Submission
2016 National Research Infrastructure Roadmap
Capability Issues Paper

Name	**Amanda Lawrence**
Title/role	**Research and Strategy Manager**
Organisation	**APO**

Introduction

Australian Policy Online (APO) is award-winning, multidisciplinary, research infrastructure specialising in policy and practice grey literature and data from Australia, New Zealand and around the world. APO currently addresses some of the research infrastructure needs for collecting and analysing policy and practice related research resources but is severely limited by the limited scale of its operations and an insecure funding model.

APO currently hosts 38,000 records including commissioned reports, discussion papers, working papers, briefings, conference papers, evaluations, case studies, data sets, infographics, audio and video, and links to websites and other databases and research tools. APO features the work of over 4,000 organisations and 16,000 authors. It is able to mint DOIs for documents and data, integrates with Orcid and Datacite, is interoperable with Trove and other databases via OAIPMH and APIs, and by early 2017 will include a CKAN datastore, linked data relationships and visualisations and some capacity for automated indexing using text mining tools.

APO is a not-for-profit organisation that has been sustained by a variety of institutional supporters, grants, partnerships and advertising since it was established by Swinburne Institute for Social Research in 2002. APO has been awarded multiple ARC Linkage Infrastructure and Equipment Grants (LIEF) grants, most recently in 2016, and two ANDS grants. Partner organisations include: Swinburne University of Technology, the University of South Australia, the Australia and New Zealand School of Government, Henry Halloran Trust, University of Sydney, University of Canberra, University of Melbourne, RMIT University, Victoria University of Wellington, the Internet Archive and the National Library of Australia.

This submission is made with the support of the APO Advisory Group:

- Mr Glenn Campbell, Business Manager, ANZSOG
- Prof Jago Dodson, Director, Centre for Urban Research, RMIT University
- Prof Gerard Goggin, Director, Department of Media and Communications, University of Sydney
- Prof Denise Meredyth, Pro Vice Chancellor, Division of Education, Arts and Social Sciences, University of South Australia
- Prof Peter Phibbs, Director, Henry Halloran Trust, University of Sydney
- Assoc Prof Ellie Rennie, Deputy Director, Swinburne Institute for Social Research, Swinburne University of Technology

- Prof Julian Thomas (Chair), Director, Swinburne Institute for Social Research, Swinburne University of Technology
- Assoc. Prof Jerry Watkins, Director, News & Media Research Centre, University of Canberra
- Mr Derek Whitehead, Chair of the Australian Digital Alliance

Question 1: Are there other capability areas that should be considered?

Many of the policy and research imperatives facing Australia in the coming decades are complex, multidisciplinary and beyond the capacity of any one sector – whether it is education, government, industry or civil society – to address effectively. Issues such as sustainable urban growth, regional economic development, social cohesion and national security, ehealth, Indigenous reconciliation, climate change adaptation, changing education and employment needs, transport, digital technologies and many more, involve researchers across the sciences, social sciences and the humanities and require engagement and participation from industry, government, civil society and the general public.

To undertake this work involves sources that are highly varied, existing in both digital and physical forms and in a wide range of formats, from historical and archival documents, artefacts and objects to large scale data sets and visualisations. Research data and primary sources may be produced or owned by researchers and institutions themselves, generated by machinery or located in institutions and archives around the world, or be scatted across federal, state and local governments, civil society organisations (CSOs), industry and business.

Grey literature is a term used to describe publications produced and disseminated directly by organisations, including government departments and agencies, academic research centres, NGOs, think tanks and industry – in academic terms 'non-traditional research objects' or NTROs. They are an essential part of the policy process, providing the evidence-base for many policy decisions. Grey literature is widely recognised as essential to the research process in health, criminology, archaeology, engineering, environmental science and many other disciplines. Recent research indicates that grey literature reports are the most used and important source of policy and practice related research and grey literature's use value in Australia could be as high as $30 billion p.a.[1] One of the challenges presented by grey literature is that it does not flow through traditional publishing channels and is therefore dispersed and disaggregated across the internet, difficult to find and evaluate and often lost in what has been described as a digital black hole.

