APO ANALYSIS & POLICY OBSERVATORY

Linked Semantic Platforms for Policy and Practice

ARC LIEF Project 2018 - 2019

Summary

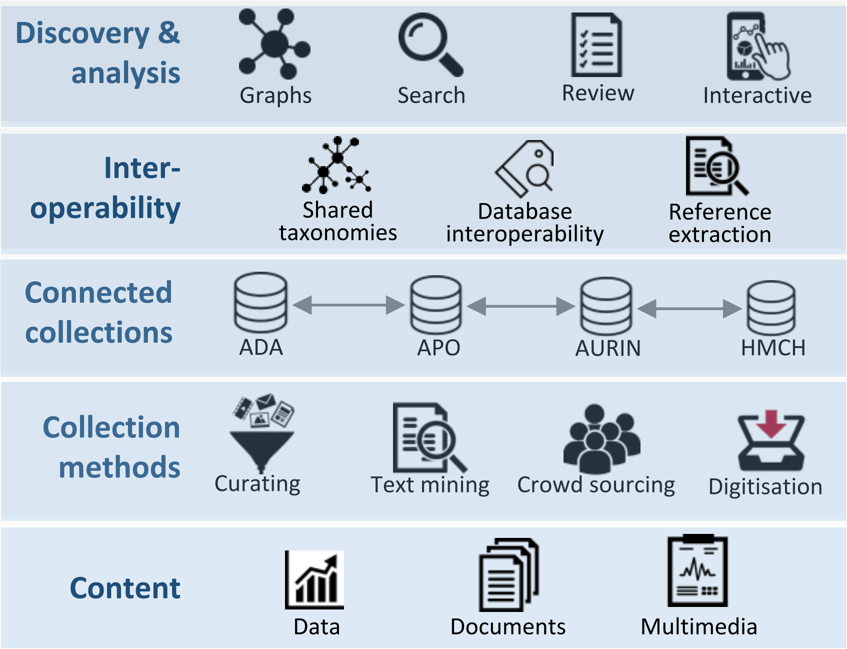
The Linked Semantic Platforms project is an ambitious, two-year multi-institutional and multi-database project that aims to revolutionise the way researchers are able to access, and analyse policy documents and data. The project aims to develop the next generation of decision-support tools for interdisciplinary research on critical public policy issues. The project will apply linked open data, knowledge graphs and collaborations across existing research infrastructure projects to improve interoperability across major social science databases and develop new analytical tools that will transform the research capabilities for evidence-based policy making. The project focus areas include sustainable built environments and transport in urban and regional communities, social care and health in the community, work and wellbeing, digital inclusion and digital health.

There are an increasing number of critical societal challenges and opportunities facing decision makers in public and private sectors that embody complexity and linkages that are multi-level, multi-scale, multi-stakeholder, multi-disciplinary. The information and knowledge required to undertake this kind of interdisciplinary policy research is often in the grey literature and across multiple datasets which are diverse, dispersed and difficult to find and analyse with traditional methods and tools (Lawrence et al 2014).

The Linked Semantic Platforms (LSP) aims to solve this problem through integrated systems, national and international collaborations and cutting edge information technologies involving four major platforms – Analysis & Policy Observatory (apo.org.au), the Australian Data Archive (ada.edu.au), the Australian Urban Research Infrastructure Network (aurin.org.au) and the Home Modification Clearing House (homemods.info).

The LSP project will use text mining and expert curators to create large-scale open access collections of key policy documents and data (grey literature), house them in linked databases with interoperable ontologies and standards, and apply cutting edge technologies such as semantic graphs, open notebooks and open peer review to enable researchers to see the relationships between entities in ways that are not currently possible.

From documents to data: maximising the benefits of textual materials



The digital world is growing at an exponential pace from two billion objects in 2006 to a projected 200 billion by 2020. What is often overlooked in discussions of Big Data is that an estimated 80 to 90% of the data in any organisation is to be found in 'unstructured data', text files, PDFs, presentations, web pages etc, and that this is growing faster than structured data. With a deluge of unstructured documents and diverse data to sift and analyse, researchers working on multidisciplinary public policy issues urgently need new digital research methods and integrated data solutions if they are to provide the evidence needed to have an impact on policy decisions and practices. To do this a comprehensive, multidisciplinary knowledge base is needed, along with intelligent online analytic infrastructure and cutting edge semantic knowledge systems. This will enable university researchers, as well those in government industry and civil society to analyse the wealth of information and explore the relationships and connections between diverse entities in a way that is not currently possible.

