols 3, 5) for UniMelb. Please note the last row, **2007–12**, may show figures greater than expected. This is because the number of grants and average dollar amounts may be "diluted" when looking at small ranges of dates such as one year intervals. One person may have received 5 grants across 2007–2012, and hence would average one a year, but when considering the entire range, s/he would be noted as receiving 5 grants. This is similar for dollar amounts.

Table 11: Grant Information per Chief Investigator, University of Melbourne

Year	Avg # Grants per Unique CI		Avg $ per Grant per Unique CI (000)	
	Primary CIs, Internal	All CIs, Internal	Primary CIs, Internal	All CIs, Internal
2012	1.1128	1.207	$496.02	$580.62
2011	1.1024	1.154	$508.39	$561.19
2010	1.0864	1.138	$451.63	$499.54
2009	1.0841	1.140	$517.51	$603.56
2008	1.0871	1.141	$445.66	$477.35
2007	1.085	1.126	$394.11	$439.83
2007–12	**1.636**	**1.914**	**$715.88**	**$898.62**

[17] Since we did not receive id numbers for external CIs, the analysis only includes non-external CIs. Therefore, these tables differ from their ANU counterparts.

Publication Data

Tables 12–15 use publication data from linking grants to publications through the same people: Grants -> CIs = Authors -> Publication. Only data originating from grants through their related internal CIs (since we can only match through internal id numbers) then matched to authors (with the same ids) with publications subsequent to the grant year are considered. For example, a grant from 2009 will only have figures for publications after 2009, i.e. 2010–2012. The **Summary** rows (6) reflect data only for the years from 2008–2012 that are relevant to the grant year listed.

Table 12 shows counts for all unique publications by year (2008–2012) by publication type for each grant year (blocked off into sections from 2007–2011). As with the ANU, journal articles make up an overwhelming majority of all publications and books make up the least. The **Summary** row (6) is the sum for the column across all the relevant years. Table 13 shows the figures from Table 12 divided by the number of unique grants. As in Table 12, a greater number of journal articles are produced per grant than books. Instead of publication figures in terms of unique grants, Tables 14–15 are presented in terms of unique internal authors (the same people as CIs for the related grants). Table 14 gives the counts of unique internal authors for the University of Melbourne by publication year and publication type for each grant year. Finally, Table 15 links these tables to show funding by publication type. Using the results from Table 14, Table 15 presents the total dollar value of grants for a particular year received by each unique internal author as an average.[18]

Table 12: Number of Publications Subsequent to Grants, University of Melbourne

Pub Year	2011 Grants				2010 Grants				2009 Grants				2008 Grants				2007 Grants			
	Jour Art	Book	Book Ch	Conf Pap	Jour Art	Book	Book Ch	Conf Pap	Jour Art	Book	Book Ch	Conf Pap	Jour Art	Book	Book Ch	Conf Pap	Jour Art	Book	Book Ch	Conf Pap
2012	253	1	13	17	351	1	10	11	263	0	17	21	342	2	14	16	280	1	10	9
2011					1414	4	104	209	1448	10	116	268	1427	5	99	166	1258	5	110	175
2010									1352	17	108	253	1344	14	93	172	1213	12	95	162
2009													1344	10	112	217	1144	13	157	188
2008																	1118	11	119	206
Summary	**253**	**1**	**13**	**17**	**1765**	**5**	**114**	**220**	**3063**	**27**	**241**	**542**	**4457**	**31**	**318**	**571**	**5013**	**42**	**491**	**740**

Table 13: Average Number of Publications per Unique Grant, University of Melbourne

Pub Year	2011 Grants				2010 Grants				2009 Grants				2008 Grants				2007 Grants			
	Jour Art	Book	Book Ch	Conf Pap	Jour Art	Book	Book Ch	Conf Pap	Jour Art	Book	Book Ch	Conf Pap	Jour Art	Book	Book Ch	Conf Pap	Jour Art	Book	Book Ch	Conf Pap
2012	2.06	1.00	1.30	1.55	2.68	0.25	0.67	0.73	2.05	0.00	0.77	1.62	3.03	0.67	0.88	1.33	2.86	1.00	0.83	1.29
2011					5.05	0.80	1.03	2.71	5.23	1.11	1.49	4.00	5.51	0.71	1.46	3.02	5.47	0.71	1.53	2.78
2010									4.79	0.71	1.27	3.37	5.13	0.82	1.69	3.44	5.21	0.86	1.48	2.79
2009													5.07	1.00	1.53	3.39	4.91	0.93	1.76	3.19
2008																	4.86	0.92	1.49	3.43
Summary	**2.06**	**1.00**	**1.30**	**1.55**	**6.24**	**0.56**	**1.11**	**2.82**	**10.3**	**0.84**	**1.93**	**6.23**	**15.7**	**1.03**	**2.56**	**7.42**	**19.7**	**1.14**	**3.25**	**7.79**

[18] For grants with multiple people attached, the full dollar amount of the grant is attributed to each person.

Table 14: Number of Unique Internal Authors, University of Melbourne																			
Pub Year	2011 Grants				2010 Grants				2009 Grants				2008 Grants				2007 Grants			
	Jour Art	Book	Book Ch	Conf Pap	Jour Art	Book	Book Ch	Conf Pap	Jour Art	Book	Book Ch	Conf Pap	Jour Art	Book	Book Ch	Conf Pap	Jour Art	Book	Book Ch	Conf Pap
2012	131	1	11	11	120	1	9	8	115	0	13	10	130	2	12	9	93	1	8	7
2011					337	4	69	75	335	10	71	67	315	5	57	60	282	5	66	59
2010									342	16	79	69	334	14	51	56	299	11	66	52
2009													326	10	74	70	300	15	89	59
2008																	285	11	80	57
Summary	**131**	**1**	**11**	**11**	**345**	**5**	**73**	**76**	**370**	**25**	**121**	**91**	**379**	**28**	**127**	**88**	**355**	**36**	**165**	**106**

Table 15: Average Amount (000) Received by Internal Authors who Received Grants, University of Melbourne																				
Pub Year	2011 Grants				2010 Grants				2009 Grants				2008 Grants				2007 Grants			
	Jour Art	Book	Book Ch	Conf Pap	Jour Art	Book	Book Ch	Conf Pap	Jour Art	Book	Book Ch	Conf Pap	Jour Art	Book	Book Ch	Conf Pap	Jour Art	Book	Book Ch	Conf Pap
2012	551	182	366	493	554	1942	676	736	594	0	666	804	520	210	482	594	484	164	418	474
2011					531	360	530	532	627	414	664	524	499	394	402	449	462	269	366	380
2010									615	535	688	627	499	347	426	450	457	381	311	339
2009													500	235	382	451	458	286	319	371
2008																	443	345	320	387
Summary	**551**	**182**	**366**	**493**	**528**	**677**	**549**	**528**	**612**	**481**	**653**	**575**	**489**	**301**	**432**	**433**	**446**	**318**	**343**	**356**

7. What Research is being Done? Topic Modeling Results

Our key finding is that ANU and University of Melbourne data have ***sufficient text information*** to describe ***what*** research is being done. The text information can be used to generate topic areas of research and directly compared to other research activities internationally. There are several ways of describing research funding by topic, in addition to the availability of FOR areas. We have found that it is possible to generate research areas from the text of grants, which can inform stakeholders about what research is being funded. It is also possible to generate research areas from the text of the publications, which can inform stakeholders about what research is being producted.

We make use of new computational techniques to describe the scientific content of what scientific teams are funded to do and what research is being produced. This technique, known as topic modeling, provides a powerful and flexible framework for representing, summarizing and analyzing the contents of large document collections: it can be used to describe research topics – and hence describe what research is being done. The topics define a simplified representation of scientific documents; the research topics are defined by the research proposal, not manual generation of taxonomies or keywords (Blei, Ng and Jordan 2003; Newman, Hagedorn, Chemudugunta and Smyth 2007). This approach has been applied in the United States to both NSF and NIH awarded grants, and

a similar approach was used with the ANU and University of Melbourne corpus of documents.[19]

The topic model is a probabilistic model that automatically learns a set of topics (categories) that describe a collection of documents, based on the words in those documents. Each document is considered to consist of a small number of topics, each of which is dominated by only a fraction of all possible words. The selected granularity of the topics depends on the needs of the user. Further detail is provided in the appendix. This approach enabled us to identify the top research areas supported by research funding at both universities, in terms of both the volume of grants and the dollar amount of funding. We present that information in the following tables.

