Data and Services Discovery projects - Transformative Data Collections

Title

AusTraits: a curated plant trait database for the Australian flora

Approach

This project builds on the “AusTraits” database, which we have been developing over the last 3 years. AusTraits standardises data from 100s of primary sources of plant trait data, thereby creating a transformative resource on the traits of Australia’s 22000 plant species. Funding was requested from the ARDC to strengthen and extend the AusTraits database in two directions: 1) Open call for new inputs from the scientific community, and 2) Improving standards, re-use and enhancing data interoperability.

To achieve the first goal we undertook the following activities: 1) Contacted researchers from across Australia advertising the project and asking if they had data to contribute. 2) Employed two Research Assistants (C Baxter, E Wenk) to assist with the processing of new data submissions. Our primary method of contact was directly emailing potential contributors. One or more of our team subsequently visited in person the following locations, where we gave presentations about the AusTraits resource and connected with contributors: University of Western Sydney (Sep 5), University of Queensland (Sep 9), Macquarie University (Sep 12), University of Western Australia (Sep 26), University of NSW Sydney (Sep 27). Slides from one of these presentations can be viewed at http://bit.ly/2ltVdM1

We received a very positive and substantial response to our emails and site visits. We were also able to add new datasets received from existing contributors. By the end of the project, we had incorporated 36 new data sources, including data for over 115k new trait records. Additionally, there were > 20 other sources still being processed at the end of the project, which will be completed in 2020. Funding from the ARDC was key in enabling to Wenk and Baxter to receive and process these into the AusTraits data resource.

Our second major activity was to improving standards, re-use and enhancing data interoperability. This was pursued through discussion with our ARDC contact (Kerry Levett). Arising from these discussions we undertook the following actions

- Reviewed our workflow, including data structures and metadata
- Conducted review of how FAIR the resource is and opportunities to enhance
- Identified suitable repository and license for broad scale data distribution
- Employed a casual (S Windecker) to advise on enhancing reproducibility of dataset construction, through the use of docker containers

- Discussions with Rowan Brownlee from ARDC on how we could formalise our vocabulary to further improve interoperability.

During the project we released v0.9.0 of the dataset, which included all data in AusTraits prior to the ARDC project, and v0.9.1, which includes some of the new data collected during the ARDC project (up to 20 Dec). Further datasets received during the project are still being processed by our team and will be included in a future release. These two versions of the data were deposited into the Zenodo data repository at http://doi.org/10.5281/zenodo.3568417.

FAIR

In reviewing our dataset and workflow with the ARDC we were pleased to learn that we already scored quite well in areas of Interoperability and some areas of Reuse (see Appendix 1, FAIR assessment). Together with the ARDC we are exploring how we can further improve Interoperability via formalisation of the vocabulary used within AusTraits. In areas of Findability, Accessibility we identified several points for improvement, which are described below.

Our plan has always been (i.e. prior to the ARDC project) to release data from AusTraits under an open source license (CC-BY) at the time our paper describing the database is published, thereby enabling easy reuse of the data. Prior to the paper being published we are keeping the resource in a private GitHub repository and under an embargoed release on Zenodo. This is necessary to ensure data is not used before the primary team assembling the data can both vouch for quality and progress a first analysis. We were almost ready to submit our paper at the start of the ARDC project, but the date for submission has been pushed back as a result of the ARDC project (from mid 2019 to early 2020), because we need sufficient time to process all the new submissions that have arisen from this project. Despite causing a short delay in making the data completely open, the eventual product will be more comprehensive and therefore useful to a wider variety of people. Currently we are planning to submit the paper towards in early 2020 for publication in mid 2020.

In the process of publishing the paper and making the dataset open, our data will become more Finable and Accessible via the following

- Globally unique, citable and persistent identifiers (DOI) for each version
- Unrestricted access to all versions of the dataset
- Data described by a comprehensive metadata record
- Using a formal machine-readable metadata schema
- Data is in one place but discoverable through several places.

Developments via this project alerted us to the fact that the metadata record could be published while access to the data was still embargoed, and that collection metadata could be published in more than one place, improving discoverability. We have therefore been able to significantly advance Findability and Accessibility ahead of our previous schedule.