Whether resources to be analysed are physical or digital, academic researchers, industry, civil society organisations and governments face considerable difficulties finding and accessing grey literature documents and data and lack adequate tools for creating and visualising new connections and insights. We need to move beyond disciplinary boundaries and also traditional concepts of what is a document and what is data to realise that documents are a form of data and data is found in documents – therefore we need to find cost effective ways to identify, curate, store, analyse and visualise documents as well as data at national and international scales.

[1] Lawrence A, Houghton J, Thomas T and Weldon P, 2014, *Where is the evidence: realising the value of grey literature for public policy and practice,* Swinburne Institute for Social Research, doi.org/10.4225/50/5580B1E02DAF9

Question 2: Are these governance characteristics appropriate and are there other factors that should be considered for optimal governance for national research infrastructure.

Question 3: Should national research infrastructure investment assist with access to international facilities?

Question 4: What are the conditions or scenarios where access to international facilities should be prioritised over developing national facilities?

Question 5: Should research workforce skills be considered a research infrastructure issue?

There is currently a substantial and growing gap in researchers' knowledge and skills for effectively participating in the scholarly and public communication and dissemination landscape. This lack of knowledge ranges from lack of familiarity with new publishing models, poor production standards for organisation-based publishing (grey literature), through to profound misunderstanding of the issues relating to intellectual property as they relate to research outputs. We think that skills in this and other areas are essential to maximise the benefit from research infrastructure projects as well as from the research system as a whole.

If we are to benefit from the investments already made in research databases and collections it is essential that a national coordinated approach to linked data and document management is made that parallels the excellent work of ANDS in training researchers and research institutions about data management and storage. One option would be to create a similar organisation for text and other forms of publishing, alternatively the remit of ANDS could be expanded. The aim would be to provide instructution, guides and funding to promote best practices for producing and managing every aspect of the research cycle including data, documents, case studies, images, audio and video, field notes, protocols, etc. and then how best to approach the production, dissemination and communication aspects, and finally what new tools for analysis of data, texts, images and many other forms should researchers be considering.

Question 6: How can national research infrastructure assist in training and skills development?

Question 7: What responsibility should research institutions have in supporting the development of infrastructure ready researchers and technical specialists?

Question 8: What principles should be applied for access to national research infrastructure, and are there situations when these should not apply?

While merit based access is an important principle, in some situations this has the potential to restrict access to researchers who have established track records, with researchers earlier in their career or trialling highly creative or innovative approaches potentially disadvantaged. It may also discriminate against industry-oriented research which is directed towards more commercial outcomes that are less traditional. Access should provide for a spectrum of activities and should not be dependent on the ability to pay.

Broad principles of open access should be applied to research infrastructure as well as research content (data and texts) wherever possible as the tools may be as valuable to other sectors of society such as government, civil society organisations and industry which are often unnecessarily excluded. This is a major issue for developing linkages with organisations outside of universities and increasing industry engagement with research. The National research infrastructure roadmap should include a statement on FAIR access with a requirement for publicly funded research infrastructure to provide public access and use (the degree may vary) or provide a reason for why this is may be limited or not possible. This includes publishing results from the use of infrastructure openly and ensuring collections of data and documents are free to access wherever possible.

Clarity of expectations and practices for Intellectual property and moral rights are key to maximising the benefits from the use of research infrastructure. However, currently for research publications the norm is that researchers sign away these rights, especially publishing rights. There is an opportunity to put in place now polices and practices that could substantially enhance the useability of information derived from National Infrastructure projects. Policies and practices could include

a. make research publications conducted using research infrastructure projects immediately free to read at the time of publication
b. make research data directly related to research publications as open as possible and as closed as necessary, in accord with the Australian Government Public Data Policy Statement
c. apply appropriate AusGOAL licenses and international metadata standards to research outputs to ensure accessibility, interoperability and reusability
d. ensure that authors/creators retain all necessary rights to enable the authorisation of publication and re-use in any format at any time
e. support the development of incentives for researchers to make research outputs available in accord with the above principles

Question 9: What should the criteria and funding arrangements for defunding or decommissioning look like?

Question 10: What financing models should the Government consider to support investment in national research infrastructure?

The pathway for research development is often presumed to be one of commercialisation however we would encourage the Government and the University sector to consider that there are a range of business models that may be suitable including those that are not-for-profit such as social enterprises, co-operatives and public service mutuals. The burden of maintaining research infrastructure such as databases and software tools can be onerous once the initial investment has ended and the expectation that an income from sales or services will be found often unrealistic. As a result many projects end and some completely disappear, resulting in a huge loss of public investment and use value.