Topic models were computed for the ASTRA collection of ARC and NH&MRC grants, and publications from UniMelb and ANU combined. Topic models were learned at three different topic resolutions of 400, 500 and 600 topics. Note that there is no one 'correct' resolution for any given collection. A higher number of topics will produce more finely grained topics, but each finely grained topic covers a smaller number of documents. One typically seeks a balance between the specificity and generality of the topics, to the point where the topics are most useful to end-users for a particular purpose. A qualitative review of the T=500 topic model (where T=number of topics) suggested that this resolution model was suitable for initial use for ASTRA.

When one sets the topic model algorithm to learn about data from real-world content, approximately 10–20% of the topics 'learned' are not particularly useful. The topic model tags every single word in the corpus with a topic ID, and thereby needs to topically explain every single word occurrence. For example, with our proposal content there are topics that do not pertain to academic disciplines. For our T=500 topic resolution model, a preliminary review indicated that there were approximately 60 topics that were not useful for analysis. These topics were suppressed from subsequent analyses reported in tables (and are indicated by NULL in the list of topics).

The topic model assumes that each separate document in the collection exhibits a small number of topics. We typically represent each document by its top-4 to top-6 topics, where those top four to six topics are ranked by their relevance. When doing aggregates over subsets of documents (e.g. by year, or grant type, or by institution), there are several ways of using this top-k topic information. One can adopt a fully weighted approach that uses the topic-model weighting of the top-k topics to compute the aggregate ranking of topics for that subset. Alternately one can use just the single-top topic to represent each document. Both are valid approaches and result in a different focus. For this draft report we used the single-top topic to represent each document (which assigns 100% of each document to the top topic). This top-topic aggregate focuses on the main topic of interest, rather than the range of topics, so it tends to be a more peaked distribution on fewer topics.

The topic modeling results presented here are preliminary and could change in subsequent iterations of this report, and/or as the ASTRA project continues.

[19] The topic model was applied to all ARC Discovery (937 documents), ARC Linkage (323 documents), NH&MRC Projects (504 documents) and available publications (20,490 documents) made available to us by ANU (17,234 documents) and the University of Melbourne (6,648 documents) for the period 2007 to 2011.

Table 16: Preliminary Top Ten Topic Areas by Grant Volume: Australian National University	
Topic Area	**Ranking**
t130 grant cancer funding state	1
t74 pathogen plant host fungal resistance disease	2
t139channel ca2 receptor release calcium ryanodine	3
t216 neuron synaptic current channel potential cell	4
t475 synthesis chem compound lett product tetrahedron	5
t232 laser frequency cavity mode noise optical	6
t20 co2 rubisco plant von photosynthesis chloroplast	7
t367 archaeological artefact archaeology stone site excavation	8
t31 surface annealing ion layer thin sample substrate	9
t254 fertility population birth demographic age family	10

Table 17: Preliminary Top Ten Topic Areas by Grant Volume: University of Melbourne	
Topic Area	**Ranking**
t98 drosophila mutant cell gene phenotype regulate polarity	1
t96 alzheimer_disease amyloid master brain copper app	2
t189 species ecology invasive ecological biology invasion	3
t324 public_health qualitative screening interview health_promotion primary_care	4
t327 brain receptor disorder rat memory behavioural hippocampus	5
t130 grant cancer funding state council animal	6
t280 history historical war historian british melbourne sydney histories national archive	7
t34 immunol mice cell antigen cd4 immune lymphocyte	8
t302 neuron nerve receptor sympathetic rat nervous_system central	9
t419 epilepsy seizure brien epileptic brain epilepsia eeg	10

Table 18: Preliminary Top Ten Topic Areas by Funding Value: Australian National University	
Topic Area	**Ranking**
t130 grant cancer funding state council animal health	1
t377 landscape conservation habitat species vegetation lindenmayer tree woodland bird	2
t203 cognitive aging older age dementia ageing aged	3
t139 channel ca2 receptor release calcium ryanodine glutathione board	4
t475 synthesis chem compound lett product tetrahedron reaction natural	5
t371 economy investment foreign growth export economic domestic trade	6
t232 laser frequency cavity mode noise optical measurement	7
t216 neuron synaptic current channel potential cell recording neurosci	8
t367 archaeological artefact archaeology stone site excavation tool occupation assemblage	9
t290 complexe ligand organometallic complex chem metal tran nmr	10

Table 19: Preliminary Top Ten Topic Areas by Funding Value: University of Melbourne	
Topic Area	**Ranking**
t98 drosophila mutant cell gene phenotype regulate polarity	1
t324 public_health qualitative screening interview health_promotion primary	2
t96 alzheimer_disease amyloid master brain copper app cappai disease barnham	3
t327 brain receptor disorder rat memory behavioural hippocampus schizophrenia deficit	4
t306 parasite host parasitology parasitic nematode worm parasitol helminth gasser	5
t130 grant cancer funding state council animal health request list psp	6
t189 species ecology invasive ecological biology invasion conservation habitat modelling distribution	7
t247 physic device quantum fabrication optic physical electronic photonic	8
t34 immunol mice cell antigen cd4 immune lymphocyte cd8 responses cytokine spleen	9
t302 neuron nerve receptor sympathetic rat nervous_system central sensory pathway cell	10

Table 20: Preliminary Top Ten Topic Areas by Volume of Publications: Australian National University	
Topic Area	**Ranking**
t321 mode structure wave waveguide dielectric transmission coupling resonance	1
t8 soliton nonlinear beam optical mode kivshar lattice propagation opt	2
t136 image feature pixel detection recognition region vision method	3
t31 thin sample substrate material temperature	4
t298 system control stability time function state nonlinear stable feedback	5
t290 complexe ligand organometallic complex chem metal tran	6
t103 female male mating mate sex reproductive offspring	7
t377 landscape conservation habitat species vegetation lindenmayer tree woodland bird site	8
t383 waveguide optical glass device chalcogenide nonlinear power loss wavelength davies	9
t145 elix lichen acid species thallus ascospore apothecia genus minor	10

Table 21: Preliminary Top Ten Topic Areas by Volume of Publications: University of Melbourne	
Topic Area	**Ranking**
t210 student learning teacher teaching education skill curriculum school	1
t87 optical power fiber transmission ghz ber channel system dispersion	2
t378 music musical song sound performance musician composer instrument concert art	3
t61 cultural culture identity politic political context contemporary relation form practice	4
t463 management strategic business strategy organizational performance industry resource manager	5
t329 school education student educational achievement teacher schooling curriculum system social	6
t219 resource grid computing scheduling service distributed system user job task environment	7
t402 income economic wage employment household survey unemployment labour_market	8
t207 art artist painting image gallery exhibition collection time melbourne	9
t435 mathematic mathematical teacher student lesson classroom teaching problem school	10

The data schema structure also makes it possible to compare the topic areas generated by processing the text of awards with the manually generated fields of research for each document. This is the same principle that was used to produce NSF's Portfolio Explorer (http://smetrics.org/pe) without requiring either researchers or program managers to create scientific areas.

Tables 22 and 23 show the mapping for the top ten FORs for ANU and University of Melbourne grants to topics (ranked by grant volume); Tables 24 and 25 show the same mapping based on grant value. The appendices also show the results of the reverse exercise; the mapping of topics to fields of value.

The same approach has been used to classify publication information; those results are shown in Tables 26 and 27. Additional information is available in the appendices as, well as at http://www.ics.uci.edu/~newman/astra/20121114/.

