Platforms EOI: NSW Cultural Data Commons, a platform for the State and NSW public libraries to make their cultural datasets collectively available for researchers.

**Project title**

NSW Cultural Data Commons, a platform for the State and NSW public libraries to make their cultural datasets collectively available for researchers.

**Field of Research code(s)**

- 20 LANGUAGE, COMMUNICATION AND CULTURE
- 21 HISTORY AND ARCHAEOLOGY

**EOI Lead Name**

Robin Phua

**EOI lead Research Group**

Digital Experience Division

**EOI lead Organisation**

State Library of New South Wales

**EOI lead Email**

Collaborator details

<table>
<thead>
<tr>
<th>Name</th>
<th>Research Group</th>
<th>Organisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cameron Morley</td>
<td>Public Library Services</td>
<td>State Library of New South Wales</td>
</tr>
</tbody>
</table>

**Project description**

Cultural information is traditionally described and managed using domain-specific standards, such as MARC 21 for library collections and Rules for Archival Description (RAD) for archives. Whilst metadata facilitates efficient exchange of information between institutions, publishing this information in the context of data or datasets has been challenging. Cultural data includes digital assets such as manuscripts (i.e. First fleet & WW1 diaries), oral histories and pictures (i.e. local historic and contemporary collections).

We propose a data-first approach to harvesting, transforming and publishing documentary data from NSW libraries. The data will be available to researchers via a Cultural Data Commons, which standardises, categorises and brings together collections from participating institutions. Local studies collections and the State Library collection will be harmonised, providing opportunities to analyse, experiment and derive insights both individually or as a larger corpus for machine learning and natural language processing.

**Existing technology**

**Adopt**

- Existing ontologies/vocabularies will be used to transform and normalise disparate sources of cultural data. Adoption may include (but not limited to) investigation of linked data schemas such as Bibframe, schema.org, and...
DCAT
- Cloud compute and storage services may be adopted to address scale requirements for the project.

Adapt
- Extract, transform and publish scripts to harvest data via AWS Lambda.
- RDS - scalable relational db is used by the Library to store back-end data.
- APIGEE - API mediation layer for system and UX APIs will be adapted to support ETL processes and publishing of machine-readable data.
- Amplify (Library oral histories transcription platform) and image tagging platforms will be adapted and used as tools. Voyant Tools will be adapted for analysis.
- Elasticsearch will be extended to support search.

Build
- ETL scripts will be developed to harvest, normalise and publish data.
- User interfaces for the Commons will be developed using Node and React frameworks

Anticipated requirements

<table>
<thead>
<tr>
<th>Annual funding</th>
<th>$100,000 - $199,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed length</td>
<td>2 years</td>
</tr>
</tbody>
</table>

Other information

Other information you wish to provide
- Cultural datasets are not well represented in the current ecosystem of open data. Cultural datasets include digitised and born digital material contained in library special collections.

- Implementing a Data Commons will aid in the discoverability, use and re-use of local studies collections whilst also complimenting existing discovery platforms such as Trove.

- The development of this platform may also open potential collaboration and partnership with the wider GLAM sector in NSW.

Terms

I agree to the terms
- Yes