Supporting and rewarding universities to develop not-for-profit, collaborative enterprises may provide a way for industry, civil society organisations and governments departments and agencies, as well as multiple universities to engage with research infrastructure and research investigations in new and innovative ways outside of the narrow focus of economic gain. Open source software is an example of how successful this process can be and the potential for national and international partnerships to work together when profit is not the driver and open collaboration beyond universities is supported.

Question 11: When should capabilities be expected to address standard and accreditation requirements?

Interoperable metadata standards and recognised schemas and classification systems should be adopted or where unavailable, developed for all databases and collections supported as national infrastructure. Effort should be made to ensure Australian terminology and taxonomies are developed while also aiming for international compatibility.

Interoperability, standards and linked data

There is a need for an investment in national database infrastructure that is built on internationally recognised identifiers, metadata schema and standards, and linked data ontologies and that are capable of, or integrated with, systems able to undertake automated indexing and text mining. These two elements can we developed independently however their real potential comes with integration where structured metadata is able to combine with unstructured texts to produced further structured data that is then able to be reintegrated into database systems.

Linked data is an approach to digital information infrastructure which aims to enhance the utility of data on the web by making it more consistent, structured and linkable, and therefore discoverable and able to be analysed within and across information systems. Key elements are the need for information rich structured data, standardised classification systems and stable URIs for linking across information systems. Linked data relies on nationally and internationally recognised, well managed ontologies and identification systems with globally unique and stable URIs. Ontologies enable the unambiguous identification of entities in heterogeneous information systems and assertion of applicable named relationships that connect these entities together. Linked data provides the opportunity to go beyond standard bibliographic information on publications and data to consider every piece of structured and unstructured information as a potential source of data that can be analysed and visualised to provide new knowledge. Not only documents, but people, organisations, projects and policies can be enriched with the addition of spatial, temporal, topical and other details.

Question 12: Are there international or global models that represent best practice for national research infrastructure that could be considered?

Question 13: In considering whole of life investment including decommissioning or defunding for national research infrastructure are there examples domestic or international that should be examined?

Question 14: Are there alternative financing options, including international models that the Government could consider to support investment in national research infrastructure?

Health and Medical Sciences

Question 15: Are the identified emerging directions and research infrastructure capabilities for Health and Medical Sciences right? Are there any missing or additional needed?

We agree with the general outline of emerging directions and capabilities however we would also like to see recognition that there are vast amounts of crucial health grey literature that is currently uncollected or curated, difficult to find and evaluate which is causing costly delays to conducting systematic reviews and other research, particularly into population needs and public health.

Grey literature collections for health

There is a need for large scale, open access, interoperable collections of health related grey literature that is able to be evaluated and rated, text mined for automatic indexing and entity extraction and interrogated with new research tools as they develop. An example is the Cochrane Collaboration's use of text mining to evaluate if RCTs have been used as the evidence source in systematic reviews that have been conducted. The biomedical field has been leading the way in text mining capabilities and these techniques can be brought into the public health grey literature corpus to provide new insights as well as huge efficiencies in curation and analysis.

eHealth is an example of an emerging direction with many policy implications beyond health including legal and ethical issues, health policy and management, digital communication capacities, regional and remote community access, digital literacy and many others. Bringing these resources together will be a huge productivity saving and ensure that documents and data can be analysed together. APO is in a position to extend its existing collection if public health, health policy and other resources to provide collaborative, open access storage and tools for grey literature documents and link them to relevant data sources and tools.

Question 16: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Question 17: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Health and Medical Sciences capability area?

Environment and Natural Resource Management

Question 18: Are the identified emerging directions and research infrastructure capabilities for Environment and Natural Resource Management right? Are there any missing or additional needed?

We agree with the need for greater emphasis on people and networks/collaboration and integration across people, networks and infrastructure. There is currently a great deal of concern that there needs to be a better evidence-base for policy decisions on the built and natural environment with a range of projects occurring around the world. APO is working with the CRC for Low Carbon Living to develop a knowledge hub that will be interoperable with the APO repository and other data sources such as CSIRO to bring together resources but also provide filtering services that will assist government and industry to evaluate evidence and conduct systematic reviews. Building on existing investments in this way is an example of how infrastructure can cost effectively scale up to the national and international level. At the same time we are working to ensure that any system created will be compatible with existing and new projects occurring around the world.