Table 22: Top FoR Codes (by Number of Grants) and their Most Commonly Occurring Topics, University of Melbourne

FoR Codes		Topics
1109	1st	[t216] neuron synaptic current channel potential cell recording neurosci voltage sl
Neurosciences	2nd	[t491] mice expression cell protein mouse model gene treatment control vivo ...
	3rd	
1103	1st	[t86] patient clinical treatment diagnosis hospital disease med blood acute severe
Clinical Sciences	2nd	
	3rd	
1117	1st	[t315] patient medical care health hospital clinical health_care doctor medicine pr
Public Health and	2nd	[t92] trial intervention outcome group participant treatment recruitment baseline p
Health Services	3rd	[t334] health public_health disease population social diseases countries poor globa
1112	1st	[t376] tumour cancer tumor metastasis growth breast_cancer human metastatic cell ma
Oncology and	2nd	[t480] cancer tumour ovarian_cancer ovarian breast_cancer mutation patient gene dna
Carcinogenesis	3rd	[t372] patient cancer therapy cell survival tumour clinical response treatment lymp
1107	1st	[t34] immunol mice cell antigen cd4 immune lymphocyte cd8 responses cytokine ...
Immunology	2nd	[t471] cell anti antibody control mouse mice staining pbs expression antibodies ...
	3rd	[t155] cytokine macrophage inflammatory immune tnf factor cell induced production monocyte ...
601	1st	[t200] residues binding protein interaction structure domain structural terminal amino_acid bind ...
Biochemistry and	2nd	[t255] min concentration buffer solution added protein cell method sample culture ...
Cell Biology	3rd	
1102	1st	[t491] mice expression cell protein mouse model gene treatment control vivo ...
Cardiovascular	2nd	[t148] rat hypertension vascular angiotensin induced blood pressure blood endotheli
Medicine and	3rd	[t202] heart cardiac failure myocardial ventricular cardiovascular hypertrophy myoc
604	1st	[t69] pcr dna primer sequence gene product region ampli reaction fragment ...
Genetics	2nd	[t102] sequence genome protein gene alignment conserved amino_acid genomic database family ...
	3rd	[t133] chromosome mammal gene marsupial sex grave evolution genome human eutherian ...
1114	1st	[t197] pregnancy maternal women fetal birth pregnant mother placental obstetric pla
Paediatrics and	2nd	[t86] patient clinical treatment diagnosis hospital disease med blood acute severe
Reproductive	3rd	[t461] infant birth weight neonatal pediatric week preterm neonate newborn outcome
602	1st	[t103] female male mating mate sex reproductive offspring evolution success sexual
Ecology	2nd	[t75] bird nest breeding species egg avian fairy wren behaviour nesting ...
	3rd	

Table 23: Top FoR Codes (by Number of Grants) and their Most Commonly Occurring Topics, Australian National University		
FoR Codes		Topics
1117	1st	[t315] patient medical care health hospital clinical health_care doctor medicine practice ...
Public Health and Health Services	2nd	[t92] trial intervention outcome group participant treatment recruitment baseline program data ...
	3rd	[t334] health public_health disease population social diseases countries poor global mortality ...
601	1st	[t200] residues binding protein interaction structure domain structural terminal amino_acid bind ...
Biochemistry and Cell Biology	2nd	[t255] min concentration buffer solution added protein cell method sample culture ...
	3rd	
604	1st	[t69] pcr dna primer sequence gene product region ampli reaction fragment ...
Genetics	2nd	[t102] sequence genome protein gene alignment conserved amino_acid genomic database family ...
	3rd	[t133] chromosome mammal gene marsupial sex grave evolution genome human eutherian ...
607	1st	[t93] leaf co2 leave plant water stomatal photosynthesis conductance exchange species ...
Plant Biology	2nd	[t500] arabidopsis plant root mutant plant_cell thaliana development gene growth physiol ...
	3rd	[t389] metabolism metabolic pathway metabolite enzyme acid mitochondrial stress synthesis condition ...
205	1st	[t8] soliton nonlinear beam optical mode kivshar lattice propagation opt waveguide ...
Optical Physics	2nd	[t321] mode structure wave waveguide dielectric transmission coupling resonance light frequency ...
	3rd	[t442] solution wave nonlinear soliton parameter amplitude evolution equation initial instability ...
2103	1st	[t343] colonial empire british imperial history european trade slave world britain ...
Historical Studies	2nd	
	3rd	
2101	1st	[t367] archaeological artefact archaeology stone site excavation tool occupation assemblage cave ...
Archaeology	2nd	[t260] island lapita site archaeological archaeology paci oceania anderson pacific prehistoric ...
	3rd	[t193] origin neolithic human region philippine island southeast_asia bellwood evidence dispersal ...
1107	1st	[t34] immunol mice cell antigen cd4 immune lymphocyte cd8 responses cytokine ...
Immunology	2nd	[t471] cell anti antibody control mouse mice staining pbs expression antibodies ...
	3rd	[t155] cytokine macrophage inflammatory immune tnf factor cell induced production monocyte ...
101	1st	[t179] group map module algebra math element product isomorphism algebras theory ...
Pure Mathematics	2nd	[t493] solution equation function estimate convex smooth constant boundary elliptic math ...
	3rd	[t80] manifold invariant lie group map space metric structure geometry bundle ...
202	1st	[t405] scattering energy electron cross_section measurement beam elastic collision angle energies ...
Atomic, Molecular, Nuclear, Particle and Plasma Physics	2nd	[t271] plasma ion density electron plasmas current source pressure phy magnetic_eld ...
	3rd	[t17] state kev transition decay ray phy energy neutron spin con_guration ...

Table 24: Top FoR Codes (by Grant Dollars) and their Most Commonly Occurring Topics, University of Melbourne		
FoR Codes		Topics
1109	1st	[t216] neuron synaptic current channel potential cell recording neurosci voltage slice ...
Neurosciences	2nd	[t491] mice expression cell protein mouse model gene treatment control vivo ...
	3rd	
1107	1st	[t34] immunol mice cell antigen cd4 immune lymphocyte cd8 responses cytokine ...
Immunology	2nd	[t471] cell anti antibody control mouse mice staining pbs expression antibodies ...
	3rd	[t155] cytokine macrophage inflammatory immune tnf factor cell induced production monocyte ...
1103	1st	[t86] patient clinical treatment diagnosis hospital disease med blood acute severe ...
Clinical Sciences	2nd	
	3rd	
1112	1st	[t376] tumour cancer tumor metastasis growth breast_cancer human metastatic cell mammary ..
Oncology and Carcinogenesis	2nd	[t480] cancer tumour ovarian_cancer ovarian breast_cancer mutation patient gene dna sample
	3rd	[t372] patient cancer therapy cell survival tumour clinical response treatment lymphoma ...
1117	1st	[t315] patient medical care health hospital clinical health_care doctor medicine practice .
Public Health and Health Services	2nd	[t92] trial intervention outcome group participant treatment recruitment baseline program d
	3rd	[t334] health public_health disease population social diseases countries poor global mortal
903	1st	[t188] mechanical stress material strain properties shear force deformation elastic load ..
Biomedical Engineering	2nd	[t202] heart cardiac failure myocardial ventricular cardiovascular hypertrophy myocyte circ
	3rd	[t274] polymer material poly synthesis properties surface chemical gel mater particle ...
601	1st	[t200] residues binding protein interaction structure domain structural terminal amino_acid
Biochemistry and Cell Biology	2nd	[t255] min concentration buffer solution added protein cell method sample culture ...
	3rd	
202	1st	[t405] scattering energy electron cross_section measurement beam elastic collision angle
Atomic, Molecular, Nuclear, Particle and	2nd	[t271] plasma ion density electron plasmas current source pressure phy magnetic_eld ...
	3rd	[t17] state kev transition decay ray phy energy neutron spin con_guration ...
1101	1st	[t200] residues binding protein interaction structure domain structural terminal amino_acid
Medical Biochemistry and Metabolomics	2nd	[t255] min concentration buffer solution added protein cell method sample culture ...
	3rd	
604	1st	[t69] pcr dna primer sequence gene product region ampli reaction fragment ...
Genetics	2nd	[t102] sequence genome protein gene alignment conserved amino_acid genomic database family ...
	3rd	[t133] chromosome mammal gene marsupial sex grave evolution genome human eutherian ...