Partner organisations

Swinburne University of Technology, University of South Australia, RMIT University, University of Melbourne, Australian National University, University of NSW, and the Australia and New Zealand School of Government.

Chief investigators

Swinburne University: Professor Jane Farmer, Professor Peter Newton; Professor Penelope Schofield; Professor Peter Graham; Professor Timoleon Sellis;
RMIT University: Professor Julian Thomas; Professor Jago Dodson; Professor Mark Sanderson
UniSA: Professor Kerry London; Professor Ian Olver; Professor Maureen Dollard; Professor Susan Luckman;
University of Melbourne (AURIN): Professor Richard Sinnott
ANU (ADA): Dr Steven McEachern
UNSW: Associate Professor Catherine Bridge

Project team

Project manager: Amanda Lawrence, Director, APO
alawrence@apo.org.au

Technical lead: Camilo Jorquera, Senior Developer, APO
cjorquera@apo.org.au

Australian Government
Australian Research Council

Project details

This project involves four key strategies: **Collections, Connections, Discovery and Analysis**.

Collections

Research on public policy issues often involves the collation and synthesis of grey literature - reports and publications produced by NGOs, government departments and agencies, research centres, think tanks and so on. Many of these are not curated or managed in a way that allows for efficient analysis or correlation. Collections will be developed using four main methods: Expert curators, text mining, crowd-sourcing and digitisation.

Creating specialist collections requires a level of domain expertise to understand the specific needs of researchers and the types of resources required such as international case studies, evaluations, submissions, technical reports, historical materials, comparative data, government reports and policies. Collection curators will be employed across three partner universities, Swinburne, RMIT and UniSA, to select and add resources within the overlapping themes of social and physical and digital infrastructure.

Given the scale of materials being published online that require curation and long term management, the LSP project also involves applying text mining and entity extraction techniques to create structured data and automatic classification of resources. This has huge potential for transforming the way policy research is conducted.
APO and the other platforms all receive contributions from research networks and partners and this is a valuable part of collection development and user engagement.

APO's digitisation of print publications will continue using the Internet Archive Table top scribe. This involves OCR processing by partners at the Internet Archive (archive.org) in the US which hosts a collection of APO digitised resources (https://archive.org/details/apoanalysisandpolicy).

Social infrastructure collections will cover issues such as social care, health in the community, work and wellbeing, social service delivery, community services and disability policy. Resources collected include evaluations, case studies, strategic plans, surveys and data including interactive access to datasets produced by Prof Maureen Dollard's on work and wellbeing (the Australian Workplace Barometer) and Julian Thomas' on digital inclusion.

Physical infrastructure collections will cover issues such as urban and regional planning, housing and built environment co-benefits, precinct design, smart cities policy. Resources types include: documents associated with urban and regional strategic plans, transport, infrastructure, local and state government reports, historical documents, and case studies on sustainable building design, reduced carbon emissions, housing strategies and affordability and other key issues.

Digital infrastructure collections will cover issues such as digital health, digital inclusion, knowledge translation and communication. Resources types include: digital health and inclusion strategies for self-management and promotion; comparative case studies and evaluations from around the world on eHealth and knowledge translation initiatives; online health applications; internet policies and strategic plans; industry reports and white papers; comparative data on education; collective intelligence projects and government consultations.

Connections

The LSP project is a significant and ground-breaking step in the history of two long standing national eResearch infrastructure projects, APO and ADA, and provides a unique opportunity to connect policy grey literature and data in a way that will have enormous benefits both for researchers and the wider community. The project also continues the collaboration between AURIN and APO established with previous ARC LIEF grants and connects APO and HMCH, building on collaborations occurring as part of the CRC for Low Carbon Living Knowledge Hub project Built Better (builtbetter.org).

