A Data Harmonisation Framework: Organising Political Science Data for Research in Australia and New Zealand

Framework version: 1.0

Framework contributors: Steven McEachern, Ingrid Mason
ANZLEAD project partners
ADA team members

FRAMEWORK CONTENTS **3**

RESEARCH PRACTICE **3**

PROBLEM 3
DATA SOURCES 4
RESEARCH REQUIREMENTS ..... **ERROR! BOOKMARK NOT DEFINED.**

INFORMATION MODELLING **6**

HARMONISATION GOALS 6
MODELLING APPROACH 7
RELATED MODELS 7
CONCEPTUAL MODEL 7
DATA MAPPING ..... **ERROR! BOOKMARK NOT DEFINED.**

DATA MANAGEMENT **9**

DATA PROVENANCE 9
DATA VERSIONING 9
DATA GOVERNANCE 9
DOCUMENTATION STANDARDS 9
DATA ACCESS 9

Framework Contents

This framework operates in two sections: the first section outlines research practices in political science and requirements for a data harmonisation framework and national data asset to be developed and the second section outlines the rationale for the information modelling and techniques for curating and combining data to establish a national data asset.

Research practice

- Problem outline
- Secondary data sources
- Research requirements

Information modelling

- Harmonisation goals
- Modelling approach
- Related models
- Conceptual model
- Data integration
- Data mapping

Data management

- Data provenance
- Data versioning
- Data governance
- Documentation standards
- Data access

Research Practice

Political science research into political and social systems in Australia and New Zealand uses empirical research methods and national data sources to aid enquiries into leaders, elections and democracy. Four populations are commonly investigated to understand the dynamics of political activity in Australia and New Zealand:

- The Voting Public: voters, citizens and the Australian and New Zealand populations
- Political Elites: candidates, parliamentary members and bureaucrats
- Elections: Federal and State election results and enrolments
- Electorates: geographic and administrative characteristics of the Federal and State electorates

Problem

Political science researchers are active and regular users of secondary public research and administrative data. There are well established data sources, national in scope and extensive in time (from Federation to current day), covering a breadth of the political

science discipline. But these datasets have largely been developed in isolation, by different organisations, using bespoke vocabularies, identifier systems and data standards. Significant effort is required by researchers to find, access and integrate disparate sources of data to allow the study of complex political science problems. Particularly the integration of these sources across units of analysis, jurisdictions (that can change) and over time. For example: there is no shared or common controlled vocabulary of political parties used across the discipline (Australian Liberal Party, Australian Liberals, Liberal Party, Australian Liberal Party to name a few). Add to this semantic variation in terminology the fact that electoral boundaries frequently change which means one jurisdiction can lose/gain data from year to year. The result is that significant harmonisation work is required and repeated each time a researcher wishes to analyse data, for example, the changes in voting patterns for a political party over time.

Data Sources

Multiple data sources that already serve political science research in Australia and New Zealand are the basis upon which a data harmonisation process is being established. A selection of these sources are included in the first release of the ANZLEAD data portal.

Dataset	Custodian/Owner	Coverage	Country	ANZLEAD Release
ANUPoll	Australian National University	2008-Present		Some
Australian Election Study	Australian National University	1987-2022	Australia	All
New Zealand Election Study	University of Auckland and Victoria University of Wellington	1990-2020	New Zealand	All
SmartVote (?)	Australian National University		Australia	
Australian Candidate Study	Australian National University		Australia	
Cooperative Australian Election Study	University of Sydney		Australia	
SEDEPE	Australian National University (and others)		International	
AEC Enrolment Data	Australian Electoral Commission	1901-Present	Australia	1993 onwards
AEC Spatial Data	Australian Electoral Commission		Australia	
ECNZ Enrolment Data	Electoral Commission (of New Zealand)		New Zealand	

Dataset	Custodian/Owner	Coverage	Country	ANZLEAD Release
ECNZ Spatial Data	Electoral Commission (of New Zealand)		New Zealand	
Australian Elections Database	University of Western Australia		Australia	
Australian Election Results	AEC, State Commissions, TBC	1993-Present	Australia	Subset 1993 onwards
New Zealand Election Results	TBC		New Zealand	

Information Modelling

Information modelling involves understanding the requirements of information users and information that arises in a specific domain and applying information structuring techniques to enable that information to be managed, made interpretable and usable. Modelling may involve assessing structuring requirements at different levels to understand:

- The relationship between concepts embedded in metadata to enable domain knowledge modelling of vocabularies that are captured as *part of* a dataset.
- The relationship between metadata and value data in datasets to support data analysis *within* a dataset.
- The alignments in metadata and value data in multiple datasets to fuse data and enable data analysis *across* datasets.
- The aggregation of metadata about multiple datasets to make the data more discoverable and identify semantic and analytic relationships *between* datasets.
- The relationship of metadata about metadata that describes datasets to act as infrastructure that enables *organisation* of datasets.

The documentation and interpretation of information involves the capture of key concepts, the relationships between concepts, any constraints, structures, and rules that encode and express important information in the arrangement of concepts. Information modelling is undertaken with a view to understanding the impact of semantics on any operations that might be undertaken upon that information that support system functions or actions and/or generate insights (analytics).

