Version Control for Data Analysis

at Sheffield City Council

Laurie Platt

08 December, 2023

Table of contents

Preface

This is a Quarto book.

To learn more about Quarto books visit https://quarto.org/docs/books.

```
1 + 1
```

```
[1] 2
```

1 Introduction

This is a book created from markdown and executable code.

See Knuth (1984) for additional discussion of literate programming.

```
1 + 1
```

```
[1] 2
```

2 Status

This documentation is at a very early stage of development.

2.1 Move to Azure DevOps

The expectation is that this book, its repo, along with most repos in this GitHub organisation, will move to an Azure DevOps within the Council's tenancy. Establishing the Council Azure DevOps for data analysis is currently on hold in order to first establish a new data platform.

2.2 Information Management

Once we move to Azure DevOps we may take advantage of the project management tools. However, for now we'll keep track of actions on this page.

Preliminary discussions with the Council's Information Management team have been held. Below are notes of (italicised) questions raised and (bulleted) actions agreed.

1. Are there retention schedules for scripts (and their versions)?
No. Current housekeeping is to periodically (maybe every 6 month) review the list of GitHub repos (repositories) and archive what's no longer in use.

- Formalise and document proposed housekeeping, including retention schedules.
- Extend periodic reviews of repos to include what should be archived, and what should be deleted from the archive.

2. When you publish your data analyses do you indicate anything about the script, its version and a plain English statement?
Each repository includes, as a minimum, a README.md file with a plain English statement. It will also often include further documentation about the data analysis pipeline. The current version and previous versions are an intrinsic aspect of GitHub and Azure DevOps.

- Include the README.md and documentation requirements in the version control housekeeping documentation.

3. When you publish to GitHub (in future) is it checked? attributed to SCC? and come with a disclaimer?

- Unit tests and code reviews are RAP (Reproducible Analytical Pipeline) practices that we're considering adopting.

- Our GitHub repos are held under the scc-pi GitHub organisation and are clearly attributable to SCC.

- Need to consider the different licenses that can be applied to published repos and whether this would cover the disclaimer.

Main action from meeting Information Management:

- Complete mini 2-page DPIA and make MW aware. It might sit under the full DPIA of the new Data Platform.

Other actions:

- Move GitHub repos to Azure DevOps.

- Draft a fit-to-publish (publically on GitHub) checklist.

- pre-commit checks (scripts) or "hooks" that screen for CSVs, PID, {secrets} (e.g. passwords) etc.
- Publish guidance and tools for annonymising datasets for data analysis to encourage working with PID as only by exception.

- Protocols in case of a data breach from incorrect use of version control.

2.3 Other ToDos

- Move version control content from pinsheff.

- Link to related unmoved pinsheff content.
- Note Version Control Show & Tell e.g. .gitignore /data.

- Add link to Show & Tell once this dpocumentation is moved to Azure DevOps.

3 Summary

In summary, this book has no content whatsoever.

```
1 + 1
```

```
[1] 2
```

References

Knuth, Donald E. 1984. “Literate Programming.” *Comput. J.* 27 (2): 97–111. https://doi.org/10.1093/comjnl/27.2.97.