Using Version Control Systems to Support High-Quality Phenotype Definitions

Martin Chapman[1], Luke V. Rasmussen[2], Jennifer A. Pacheco[2], Vasa Curcin[1]
[1]King's College London, London, UK; [2]Northwestern University, Chicago, Illinois, USA

Background and Problem Statement

High-quality phenotype definitions—which are reproducible, portable and valid—are essential in providing access to accurate disease cohorts for research and, going forward, will support clinical care. It was recently identified that the *libraries* in which definitions are stored are capable of contributing to their quality; so-called *next-generation* libraries[1]. A library can contribute to the quality of its definitions by, for example, recording how they evolve over time (including the extension of existing definitions), increasing reproducibility. However, building these features into a library, or adding them to an existing one, is not a straightforward task.

Building a VCS-based Phenotype Library

A version control system (VCS) can support many of the features associated with a next-generation phenotype library (Table 1). However, to date, VCSs have mostly been used for storing and disseminating phenotype definitions—typically in individual repositories associated with a publication—with the most robust example of this being the OHDSI Phenotype Library[2].

Table 1: VCS support for phenotype libraries

VCS feature	Library feature
Versioning	View definition evolution
Branching	Access multiple definition implementations (for multiple data sources).
Cross-repository *diffs*	View definition intersection (for the same disease).
Repository visualisation	View an *abstract* representation of a definition.
Code attribution	Understand contributions and author roles.
Submodules	View parent/child definition relationships.
Continuous Integration (CI)	Determine *technical validity* using unit tests; automate gold standard comparisons.

We have updated an existing library, Phenoflow (https://kclhi.org/phenoflow), to better leverage VCS features by using GitHub as a remote VCS backend. Phenoflow uses the Common Workflow Language (CWL) to support multiple programming language implementations. Upon importing a new definition into the library, its CWL implementation is generated and stored in a GitHub repository, with subsequent updates tracking its version history.

One of Phenoflow's novel contributions is that it is built to support a canonical phenotype definition that can be connected to different data sources (e.g., OMOP, FHIR, or a local data warehouse)[3]. As such, permutations of the definitions (based on a target data source) are versioned along with it as separate *branches*. These branches retain a relationship between these versions, as they often have the same core logic. An abstract view of phenotype logic is retained by using *CWL Viewer* (https://view.commonwl.org). Definitions that target the same disease yet have entirely different logic (an increasingly common occurrence as definitions are produced for a wider variety of use cases) can also be compared and understood by computing the difference between multiple repositories.

Conclusions

We have extended the Phenoflow phenotype repository to leverage a VCS (GitHub) as its backend, demonstrating the benefits of VCSs over and above the simple versioning of phenotype definitions. In doing so, we are, in effect, building a 'GitHub for phenotypes', which is an essential step towards promoting phenotype definitions as first-class research objects which can be published, assigned DOIs, reused, and referenced, thereby improving the reproducibility and transparency of health data research.

References

[1] Chapman M, Mumtaz S, Rasmussen L, Karwath A, Gkoutos GV, Thayer D, et al. Desiderata for the development of next-generation phenotype libraries. Gigascience. 2021;10(9):1-13.

[2] Rao, Gowtham. The OHDSI Phenotype Library; 2022. https://ohdsi.github.io/PhenotypeLibrary, Last accessed on 2022-09-13.

[3] Chapman M, Rasmussen LV, Pacheco JA, Curcin V. Connecting computable phenotypes with multiple Health IT Standards using the Phenoflow library. In: AMIA Clinical Informatics Conference; 2022. .