Table 25: Top FoR Codes (by Grant Dollars) and their Most Commonly Occurring Topics, Australian National University		
FoR Codes		**Topics**
403	1st	[t236] zircon age grain analyses shrimp geochronology biotite core granite sample ...
Geology	2nd	[t191] belt tectonic precambrian group rock basement craton terrane neoproterozoic age ...
	3rd	[t190] trace_element element rock ree composition sio2 geochemical rich magma cao ...
201	1st	[t18] telescope star image band observation survey line data optical color ...
Astronomical and Space Sciences	2nd	[t263] galaxies galaxy mass apj mnras velocity stellar gas population ngc ...
	3rd	[t99] star cluster halo galactic distance dwarf velocity radial distribution kpc ...
1117	1st	[t315] patient medical care health hospital clinical health_care doctor medicine practice .
Public Health and Health Services	2nd	[t92] trial intervention outcome group participant treatment recruitment baseline program d
	3rd	[t334] health public_health disease population social diseases countries poor global mortal
1107	1st	[t34] immunol mice cell antigen cd4 immune lymphocyte cd8 responses cytokine ...
Immunology	2nd	[t471] cell anti antibody control mouse mice staining pbs expression antibodies ...
	3rd	[t155] cytokine macrophage inflammatory immune tnf factor cell induced production monocyte ...
601	1st	[t200] residues binding protein interaction structure domain structural terminal amino_acid bind ...
Biochemistry and Cell Biology	2nd	[t255] min concentration buffer solution added protein cell method sample culture ...
	3rd	
1605	1st	[t370] government policy state development national local policies political problem instit
Policy and Administration	2nd	
	3rd	
1601	1st	[t365] anthropology ritual anthropological social anthropologist practice ethnographic exch
Anthropology	2nd	[t61] cultural culture identity politic political context contemporary relation form practi
	3rd	
1007	1st	[t31] surface annealing ion layer thin sample substrate material temperature amorphous ...
Nanotechnology	2nd	[t170] growth gaas nanowire device layer quantum semiconductor electron energy tan ...
	3rd	
1801	1st	[t2] act legislation law provision regulation legal protection state committee amendment ..
Law	2nd	[t358] law court legal judicial judge justice constitutional lawyer power constitution ...
	3rd	
603	1st	[t57] tree phylogenetic sequence molecular clade evolution rate analyses dna lineage ...
Evolutionary Biology	2nd	[t144] species genus taxa genera systematic plant group taxonomic specimen taxon ...
	3rd	[t145] elix lichen acid species thallus ascospore apothecia genus minor thick ...

Table 26: Top FoR Codes (by Number of Pubs) and their Most Commonly Occurring Topics, University of Melbourne		
FoR Codes		**Topics**
1103	1st	[t86] patient clinical treatment diagnosis hospital disease med blood acute severe ...
Clinical Sciences	2nd	
	3rd	
1117	1st	[t315] patient medical care health hospital clinical health_care doctor medicine practice .
Public Health and Health Services	2nd	[t92] trial intervention outcome group participant treatment recruitment baseline program d
	3rd	[t334] health public_health disease population social diseases countries poor global mortal
1109	1st	[t216] neuron synaptic current channel potential cell recording neurosci voltage slice ...
Neurosciences	2nd	[t491] mice expression cell protein mouse model gene treatment control vivo ...
	3rd	
1102	1st	[t491] mice expression cell protein mouse model gene treatment control vivo ...
Cardiovascular Medicine and	2nd	[t148] rat hypertension vascular angiotensin induced blood_pressure blood endothelial arter
	3rd	[t202] heart cardiac failure myocardial ventricular cardiovascular hypertrophy myocyte circ
1113	1st	[t101] eye retinal retina vision visual photoreceptor ophthalmology age degeneration cone .
Ophthalmology and Optometry	2nd	[t171] visual eye responses response stimulus cell vision temporal system stimuli ...
	3rd	
1402	1st	[t473] growth economic price demand capital productivity factor income production output ..
Applied Economics	2nd	
	3rd	
1904	1st	[t378] music musical song sound performance musician composer instrument concert art ...
Performing Arts and Creative Writing	2nd	[t90] theatre performance art drama play audience dance form production artist ...
	3rd	
705	1st	[t467] forest tree wood timber stand forestry fire logging management res ...
Forestry Sciences	2nd	[t39] vegetation forest plant species soil ecology site tropical dry fire ...
	3rd	[t489] environmental conservation management environment land ecological biodiversity susta
1701	1st	[t33] cognitive task processing memory attention psychology visual participant cognition br
Psychology	2nd	[t248] social psychology group social_psychology psychological personality theory identity
	3rd	
601	1st	[t200] residues binding protein interaction structure domain structural terminal amino_acid bind ...
Biochemistry and Cell Biology	2nd	[t255] min concentration buffer solution added protein cell method sample culture ...
	3rd	

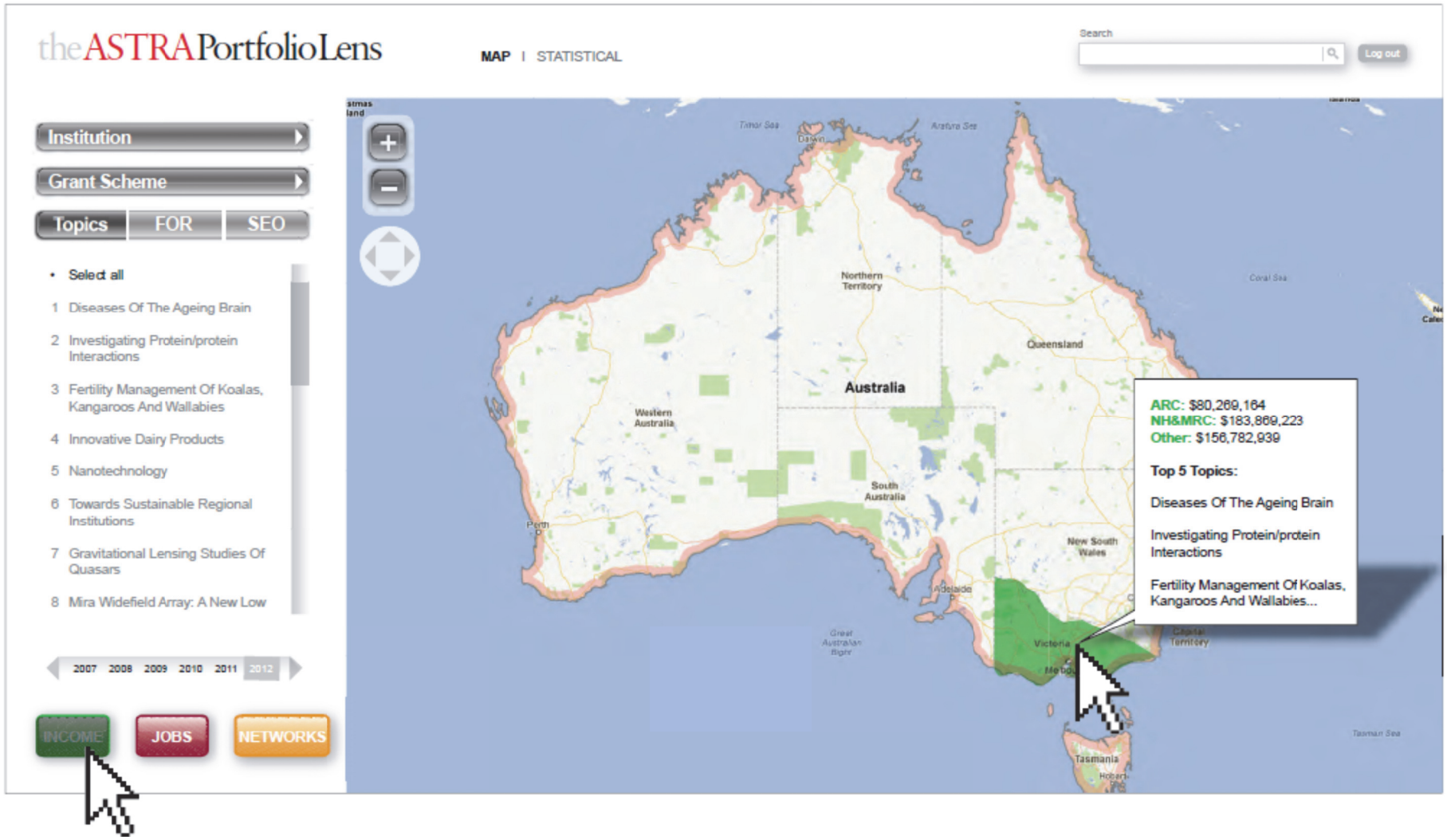
Table 27: Top FoR Codes (by Number of Pubs) and their Most Commonly Occurring Topics, Australian National University		
FoR Codes		Topics
201	1st	[t18] telescope star image band observation survey line data optical color ...
Astronomical and Space Sciences	2nd	[t263] galaxies galaxy mass apj mnras velocity stellar gas population ngc ...
	3rd	[t99] star cluster halo galactic distance dwarf velocity radial distribution kpc ...
1117	1st	[t315] patient medical care health hospital clinical health_care doctor medicine practice .
Public Health and Health Services	2nd	[t92] trial intervention outcome group participant treatment recruitment baseline program d
	3rd	[t334] health public health disease population social diseases countries poor global mortal
1402	1st	[t473] growth economic price demand capital productivity factor income production output ..
Applied Economics	2nd	
	3rd	
205	1st	[t8] soliton nonlinear beam optical mode kivshar lattice propagation opt waveguide ...
Optical Physics	2nd	[t321] mode structure wave waveguide dielectric transmission coupling resonance light
	3rd	[t442] solution wave nonlinear soliton parameter amplitude evolution equation initial
1103	1st	[t86] patient clinical treatment diagnosis hospital disease med blood acute severe ...
Clinical Sciences	2nd	
	3rd	
403	1st	[t236] zircon age grain analyses shrimp geochronology biotite core granite sample ...
Geology	2nd	[t191] belt tectonic precambrian group rock basement craton terrane neoproterozoic age ...
	3rd	[t190] trace element element rock ree composition sio2 geochemical rich magma cao ...
202	1st	[t405] scattering energy electron cross section measurement beam elastic collision angle
Atomic, Molecular, Nuclear, Particle and	2nd	[t271] plasma ion density electron plasmas current source pressure phy magnetic eld ...
	3rd	[t17] state kev transition decay ray phy energy neutron spin con guration ...
502	1st	[t489] environmental conservation management environment land ecological biodiversity susta
Environmental Science and Management	2nd	[t377] landscape conservation habitat species vegetation lindenmayer tree woodland bird sit
	3rd	
402	1st	[t388] isotope sample ratios analytical analyses element standard isotopic concentration ic
Geochemistry	2nd	[t190] trace_element element rock ree composition sio2 geochemical rich magma cao ...
	3rd	[t236] zircon age grain analyses shrimp geochronology biotite core granite sample ...
602	1st	[t103] female male mating mate sex reproductive offspring evolution success sexual ...
Ecology	2nd	[t75] bird nest breeding species egg avian fairy wren behaviour nesting ...
	3rd	