We believe this may be a model for creating specialised portals that filter evidence for target audiences while at the same time enhancing and building on the underlying infrastructure and document and data collections and tools.

Question 19: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Question 20: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Environment and Natural Resource Management capability area?

Advanced Physics, Chemistry, Mathematics and Materials

Question 21: Are the identified emerging directions and research infrastructure capabilities for Advanced Physics, Chemistry, Mathematics and Materials right? Are there any missing or additional needed?

Question 22: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Question 23: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Advanced Physics, Chemistry, Mathematics and Materials capability area?

Understanding Cultures and Communities

Question 24: Are the identified emerging directions and research infrastructure capabilities for Understanding Cultures and Communities right? Are there any missing or additional needed?

The emerging directions and infrastructure capabilities are generally appropriate however there are important infrastructure resources for digital humanities that are not recognised in the issues paper.

Digitisation and web archiving

Large-scale digitisation and web archiving systems need to be established extending current capabilities. These need to be owned and operated by the University sector and able to be accessed and used by researchers rather than limited by the resources, timelines, institutional priorities and copyright issues of libraries and archives. The physical collections are vast and therefore choices and priorities about what is to be digitised will be need to be set. These choices should be able to be made by researchers, many of whom have their own stores of key documents, many of them unique.

Examples of this are occurring in projects developed by APO and AustLII (the legal resources database) where both organisations are investing in scanners and digitisation tools in order to provide researchers with the opportunity to ensure key historical research materials in law and public policy are transformed into digital content which will then be able to be analysed and manipulated in ways that are impossible in physical forms. With ARC LIEF funding, APO is working with the US based internet archive to purchase infrastructure that will be available to researchers in urban and regional policy and planning, infrastructure, media and communications and public administration who will be able to prioritise and arrange for the resources they want to be digitised.

These types of university-based digitisation projects could be expanded to other disciplines and made open to use by government, industry and civil society organisations, which are similarly needing to digitise and analyse key historical policy documents and bring them together into a consolidated collection able to text mined and analysed for years to come. APO and Austlii's digitisation projects show both the research need and possible pathways toward implementation of large-scale research-focussed university-based digitisation operations. Such operations and any new facilities should be supported to operate as a network across universities and partner and integrate with the outputs of the GLAM sector, CSOs, government and industry contributing content to a publicly accessible collection of national and international research significance.

Question 25: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Australia has signed up to the international Open Government Partnership (OGP) and is currently developing a national action plan. This should be a consideration in developing the roadmap as information access, data management and transparency are key elements on the agenda. There is a huge need for improved management and accessibility to a range of information resources as well as the collection of new kinds of data and improved processes. This presents many opportunities for researchers to benefit from improve access to materials previously difficult if not impossible to work with. In addition governments will need advice and support to ensure their approaches are compatible with the needs of researchers.

Question 26: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Understanding Cultures and Communities capability area?

National Security

Question 27: Are the identified emerging directions and research infrastructure capabilities for National Security right? Are there any missing or additional needed?

Question 28: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Question 29: Is there anything else that needs to be included or considered in the 2016 Roadmap for the National Security capability area?

Underpinning Research Infrastructure

Question 30: Are the identified emerging directions and research infrastructure capabilities for Underpinning Research Infrastructure right? Are there any missing or additional needed?

Curated collections of born-digital grey literature

A vast amount of born digital content is not collected or curated and since the introduction of the internet huge amounts of materials have disappeared from online access in what is known as the digital black hole. Web archiving via Pandora and the Internet Archive has helped to stem the loss to some extent but these systems do not replace the need for fully curated full text and data collections. To date the approach has been piecemeal at best with a range of funded projects attempting to collect, curate and provide access to research resources. Many of these are long running projects attempting to find sustainable models of operation between grants and university partnerships and could disappear at any stage without recognition and investment as national infrastructure.

Public policy and practice document and data collection

Australia's ability to respond effectively to future economic, social and environmental challenges depends on national capacity to develop and implement efficient and effective public policy. The national research infrastructure roadmap should aim to build on existing investments in open access knowledge infrastructure to develop large-scale collections of policy documentation and data that will support new tools for problem solving and analysis. Efficient universal access to historical and archived policy material requires critical research infrastructure that will support innovative and applied approaches to Australian public policy research.

