Platforms EOI: Language Data Analytics Platform (LaDAP): Access to Big, Sensitive and Complex Language Data in Australia for Humanities Researchers

Project title
Language Data Analytics Platform (LaDAP): Access to Big, Sensitive and Complex Language Data in Australia for Humanities Researchers

Field of Research code(s)
- 20 LANGUAGE, COMMUNICATION AND CULTURE

EOI Lead Name
Michael Haugh

EOI lead Research Group
Language Technology and Data Analytics Lab (LADAL)

EOI lead Organisation
The University of Queensland

EOI lead Email

Collaborator details

<table>
<thead>
<tr>
<th>Name</th>
<th>Research Group</th>
<th>Organisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marco Fahmi</td>
<td>Research Computing Centre</td>
<td>The University of Queensland</td>
</tr>
<tr>
<td>Martin Schweinberger</td>
<td>Language Technology and Data Analytics Lab (LADAL)</td>
<td>The University of Queensland</td>
</tr>
<tr>
<td>Simon Musgrave</td>
<td></td>
<td>Monash University</td>
</tr>
<tr>
<td>Nicholas Evans</td>
<td>ARC CoEDL</td>
<td>ANU</td>
</tr>
<tr>
<td>Jane Simpson</td>
<td>ARC CoEDL</td>
<td>ANU</td>
</tr>
<tr>
<td>Catherine Travis</td>
<td>ARC CoEDL</td>
<td>ANU</td>
</tr>
<tr>
<td>Nick Thieberger</td>
<td>ARC CoEDL</td>
<td>University of Melbourne</td>
</tr>
<tr>
<td>Monika Bednarek</td>
<td>Sydney Corpus Lab</td>
<td>University of Sydney</td>
</tr>
<tr>
<td>Celeste Rodriguez Louro</td>
<td></td>
<td>University of Western Australia</td>
</tr>
<tr>
<td>Dan Angus</td>
<td>DMRC</td>
<td>QUT</td>
</tr>
<tr>
<td>Julia Hickie</td>
<td>Trove</td>
<td>National Library of Australia</td>
</tr>
<tr>
<td>Amanda Lissarrague</td>
<td></td>
<td>AIATSIS</td>
</tr>
</tbody>
</table>
Project description

The objective of the Language Data Analytics Platform (LaDAP) is to enhance language-based digital research building on extant language data collections of (inter)national significance. The LaDAP will offer (1) a policy framework to integrate and facilitate access to collections of national significance not currently readily available to researchers, (2) web-based services that enables researchers to analyse types of language data that cannot be analysed using standard desktop software, and (3) an advisory and support service for researchers undertaking language analytics to facilitate broader uptake of LaDAP services and big data by researchers across the humanities.

LaDAP services will focus on: (a) Big data collections that are too large to be downloaded by individual researchers; (b) Sensitive language data collections that cannot simply be made publicly accessible due to rights restrictions (cultural, moral, copyright); (c) Linguistically complex data collections that otherwise require advanced computational skills to query. This will be provided through access to collections for targeted analysis using a web-based interactive development environment (via Jupyter notebooks).

Existing technology

Adopt

- A cloud-based Jupyter notebooks environment for text-based language data analytics
- Training and education resources provided by the Language Technology and Data Analytics Laboratory at the University of Queensland
- International best practice for access to language data through tiered Creative Commons licensing (commensurate with data access conditions)
- PIDs enabling fine-grained data citation in published research

Adapt

- Scripts for undertaking combinatorial queries of complex linguistic data developed through an ARDC Data and Services project (TD76)
- Connect to collections of language data maintained by AIATSIS, Australian National Corpus, ARC CoEDL, Digital Observatory at QUT, National Library of Australia, PARADISEC, Sydney Corpus Lab, and the prototype Language Data Commons of Australia, using APIs for language analytics (where access conditions allow)
- MoU policy framework for depositing data in Trove developed by National Library of Australia

Build

- Prototype of the Language Data Commons of Australia to host nationally significant collections of language data not currently hosted in repositories and accessible through APIs

Anticipated requirements

Annual funding

$200,000 - $299,000

Proposed length

3 years

Other information

Other information you wish to provide

LaDAP is part of the broader Language Data Commons of Australia initiative, which aims to more fully realise the social and economic possibilities of Australia’s rich linguistic heritage, and to ensure Australia is able to keep pace with best practice in the broader international research community. LaDAP is coordinating and liaising with related research communities submitting their EoIs in Digital Cultural Heritage and Social Media Analytics through participation in the HDRP EOI. It is also complementary to the EOI for culture and community specific accessibility submitted by UTS eResearch.
I agree to the terms

Yes