Veracity Technology Spearhead

Enabling end-to-end veracity within value exchange ecosystems

Labelling code repositories for sovereignty checks

David Eyers University of Otago

Motivation—data sovereignty checks applied to code repositories

Discussion with Kevin Shedlock during a Veracity Spearhead meeting raised the potential usefulness of indigenous data sovereignty labels being applied to software artefacts, e.g., Git repositories.

Aims of including such labels include signalling appropriate use of the code or data within the repository, orthogonal to legal code licensing.

Popular source code management (SCM) systems such as GitHub, GitLab, Bitbucket, etc., do not yet directly support data sovereignty. Labelling repositories could be coupled with extensions to SCMs, e.g., to ensure that they adjust their functionality appropriate to the expressed sovereignty requirements, such as to ensure that repository data is only replicated to an approved subset of potential servers.

Ad hoc metadata within Git repositories

Over time, conventions have emerged that involve placing files with particular names in the top-level directory of a code repository, e.g.:

- README—Starting point to read about project; installation; use
- LICENSE / COPYING—Code license applied to repository (e.g., GPL)
- CONTRIBUTING—how to contribute, and related policies
- CODE_OF_CONDUCT—expectations of behaviour in community
- AUTHORS / CONTRIBUTORS—individuals who developed project
- CREDITS—acknowledge people & code indirectly contributing

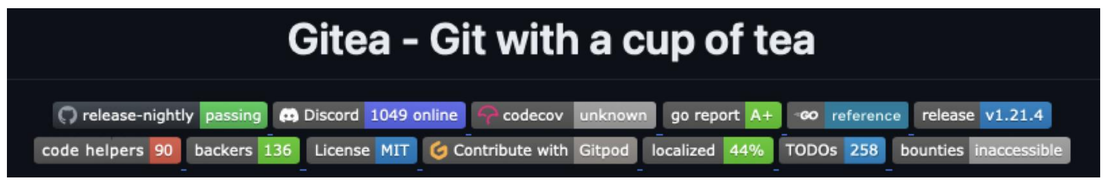
These files do not cover the cultural context we want to explain. We are keen to advocate addition of another metadata description, e.g., PROVENANCE, or CONTEXT, that can describe the cultural background.

DOI labelling

Open-source software repositories can readily acquire document object identifiers (DOIs) for specified releases. These DOIs provide a means to cite code artefacts, and link through to relevant code and data. DOIs also provide means to attached further metadata, and thus alternative or complementary means to indicate code provenance or context.

Graphical 'badges' applied to source code repositories

It is common for tools such as code testing, or security vulnerability scanners to automatically be run over the content of source code repositories when updates are made to those repositories. A popular way to display results returned from such tools is through 'badges'. These are small images that are included within a repository's documentation, that will indicate the status of the most recent evaluation of the aforementioned tools, e.g., we show here below the set of badges from the Gitea open-source SCM system:



Provenance-checking badge

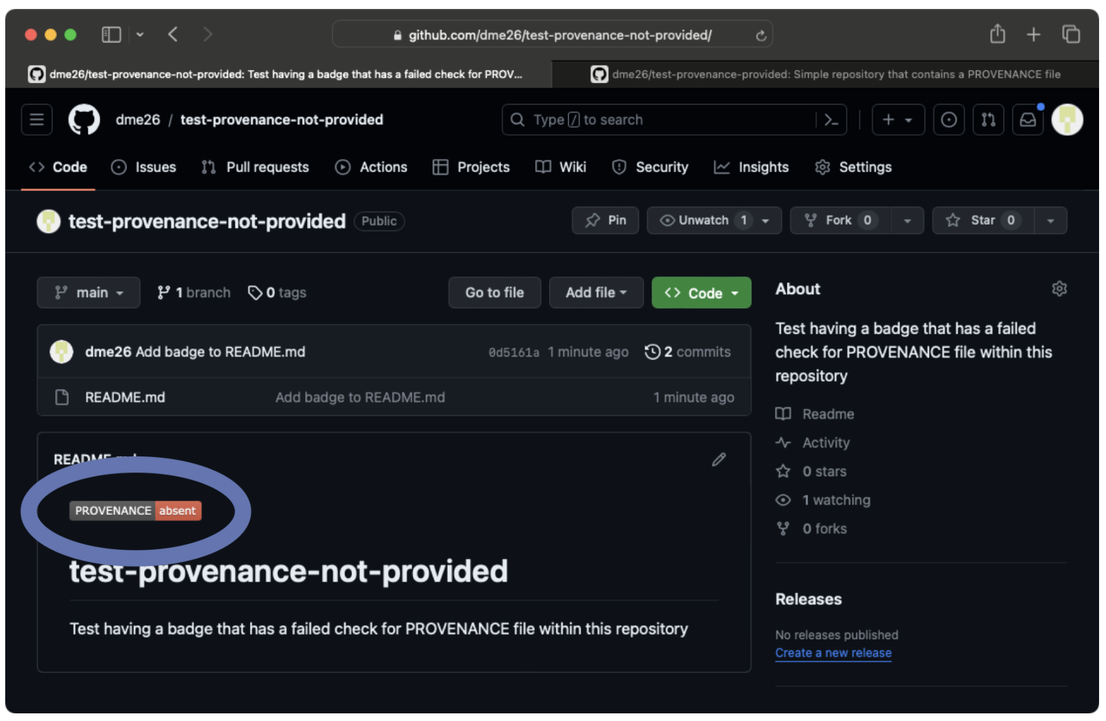
We developed a simple implementation of a badge that checks for the presence of metadata explaining the provenance of a code repository.