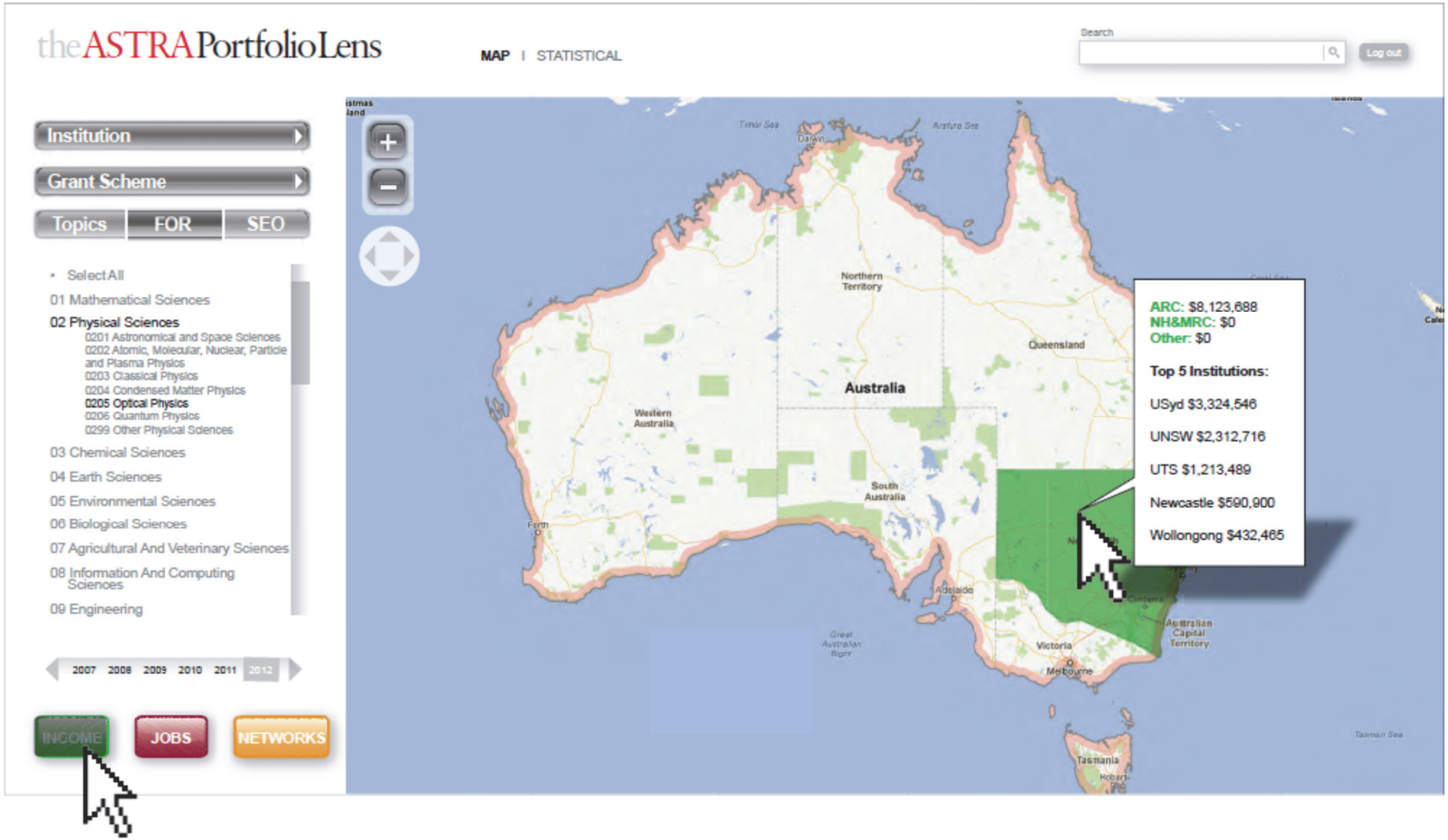
8. Where is Research being Conducted? ASTRA Portfolio Lens