A wide ranging collection of full text digital resources now exists at APO.org.au focussing on public policy and practice research across not only the humanities and social sciences but also science and technology, health and the environment. Established in 2002, APO has already had public investment of well over $5 million and includes 30,000 records from 5000 organisations and 12,000 authors. The services has expanded to include New Zealand research and policy resources and increasingly covers research from around the world given the nature of global policy issues – it therefore has the potential to expand from being a national resource to an international collaboration with the right investment and strategic approaches.

APO is currently developing the capacity to host datasets and link these with publications and has led the way in minting DOIs for grey literature in Australia. With a relatively modest investment in smart technologies and curatorial systems, and a collaborative, networked approach across universities, government, CSOs and the private sector, APO’s potential to be national and international public knowledge infrastructure could be realised. This would have huge benefits for many disciplines that have identified this issue in the discussion paper such as health, the environment and national security as well as for humanities and social science researchers – all of whom waste inordinate amounts of time trying to find, evaluate, collect and analyse digital content scatted across numerous websites. APO’s cross sectoral approach and open access principles also ensure it is heavily used and highly valued by industry, government and civil society organisations who would also benefit from any expansion of the enterprise.

APO should be recognised and supported as national and international infrastructure and a world leading grey literature repository supporting applied research across the humanities, social sciences, health and science.

Text mining systems for unstructured data

While structured data is an essential tool for researchers in every discipline, unstructured and semi-structured data in the form of textual materials remains an absolutely essential research resource for all research but particularly for those in humanities, arts and social sciences. Technologies such as text mining or text analytics, the process of deriving high-quality information from text, allows unstructured texts such as documents, social media, etc to be transformed into structured or semi-structured texts, providing new analytical possibilities and discoveries for many disciplines. This is now well-known in business and industry, biomedical research and security but has not had the same focus in other disciplines despite similar needs and potential benefits. In the humanities and social sciences which relies on huge amounts of highly varied textual material produced by civil society and governments the potential may be even greater given the huge amounts of unstructured texts that continue to be produced and circulated in these arenas. Establishing integrated systems for the curation and automatic indexing and text mining of digitised and born-digital texts would then lead to the adoption of new methods of analysis and visualisation and new discoveries using the many tools already available and others that will come on the market.

A national centre for text mining with a particular focus on humanities and social sciences should be established to develop the tools and work with collections and databases to establish best practice text mining capabilties. Similar centres exist in the UK (NaCTeM) that have been able to expand their services to industry, and civil society organisations. They are however mainly focussed on biomedical and therefore there are great opportunities to bring concerted focus to humanities and social sciences in this area as tools and classification systems need to be tailored to specific domains.

Social media data collection and analysis is already occurring with the TRISMA project, providing insights for health, national security, social cohesion, emergency management and numerous other issues but requires ongoing investment in storage and analytical capabilities.

Relative to science capability areas, the costs associated with the development and maintenance of these types of infrastructure is modest, the potential uses and scope for expansion are substantial, yet there is no dedicated funding source. The costs are supported through modest funding on a semi-regular basis from ARC LIEF and institional support with around $5 million spent on APO to date. However, this limits the relevance and potential of these national resources. Recognition and support for APO as national infrastructure would allow a major expansion of the repository which is already interoperable with other systems such as Trove, World Catalogue. An investment in scale and infrastructure would also support the APO database to become an international collaboration and model for grey literature collecting in other regions such as Africa, Asia and Latin America, as well as potentially in Europe and North America.

Question 31: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Question 32: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Underpinning Research Infrastructure capability area?

Data for Research and Discoverability

Question 33 Are the identified emerging directions and research infrastructure capabilities for Data for Research and Discoverability right? Are there any missing or additional needed?

Question 34: Are there any international research infrastructure collaborations or emerging projects that Australia should engage in over the next ten years and beyond?

Question 35: Is there anything else that needs to be included or considered in the 2016 Roadmap for the Data for Research and Discoverability capability area?

Other comments

If you believe that there are issues not addressed in this Issues Paper or the associated questions, please provide your comments under this heading noting the overall 20 page limit of submissions.