The LSP collaboration across these systems will support researchers to easily find related data and publications through establishing interoperability in three key ways:
1) shared taxonomies;
2) database interoperability
3) text mining for references and enhanced metadata

Shared taxonomies involves the development of a policy terms based on FAST (Faceted Application of Subject Headings) an open linked data classification system developed and managed by OCLC based on Library of Congress Subject Headings (LCSH). The policy subset of FAST will be developed using tools such as the ANDS-supported PoolParty software (poolparty.biz) or the open source EU-funded VocBench (http://vocbench.uniroma2.it) and will be made open and accessible to all via the ANDS Research Vocabularies Australia website (https://vocabs.ands.org.au/). This work will also involve investigating cross walks and compatibility with other vocabularies such as Geonames, EuroVoc (eurovoc.europa.eu), the UN thesaurus and other vocabularies and explore the potential to efficiently add rich metadata via linked data ontologies such as spatial and demographic characteristics of cities and towns.

2) Database interoperability aims to establish a service for linking policy documents held at APO with the underlying data hosted in ADA, AURIN or APO itself. This will allow users of both facilities to easily identify and access the relevant materials associated with a publication or dataset, informing activities such as systematic reviews, meta-analysis and secondary analyses of data produced from APO-published research. Under this activity, APO and ADA will establish linked data services using the API services available at each facility to connect datasets and publications in the two collections. Each facility will support this access through embedding of DOI-based links within related metadata records for datasets and publications. The linked records will be used to update the ANDS RDA database and the DataCite service to enable further data discovery.

3) Reference extraction. Given the diversity of publication types and formats, there is currently no easy way to view and explore the citations in most policy grey literature. An exploratory aspect of this project is the use of text mining to extract references from publications hosted in APO and use these to provide further ways of relating and linking the evidence-base. This would assist with the connection between publications and data citations, allowing researchers to follow the evidence trail. Given the unstructured and immensely diverse nature of grey literature publishing, as well as the lack of bibliometric information and publishing standards, the project aims to develop a prototype for a key corpus of documents.

Discovery and Analysis

Five elements are involved in the Discovery and Analysis phase of the project:
1) graph databases;
2) semantic search;
3) open peer review and evaluation systems;
4) hosted interactive data; and
5) open notebooks.

Graph databases or 'knowledge graphs'

The power of graph databases is now becoming apparent. Companies in public health are starting to use graph-facilitated software to solve business problems. Some of the world's most knowledge-intensive organizations, including multinational banks, media companies, space agencies, and logistics companies, are also using graph databases, and intelligence agencies have been using them for a decade (PWC 2012). This project offers an opportunity for social science researchers and the wider community to explore the benefits of this cutting edge technology in an open way that can continue to be built on by others after the project has been completed. A graph database allows numerous connections or *relationships*—how people, places, and things relate to one another – to be mapped, visualised and analysed in a way not possible in a relational database. Relationship richness of this kind boosts the integration potential and the contextual relevance of the data being represented enabling researchers to draw inferences from data that is not explicit.

2) Semantic search: The LSP will utilise the enhanced taxonomies and rich metadata developed in this project to improve search relevance and retrieval using Solr and semantic search software that expresses ranking in terms that the researchers can associate with meaningful information.

3) Open peer review: The collaboration between APO and HMCH involves a shared interest in implementing interoperable, internationally-recognised open peer review or other evaluation systems that can be applied to grey literature to support evidence-based research and systematic reviews. Open software such as the Open Peer Review Module (OPRM) developed by Open Scholar (URL) for DSpace repositories or annotation systems such as Hypothes.is will be assessed for adaption to create collective intelligence tools that harness researcher and community expertise.

4) Open notebooks: The collaboration between APO and AURIN involves working with the AURIN dynamic publishing environment being built using the open source Jupyter software to publish dynamic enhanced publications integrating text, data and charts.

5) Interactive Data: will also involve AURIN and APO collaborating to develop hosted interactive data publishing tools for **Maureen Dollard**'s Australian Workplace Barometer data on workplace wellbeing and Julian Thomas' Digital Inclusion index data.