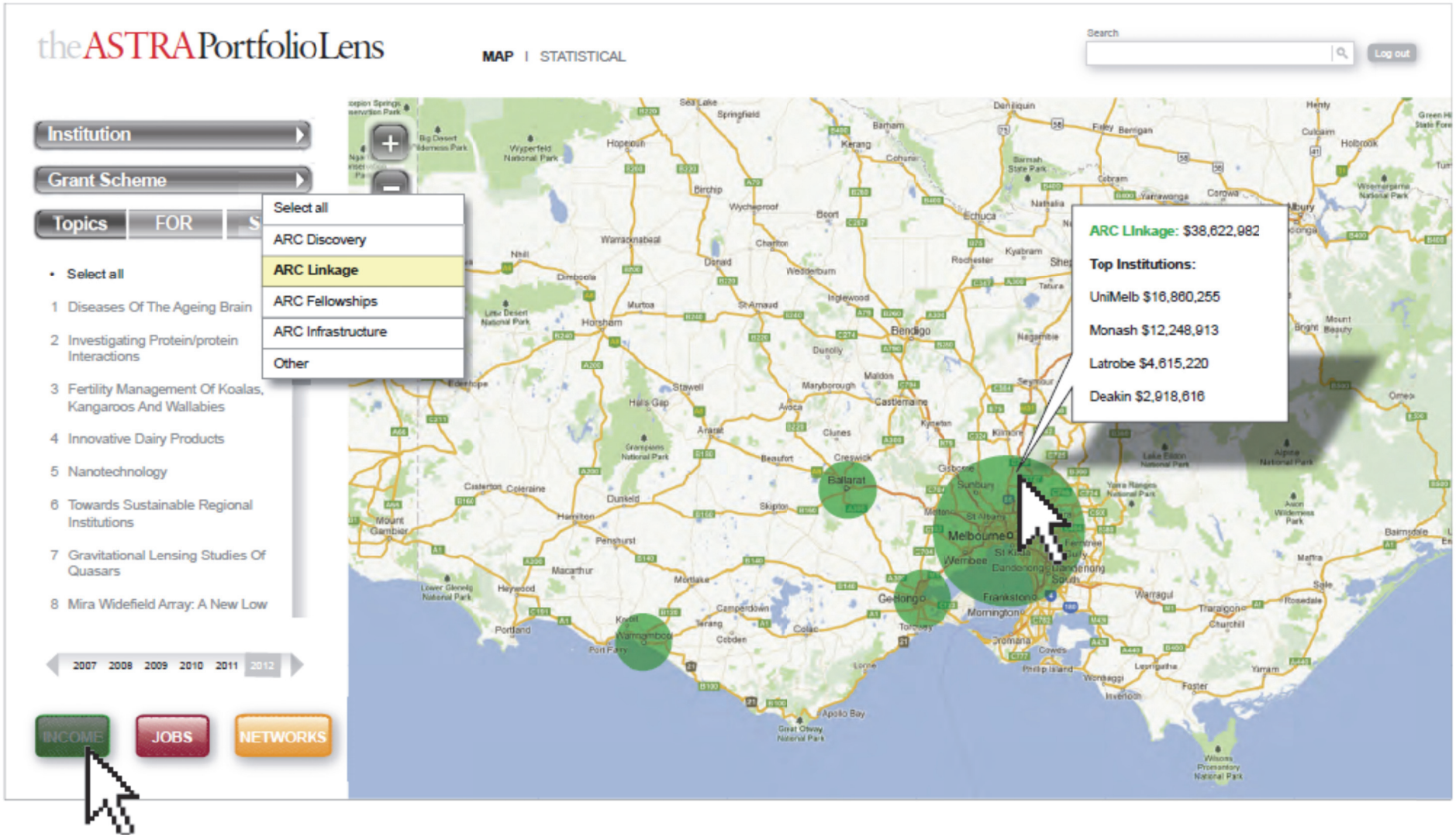
One of the key deliverables outlined in the ASTRA Feasibility Report was a stocktake of data capacity to provide a visualization of research portfolios for DVCRs and DIISRTE. The existence of these data would enable us to build a website to develop an ASTRA Portfolio Lens which displays a summary of the dollar value of research grants awarded by scheme to each university, as well as the number of jobs supported by these research grants (within the university and in external research support providers) and the collaborative networks (co-CIs) observed. In essence, the Portfolio Lens would act as a visual representation – and exploratory tool – describing the nation's research portfolio. The key finding from this part of the Feasibility Study is that it it would be indeed be possible to build such a website using the data identified. We commissioned Geronimo Creative Services to develop wireframes of the proposed website, some snapshots of which are presented below.

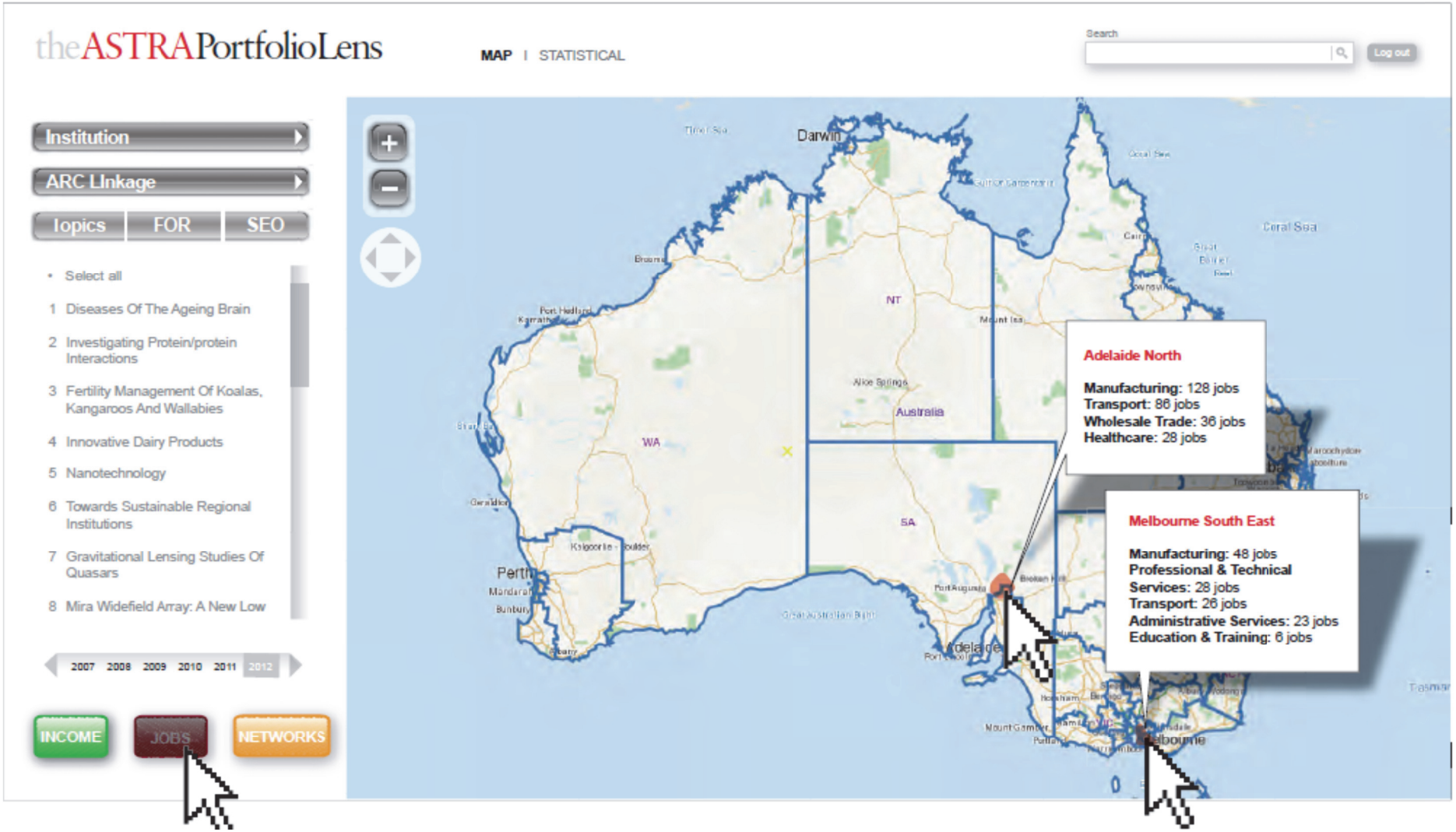
The wireframes that we present are meant to be illustrative of the type of functionality that could be built into the ASTRA Protfolio Lens. In each wireframe, it is possible to filter the search by Institution, Grant Scheme, Topics (or FOR/SEO Code) and Year. And by using the 'zoom' function, this can be observed at the city, state or national level. In the first wireframe, for example, by hovering over the state of Victoria, the amount of research income awarded in 2012 to all Victorian universities (ARC, NH&MRC and other sources) and the top topic areas relating to these research grants is displayed. Moving the hover would provide similar descriptive information on the research income obtained in other states of Australia.