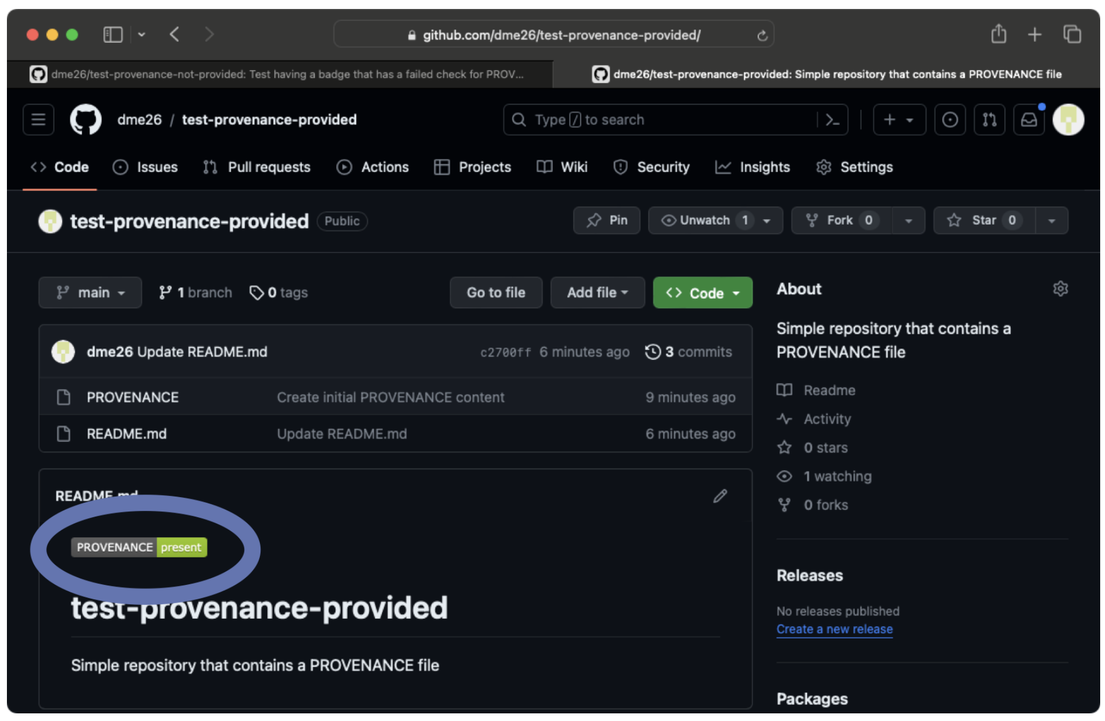
TK Labels

While we have focused on the mechanism for attaching labels, the content of the labels is crucial, and Traditional Knowledge Labels from Local Contexts would be an expressive and thorough way to describe the specific cultural context. We have had initial discussions with software developers at Local Contexts. They have not yet received a specific request for annotating code repositories, but their developers were supportive of such labelling, and exploring ways to achieve it.

Extracting provenance directly into SCMs

Some SCM systems scan for files such as LICENSE, and extract this metadata explicitly into summary information for each repository. In future we will prototype changes to an open source SCM that performs a similar function, to extract provenance metadata and act on it, e.g., automatically restricting actions consistent with data sovereignty needs.





Demonstrator: implementing a badge system that checks whether code repositories contain provenance metadata

The screen captures above show our simple prototype badge service checking for the presence of provenance information within a code repository. Ovals have been overlaid on the screen captures to highlight the badge status, which on the left pane indicates that provenance information is not present, and (b) on the right pane indicates that provenance information is present. Badges can link to web pages that provide further analysis.

The prototype service to scan a repository for provenance metadata is presently hosted on a (free) Oracle cloud instance, with plans to move it into a university hosting environment. The badge images use the functionality provided by the publicly available (and free) https://shields.io/ service.

University of Otago ŌTĀKOU WHAKAIHU WAKA

National SCIENCE Challenges — SCIENCE FOR TECHNOLOGICAL INNOVATION — Kia kotahi mai – Te Ao Pūtaiao me Te Ao Hangarau

Victoria University of Wellington
University of Auckland University of Otago
University of Canterbury University of Waikato