Moving data around: integrated research workflows for curating, and publishing data

Maude Frances (1), Steven McEachern (2), Daniel Bangert (1), Carolien van Ham (1), Janet McDougall (2), Luc Betbeder-Matibet
(1) UNSW Sydney
(2) Australian Data Archive, Australian National University

IASSIST 2017, Lawrence, Kansas, 23-26 May

Fast facts

| Founded | 1949 |
| Located in | Sydney |
| Number of Faculties | 8 |
| Faculties | Art & Design, Arts & Social Sciences, Built Environment, Business, Engineering, Law, Medicine, Science |
| Value Invested | $215m |
| Description | invested in campus infrastructure across 123 projects |
| Number of Wireless Stations | 4,200 |
| Description | wireless stations on campus and tripling of network speed |
| Number of Students | 14,392 |
| Number of Doctoral Students | 19,777 |
| Number of Alumni | 279,226 |
| Number of Universities | 47 |
| Number of Schools | 125 |
| Number of Defence Force Academy | 18 |
| Description | UNSW & external centres and institutes |
| Description | UNSW & external centres and institutes |
| Description | UNSW & external centres and institutes |

| Description | UNSW & external centres and institutes |
Overview

• Researcher perspective: current practice and needs
• Research infrastructure project for curation and publication of data
• Infrastructure and workflow
• Benefits
• Next steps

Current practice: curation and publication

• Most of my colleagues publish their datasets on their own (non-institutional) websites, along with documentation such as the codebook and publications that used the data
• Larger group efforts for data-collection publish via own institutional website, i.e. Varieties of Democracy, Polity IV, Freedom House
• Some also deposit datasets at data archives like Harvard Dataverse.
• Need to balance recognition, impact, and efficiency.
Data on Elections, Democracy and Autocracy (DEDA)

- The Quality of Elections in Democracies and Autocracies dataset. Election and election fraud in over 900 elections in 97 third and fourth wave regimes between 1974-2012, including data on presidential and parliamentary election timing, results, electoral fraud, characteristics of opposition and incumbent, turnover, and presence of civil wars and coup d'état.
- Developed for PhD (van Ham 2012)
- Currently being expanded until 2016 and to over 120 countries, funded by the Australian Research Council DECRA “Getting Elections Right”, RG142911.
ANDS High Value Collection Program

UNSW partners: School of Social Sciences, Office of PVC – Research Infrastructure, UNSW Library

External partner: Australian Data Archive, Australian National University

Goals of the project

1. Curate and openly publish several datasets on elections, democracy and autocracy collected by the researchers and colleagues at School of Social Sciences (SoSS) UNSW.

2. Create an institutional website on data on Elections, Democracy and Autocracy
   - enables sharing of data collected by researcher and UNSW colleagues on these topics
   - includes a macro data guide to other important datasets in this field
   - guides users to online data visualisation tools that would allow researchers and the general public to carry out simple data analyses online.
Section One: Title and Abstract

Study Title

Quality of Elections in Third and Fourth Wave Regimes 1974-2009

Study Abstract

The dataset Quality of Elections in Third and Fourth Wave Regimes 1974-2009 collects data on elections and election fraud on 885 elections in 97 third and fourth wave regimes between 1974 and 2009. The dataset is in country-year format and includes data at the country-level and at the election level.

At the country-level, the dataset includes indicators on region, year of independence, previous democratic experience, civil war, coup d'etats, regime breakdown as well as indicators on whether the regime held elections with de jure multi-party competition and de jure universal suffrage.

At the election level, the dataset includes indicators on presidential and parliamentary elections constituent assembly elections, election results, vote and seat shares of incumbents and
Dataset in ADA Dataverse

Quality of Elections in Third and Fourth Wave Regimes 1974-2009


Description
The dataset Quality of Elections in Third and Fourth Wave Regimes 1974-2009 collects data on elections and election fraud on 885 elections in 87 third and fourth wave regimes between 1974 and 2009. The dataset is in country-year format and includes data at the country level and at the election level. At the country level, the dataset includes indicators on regime, year of independence, previous democratic experience, civil war, coups, balance of power and regime breakdown as well as indicators on whether the regime held elections with de jure multi-party competition and de jure universal suffrage. At the election level, the dataset includes indicators on presidential or parliamentary elections, multi- or single-chamber elections, election results, vote and seat shares of incumbent and opposition, turnout; and the quality and integrity of elections. The dataset and research based on this data is described in-depth in van Ham (2012) and van Ham (2014).