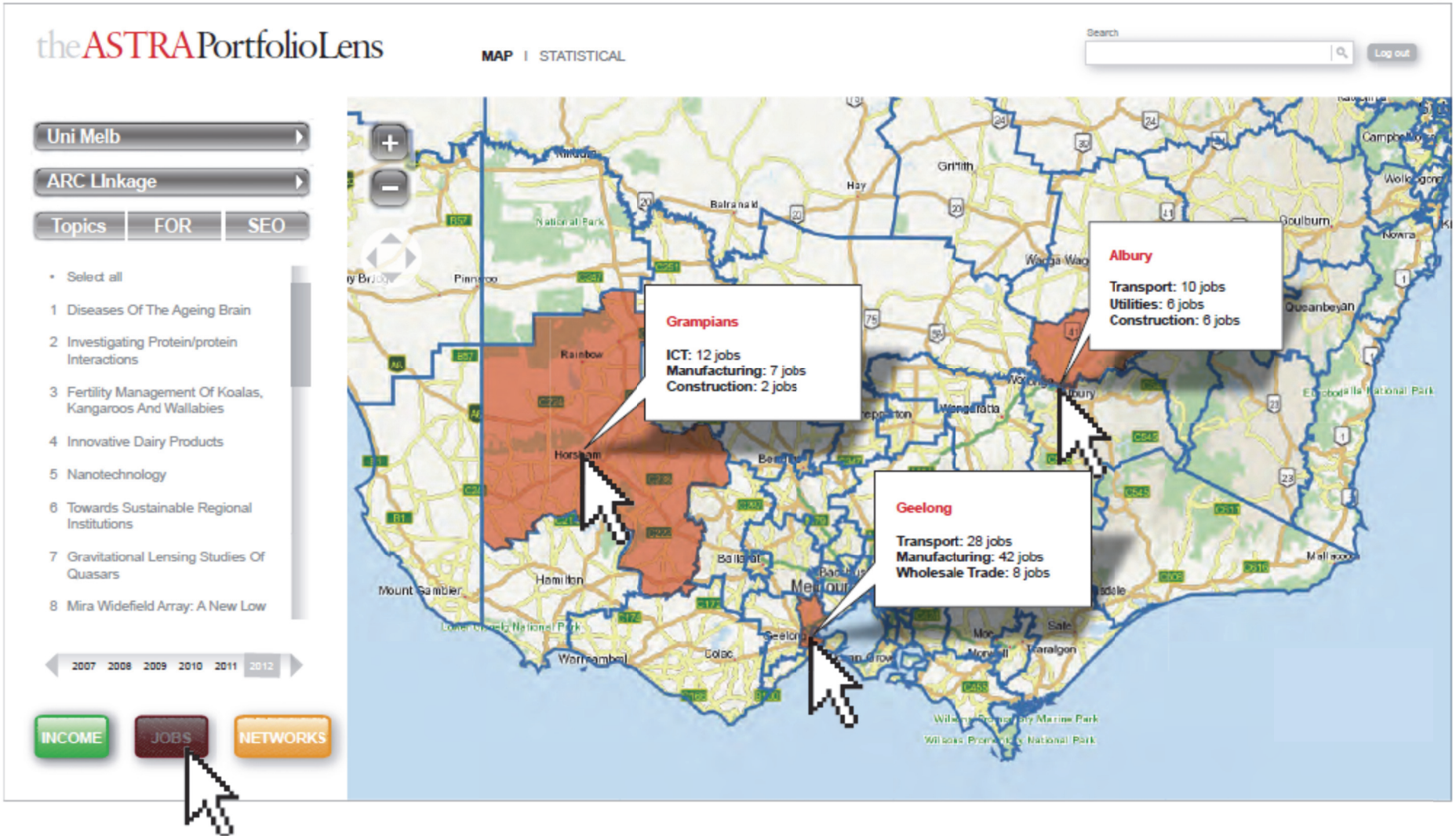
ASTRA Portfolio Lens: Research Income

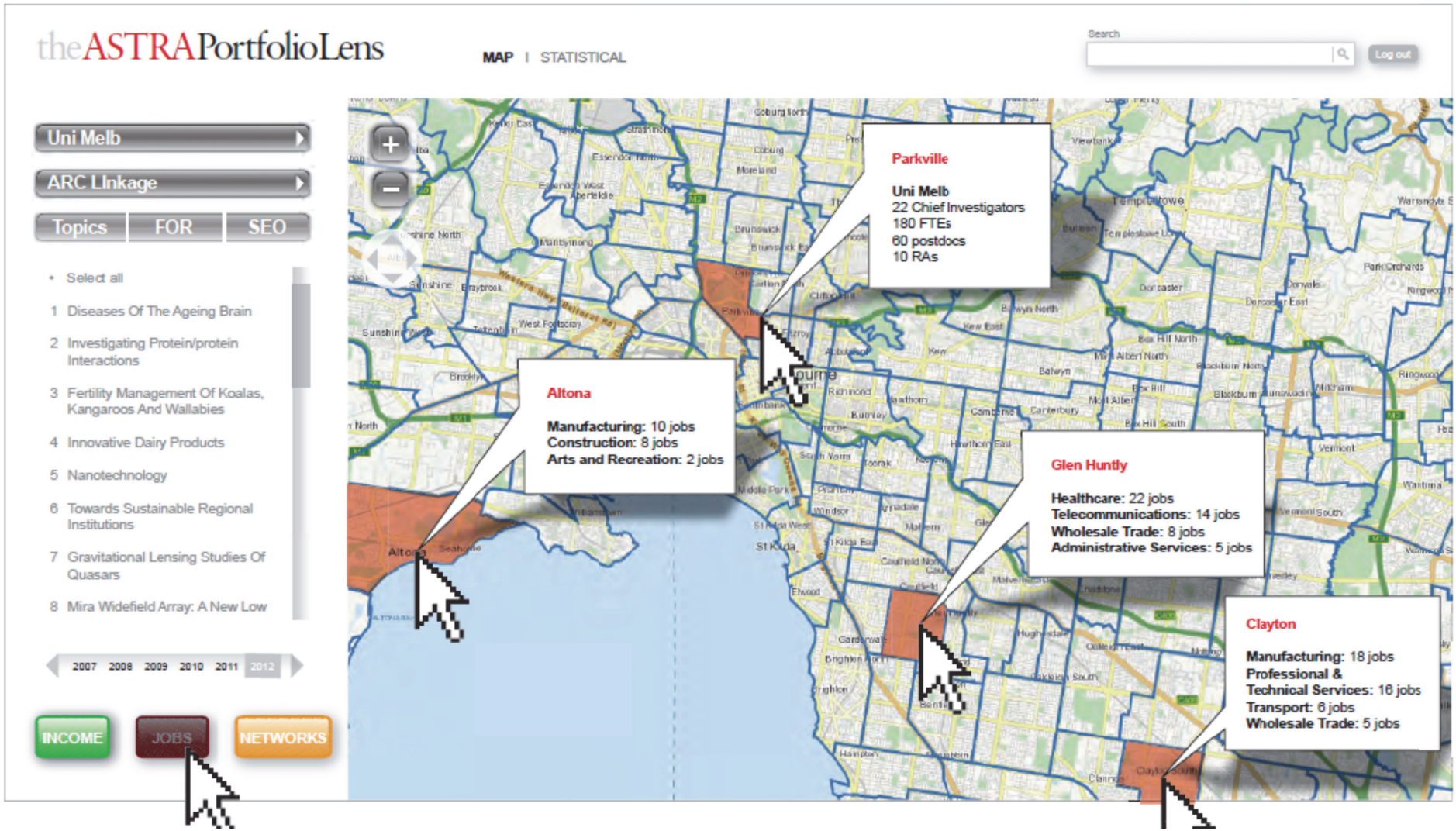


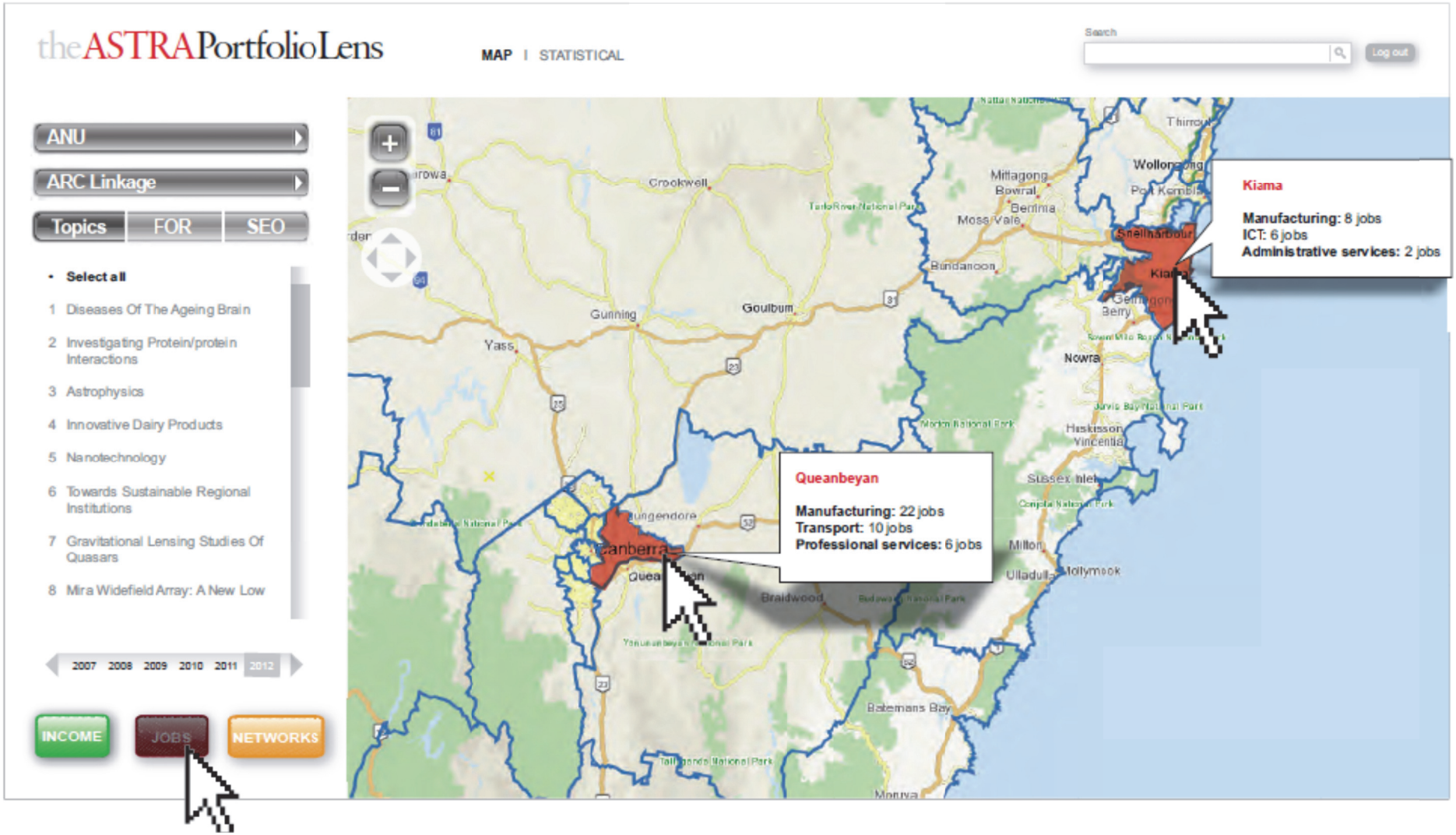

theASTRAPortfolioLens
MAP | STATISTICAL
Search
Log out
Institution
Grant Scheme
Topics
FOR
SEO
Select All
01 Mathematical Sciences
02 Physical Sciences
0201 Astronomical and Space Sciences
0202 Atomic, Molecular, Nuclear, Particle and Plasma Physics
0203 Classical Physics
0204 Condensed Matter Physics
0205 Optical Physics
0206 Quantum Physics
0299 Other Physical Sciences
03 Chemical Sciences
04 Earth Sciences
05 Environmental Sciences
06 Biological Sciences
07 Agricultural And Veterinary Sciences
08 Information And Computing Sciences
09 Engineering
2007 2008 2009 2010 2011 2012
INCOME
JOBS
NETWORKS
Timor Sea
Arafura Sea
Darwin
Northern Territory
Queensland
Australia
Western Australia
South Australia
Perth
Adelaide
Great Australian Bight
Victoria
Melbourne
Australian Capital Territory
Tasmania
Hobart
Tasman Sea
ARC: $8,123,688
NH&MRC: $0
Other: $0
Top 5 Institutions:
USyd $3,324,546
UNSW $2,312,716
UTS $1,213,489
Newcastle $590,900
Wollongong $432,465


theASTRAPortfolioLens
MAP | STATISTICAL
Search
Log out
Institution
Grant Scheme
Topics
FOR
Select all
ARC Discovery
ARC Linkage
ARC Fellowships
ARC Infrastructure
Other
Select all
1 Diseases Of The Ageing Brain
2 Investigating Protein/protein Interactions
3 Fertility Management Of Koalas, Kangaroos And Wallabies
4 Innovative Dairy Products
5 Nanotechnology
6 Towards Sustainable Regional Institutions
7 Gravitational Lensing Studies Of Quasars
8 Mira Widefield Array: A New Low
2007 2008 2009 2010 2011 2012
INCOME
JOBS
NETWORKS
ARC Linkage: $38,622,982
Top Institutions:
UniMelb $16,860,255
Monash $12,248,913
Latrobe $4,615,220
Deakin $2,918,616


theASTRAPortfolioLens
MAP | STATISTICAL
Search
Log out
Institution
ARC Linkage
Topics
FOR
SEO
Select all
1 Diseases Of The Ageing Brain
2 Investigating Protein/protein Interactions
3 Fertility Management Of Koalas, Kangaroos And Wallabies
4 Innovative Dairy Products
5 Nanotechnology
6 Towards Sustainable Regional Institutions
7 Gravitational Lensing Studies Of Quasars
8 Mira Widefield Array: A New Low
2007 2008 2009 2010 2011 2012
INCOME
JOBS
NETWORKS
Darwin
NT
WA
SA
Australia
Perth
Adelaide
Coral Sea
Adelaide North
Manufacturing: 128 jobs
Transport: 86 jobs
Wholesale Trade: 36 jobs
Healthcare: 28 jobs
Melbourne South East
Manufacturing: 48 jobs