Collaboration and coverage

Prior to ARDC project, AusTraits already had substantial coverage across Australia, with involvement from 127 contributors across 59 institutions. Even-so, the ARDC project has vastly increased the level of collaboration and national coverage. During this project we corresponded with researchers from the following institutions, most of whom were not previously involved, including:

- Western Sydney University: B Medlyn, D Ellsworth, P Rymer, D Tissue, R Smith, K Crous, B Choat, M Tjeolker, M Esperon-Rodriguez, R Nolan, J Powell, B Moore
- University of NSW Sydney: A Moles, W Cornwell, M Ooi
- University of Queensland: M Mayfield, J Dwyer, M Sams, C Bowler
- Queensland University of Technology: J Firn, J Mills
- CSIRO: K Mokany, D Metcalfe, A Richards, S Roxburgh
- Australian National University: A Nicotra, O Atkin, M Roderick, H Keith
- University of Western Australia: H Lambers, P Finnegan
- Murdoch University: R Standish, J Fontaine
- Australian National Botanic Gardens: L Guja
- University of Sydney: G Wardle
- University of Melbourne: P Vesk, L Pollock, P Baker, F Thomas
- Monash University: J Moore
- University of Auckland: C McInnes-Ng
- Macquarie University: I Wright, N Dong
- University of Tasmania: M Hovenden
- University of Technology: A Leigh

This, and other communications, led to a stubstial increase in the coverage of the dataset. The following table compares summary statistics of AusTraits at the start (v0.9.0) and end (v0.9.1) of the current project.

version	sources	contributors	species	sites	traits	records
0.9.0	139	127	19904	260	86	349,261
0.9.1	175	189	21892	3436	170	467,523

Sustainability

The ARDC project enhanced an existing project (arising from Fellowships to Falster & Gallagher). After publication of the paper and dataset, data will be publicly available and

archived in long-term repositories in an interoperable form and thus, plausibly, in little need of further support. However, our hope is to continue developing the resource over the coming years, as there will be considerable additional value realised in expanding coverage, as new data becomes available. We will therefore seek further funding to further expand the resource. In the absence of additional funding, Falster's fellowship provides some very limited resources for bare maintenance through to its completion in 2022.

Learnings

Through this project, and the broader AusTraits project, we have learnt the following major lessons about building a transformative data resource.

Value of "journal data paper" model: A common concern for researchers looking to contribute data into national or global contributions is how they will receive credit for their contributions. AusTraits provides credit through co-authorship of resulting data paper and subsequent citation of that paper. As evidenced by the enthusiastic responses to our requests for contribution, we believe this approach is satisfying for many researchers in our field. Without offer of a data paper, we believe many fewer contributions would have been received.

Harmonising diverse "small data" requires a serious workflow: AusTraits harmonises data from over 200 different sources, all using different variable names, data structures, units and sometimes methods. We developed a programmatic pipeline for standardising data without which we would have been completely overwhelmed. The raw data and code for harmonising data is all available in our github repository, https://github.com/traitecoevo/austraits.build/, which will be made open access at the time our dataset is also made open access. This exposes all the decisions made during the harmonisation process to scrutiny and potential review. Moreover, we believe the code developed can potentially be adapted by other projects seeking to harmonise tabular data.

Person time is essential: Together with the pipeline, the value of being able to employ people to process data cannot be underestimated. Each submission requires a few hours of someone's time. This is essential for those building the resource and also for contributors, as they are much more likely to contribute if they can just send us "what you have". Finding time to reformat data is a task that easily drops off the bottom of people's todo list, so anything we can do to ease this will enhance participation.

You may be defining the standards / workflows: When creating a new resource you may often be creating the standards, i.e. data structures and definitions / vocabularies / ontologies. Researchers therefore need a pragmatic approach, using tools for standardisation where available and otherwise generating them.

Impact

Once published, AusTraits will undoubtedly have major impact in our field. Examples of likely use include:

- The Atlas of Living Australia has expressed a strong interest in hosting a version of the dataset, and incorporating the trait data into their online portal
- The global dataset of plant traits TRY has already expressed interest in incorporating versions of the dataset into their global compilation
- Vegetation models are increasingly looking to predict the mixtures of plant types found across the globe. AusTraits will provide a first-class resource for developing and testing such models, especially for the Australian community land surface model.

By bringing data together on a new scale, we expect AusTraits will enable many new scientific discoveries. These will be pursued by individual researchers and research labs. As examples of projects already underway:

- Quantifying the distribution of traits across the Australian environment (led by Gallagher)
- Predicting the distribution of plant species from traits (led by Indiarto)
- Quantifying the distribution of Eucalypts in relation to climate and traits (led by Vesk)

At a broader level, AusTraits has an impact in these general areas:

- Realising the value of funds invested in previously research: Most of the data included will have been funded by public money, but has not previously been available for easy reuse
- Building community: Through the AusTraits project, we have made significant progress in forming a community within Australia around this resource, where one did not previously exist. This will raise awareness within the discipline about available resources and potentially enhance collaboration
- By releasing data under an open license, we will provide a crucial resource for outreach, enabling a curious public to engage with biodiversity in new ways

Report prepared by: Daniel Falster
Date: 20 Dec 2019