Subject
Social Sciences

Keyword
Elections

Related Publication

etadata

<table>
<thead>
<tr>
<th>ADA form</th>
<th>DDI</th>
<th>ResData</th>
<th>RIF-CS</th>
<th>DataCite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study title</td>
<td>stl</td>
<td>dc:title</td>
<td>name[TYPE=primary] &gt; namePart</td>
<td>title</td>
</tr>
<tr>
<td>Study abstract</td>
<td>abstract</td>
<td>dcterms:abstract</td>
<td>description[TYPE=full]</td>
<td>description[TYPE=abstract]</td>
</tr>
<tr>
<td>Principal investigator</td>
<td>AuthEnty</td>
<td>foaf:firstName</td>
<td>foaf:lastName</td>
<td>creatorName</td>
</tr>
<tr>
<td>Geographic coverage</td>
<td>geogCover</td>
<td>core:GeographicLocation</td>
<td>ms21:spatialInformation</td>
<td>geoLocation &gt; geoLocationPlace</td>
</tr>
<tr>
<td>Time period covered</td>
<td>timePrd</td>
<td>dcterms:PeriodOfTime &gt; core:startDate core:endDate</td>
<td>coverage &gt; temporal &gt; date[TYPE=dateFrom] [TYPE=dateTo]</td>
<td></td>
</tr>
<tr>
<td>Time dimensions</td>
<td>timeMeth</td>
<td>... Not collected ...</td>
<td>dates[TYPE=created] &gt; date[TYPE=dateFrom]</td>
<td>date[DATEYPE=Collect ed]</td>
</tr>
<tr>
<td>Funding agency</td>
<td>fundAg</td>
<td>ms21:organisationName</td>
<td>fundingReference &gt; funderName</td>
<td></td>
</tr>
<tr>
<td>Distributor</td>
<td>distrbtr</td>
<td>... Not collected ...</td>
<td>citationInfo &gt; citationMetadata &gt; publisher</td>
<td>Publisher</td>
</tr>
<tr>
<td>Related publications</td>
<td>relPubl</td>
<td>ms21:hasRelatedPublication</td>
<td>relatedInfo[TYPE=publication]</td>
<td>relatedIdentifier[relationType=IsReferencedBy, IsDocumentedBy]</td>
</tr>
</tbody>
</table>
ResData

About ResData

ResData is a Library service for managing UNSW research data. The service has two main components:

- Research Data Management Plan (RDMPlan) - allows UNSW researchers and HDR Candidates to record metadata about research projects and associated research data. Completion of an RDMPlan is a requirement for allocation of storage in the UNSW Data Archive.
- A catalogue of UNSW datasets and collections of research materials.

Catalogue records in ResData are published to Research Data Australia, a discovery service for Australian research data.

The service is supported by the Australian National Data Service (ANDS) through the National Collaborative Research Infrastructure Strategy Program and the Elevation Research Fund (ERF) Super Science Initiative, as well as through UNSW Australia.

ResData ontology

Diagram showing the relationships between Dataset, Process/Methodology, Research Data Management Plan, Party/Contributor, Project/Grant, and Research Project Lifecycle.
Benefits

• Dataset recognized as research output through DOI
• Direct link to dataset at ADA allows publicizing dataset on my website, while data safely deposited, and findable via multiple pathways (ADA, website, library research collection)
• Increased usage of our data, not only by academics but also by students and practitioners/policymakers
• Increased visibility of UNSW research in these areas, supporting future grant applications and research collaboration with other universities in this field
• Provides a platform to expand the website in the future to include data visualization tools and data on peace & conflict, development, and human rights, providing an opportunity to showcase the key research strengths of the School of Social Sciences at UNSW
• Process of data depositing at ADA encouraged thorough documentation of data collection process and improved quality of codebook.

Recognition – Impact - Efficiency

• Recognition: Academic recognition for data collection effort
  – Usually via publications about the dataset or publications that analyse data
  – Data archive mints a DOI - dataset can be directly recognised as research output
• Impact: Platform for dissemination of data
  – Link to dataset from researcher’s website - links directly to data archive source
  – Institutional repository links data to related publications and disseminates to national metadata registry
• Efficiency:
  – ADA submission process highly efficient: description of data & consent form.
What do we need to do next?

1. Shift from PDF based to direct entry of metadata into Dataverse (out of the box DV functionality)
2. Review of data and metadata by ADA staff
3. Publishing the data at ADA
   - Automated minting of DOI through Datacite (indirectly through ANDS minting service)
4. Harvesting metadata from ADA to UNSW ResData
   a) OAI-PMH from Dataverse
      ▪ Mapping already done from DDI to ResData
      ▪ Need to configure harvester (in Dataverse)
   OR
   b) Harvest from Research Data Australia / Datacite