Professional & Technical Services: 28 jobs
Transport: 26 jobs
Administrative Services: 23 jobs
Education & Training: 6 jobs


theASTRAPortfolioLens
MAP | STATISTICAL
Search
Log out
Uni Melb
ARC Linkage
Topics
FOR
SEO
Select all
1 Diseases Of The Ageing Brain
2 Investigating Protein/protein Interactions
3 Fertility Management Of Koalas, Kangaroos And Wallabies
4 Innovative Dairy Products
5 Nanotechnology
6 Towards Sustainable Regional Institutions
7 Gravitational Lensing Studies Of Quasars
8 Mira Widefield Array: A New Low
2007 2008 2009 2010 2011 2012
INCOME
JOBS
NETWORKS
Grampians
ICT: 12 jobs
Manufacturing: 7 jobs
Construction: 2 jobs
Albury
Transport: 10 jobs
Utilities: 6 jobs
Construction: 6 jobs
Geelong
Transport: 28 jobs
Manufacturing: 42 jobs
Wholesale Trade: 8 jobs
Hay
Griffith
Balranald
Goulburn
Queanbeyan
Rainbow
Horsham
Hamilton
Mount Gambier
Ballarat
Warrnambool
Colac
Sale
Traralgon


theASTRAPortfolioLens
MAP | STATISTICAL
Search
Log out
Uni Melb
ARC LInkage
Topics
FOR
SEO
Select all
1 Diseases Of The Ageing Brain
2 Investigating Protein/protein Interactions
3 Fertility Management Of Koalas, Kangaroos And Wallabies
4 Innovative Dairy Products
5 Nanotechnology
6 Towards Sustainable Regional Institutions
7 Gravitational Lensing Studies Of Quasars
8 Mira Widefield Array: A New Low
2007 2008 2009 2010 2011 2012
INCOME
JOBS
NETWORKS
Parkville
Uni Melb
22 Chief Investigators
180 FTEs
60 postdocs
10 RAs
Altona
Manufacturing: 10 jobs
Construction: 8 jobs
Arts and Recreation: 2 jobs
Glen Huntly
Healthcare: 22 jobs
Telecommunications: 14 jobs
Wholesale Trade: 8 jobs
Administrative Services: 5 jobs
Clayton
Manufacturing: 18 jobs
Professional & Technical Services: 16 jobs
Transport: 6 jobs
Wholesale Trade: 5 jobs


theASTRAPortfolioLens
MAP | STATISTICAL
Search
Log out
ANU
ARC Linkage
Topics
FOR
SEO
Select all
1 Diseases Of The Ageing Brain
2 Investigating Protein/protein Interactions
3 Astrophysics
4 Innovative Dairy Products
5 Nanotechnology
6 Towards Sustainable Regional Institutions
7 Gravitational Lensing Studies Of Quasars
8 Mira Widefield Array: A New Low
2007 2008 2009 2010 2011 2012
INCOME
JOBS
NETWORKS
Kiama
Manufacturing: 8 jobs
ICT: 6 jobs
Administrative services: 2 jobs
Queanbeyan
Manufacturing: 22 jobs
Transport: 10 jobs
Professional services: 6 jobs
Crookwell
Goulburn
Gunning
Yass
Bundanoon
Nowra
Berry
Canberra
Braidwood
Batemans Bay
Milton
Ulladulla
Mollymook

ASTRA Portfolio Lens: Collaborative Networks