Contribution ID: **294** Type: **Poster**

CAREful Linking of FAIR Language Data to Reproducible Jupyter Notebooks

Monday 13 October 2025 19:10 (20 minutes)

The reliable reuse of language data largely depends on both managing the data in ways that respect the rights, responsibilities and communities from whom it originates, and allowing any user with the appropriate skills and resources to inspect, rerun and extend the analyses that underlie the published findings.

In practice, these goals often collide where data may be preserved in one location at an institutional or disciplinary repository, while the code that generates the tables, figures and models is scattered across personal repositories.

Jupyter Notebooks are web-based shareable documents that combine code, visualisations, rich text and interactive controls, allowing users to execute code in steps directly within the notebook, making it ideal for exploratory data analysis and interactive experimentation. Over time, however, library upgrades, version changes, missing credentials and undocumented requirements can break once-working Jupyter Notebooks, which make it harder for readers to verify and reproduce results or build on them.

BinderHubs solve this problem and further enhances reproducibility by allowing users to launch pre-configured Jupyter Notebooks as interactive computing environments from code repositories with explicitly defined hardware and software requirements.

The Language Data Commons of Australia (LDaCA) in collaboration with Australia's Academic and Research Network (AARNet) has addressed this problem by developing CAREful and FAIR data infrastructure that focuses on the preservation and access of distributed, multi-modal language data collections.

The LDaCA data portal has been configured to display the Jupyter Notebooks associated with a language data collection allowing users to automatically launch them in one of several available BinderHubs. RO-Crates, a data packaging specification, are used to describe the hardware, resources and dependencies of each Jupyter Notebook for FAIR reproducibility.

Current work involves adding additional language collections to the data portal. Some material may be sensitive and require access control, which would require a user to request access to data where appropriate. Once approved by the data custodian the user can inspect, rerun and extend the research findings using the original dataset and published analyses.

Primary author: COOKE, Steele (AARNet)

Co-authors: Mr BELL, Adam (AARNet); IP, Alex (AARNet Pty Ltd); Ms SMITH, Rosanna (UQ); MUSGRAVE, Simon

Presenter: COOKE, Steele (AARNet)

Session Classification: Poster Session

Track Classification: SciDataCon2025 Specific Themes: CAREful Indigenous Data Governance