Harmonisation Goals

The main goals of data harmonisation (based on key research datasets) are to:

- Standardise key units of analysis, document the procedures and metadata requirements.
- Create and publish a controlled vocabulary for core political and demographic information.
- Develop a model for linking and harmonising data as part of curating a national data asset.

Modelling Approach

[Describe the approach to interpreting the data sources]

The approach to modelling involves iterative investigation of the:

- Relatedness of datasets in terms of coverage e.g., units of analysis.
- Extent of documentation e.g., abbreviations or concept definitions.
- State of a dataset over time e.g., consistency or variation.
- Integration points in a dataset e.g., concepts, variables or formal vocabularies.

Datasets are reviewed individually and collectively to identify:

- Key units of analysis
- Core political and demographic information

Datasets
Election Study, Australian Electoral Commission

Related Models

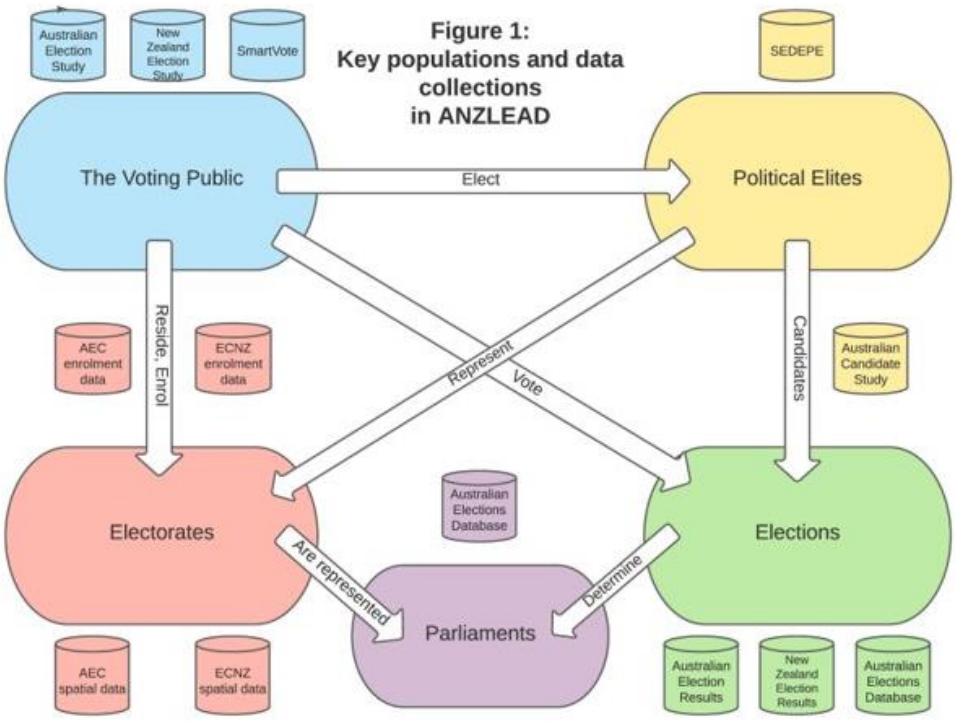
[Describe other relevant information models (ontologies) used to capture information relevant to political science including POWER, UK Parliament ontology and AGP glossary]

https://ukparliament.github.io/ontologies/
https://link.springer.com/content/pdf/10.1007%2F978-3-642-22056-2_51.pdf
https://elections.uwa.edu.au/glossarysearch.lasso

Conceptual Model

A high-level mud-map of concepts and relationships was defined to:

- ensure that research requirements were focused in the common areas for investigation
- understand potential coverage from key populations represented in the data collections



This conceptual model was then converted into a set of core concept definitions and a knowledge model. These concept definitions have been published in demonstration form in Research Vocabularies Australia (.

Data Integration

Following the establishment of the core ANZLEAD concepts, controlled vocabularies will be developed to support each of the concepts included in the model. In the first phase of ANZLEAD, the focus of these vocabularies will be in two areas – political parties and electorates.

Data Management

Data Provenance

[Describe how the mapping/transforming process is being documented and scholarly links are established and can be understood between data sources and the asset]

Content included in the data asset is sourced from public sources, including the Australian Data Archive and the Australian Electoral Commission. Where data is integrated, a content page reflecting the original author source of the data and the source location is included in the collection. A sample source table is included below.

Data source	Author/Investigator	Source location

Data Versioning

New versions of datasets will be versioned using the Australian Data Archive versioning rules, and added to the relevant ADA Dataverse. On publication, the version of the data will be incremented to reflect the current version, and records of previous versions preserved in the ADA Dataverse system

Data Governance

The ANZLEAD data assets are managed by the Australian Data Archive on behalf of the data producers. The CADRE-IRISS-ANZLEAD management committee will be responsible for overall project governance, with operational maintenance provided by the Australian Data Archive.

Documentation Standards

Datasets harmonised with ANZLEAD will be documented using the DDI standard. Study metadata and data will be published in the Australian Data Archive catalogue – https://dataverse.ada.edu.au. Variable level metadata will be published in the AUSIRISS registry – https://mdr.ausiriss.org.au.

Data Access

Access to ANZLEAD metadata will be publically accessible. Access to ANZLEAD data assets is based on existing data access requirements for the underlying datasets, and may include both open and restricted access processes.