Attendees:

- Ed Thomson: product manager for GitHub Actions (repository automation platform, CI, compute)
- Rimas Silkaitis: product manager for data and machine learning ops at GitHub
- Phil Durbin: developer for Dataverse
- Danny Brooke: program manager for product development at IQSS
- Gustavo Durand: tech lead for Dataverse
- Ana Trisovic: postdoc at IQSS, PhD at CERN, focus on data flow and reproducibility, especially in high energy physics. Found different reproducibility problems at U Chicago.

Agenda:

- Introductions
- Reproducibility, GitHub Actions

Resources:

- Transcript of Phil's talk at FOSDEM: https://groups.google.com/d/msg/dataverse-community/OJwTrGFPsUY/dyhSk4q3AgAJ
- Ed's blog series about GitHub actions: https://edwardthomson.com/blog/github_actions_advent_calendar.html
- Computational Reproducibility in Dataverse: https://docs.google.com/document/d/1xG8xAcPSOe1xCWUlhj46AKrK4MAZbY6ed96yBKHCXiA/edit?usp=sharing
- Paper draft on the Dataverse integrations with cloud services (CodeOcean and others): https://drive.google.com/file/d/1YFW4EHKHIury5N0Qh2zRfn4A7Y3K3GBP/view?usp=sharing

Notes:

- (Phil) Ed, can you please start us off? Something about my talk at FOSDEM resonated with you. Can you please tell us what you heard? I'll confirm or deny anything I said. :)
- (Ed) Phil described a real problem, the reproducibility crisis. We hear about this in medicine. Steps to validate the paper. Collaborative research is happening on GitHub.
- (Ed) Sounds like devops for data science. :)
- (Ed) I can see some sort of workflow that does a Docker container build and then pushes it up into Dataverse.
- (Phil) Let me walk you through the screenshots in the "Computational Reproducibility in Dataverse" doc that we recently shared with Code Ocean, Whole Tale, and Renku: https://docs.google.com/document/d/1xG8xAcPSOe1xCWUlhj46AKrK4MAZbY6ed96yBKHCXiA/edit?usp=sharing

- (Ed) I have a positive view of screenshots. "Other files" is a little weird. Seems branded as AJPS.
- (Phil) Branding is part of Dataverse. Individual collections (somewhat confusingly called "dataverses") can be branded by journals, departments, or whole institutions.
- (Rimas) Back to reproducibility. Where does the data live? How does it live there? Git LFS? Are databases supported? Big data? Does it belong in GitHub? Metapackage. Do they create a requirements file? Rake spec. Continuous integration script? Starting with Jupyter Notebooks. repo2docker. Does the notebook just run? Do you get output with no errors (CI test).
- (Phil) Maybe repo2docker is common ground. The good news is that repo2docker already supports Dataverse DOIs thanks to Kacper Kowalik from Whole Tale: https://github.com/jupyter/repo2docker/pull/739
- (Rimas) We've created a GitHub Action that creates a Dockerfile if you have a workflow file in the repo. The image could go into the GitHub Packages artifact store or Docker Hub. The Dockerfile itself would probably be committed back into GitHub.
- (Rimas) Docker is complicated for many people to work with.
- (Phil) Yes, when I attended my first Code Ocean workshop, I thought, "This is a GUI tool for creating a Dockerfile. Great!" Whole Tale does this too.
- (Ana) Code Ocean and Whole Tale allow depositing code and data. However, GitHub is used the most. There is a huge push to learn git. For economists, everybody knows about GitHub. People don't know Code Ocean or Whole Tale, for example. Setting up continuous integration would be great.
- (Rimas) Repo2Docker Action: https://github.com/machine-learning-apps/repo2docker-action : https://github.com/machine-learning-apps/repo2docker-action/pull/8
- (Phil) Let's talk more about code deposit, Zenodo-style: https://github.com/IQSS/dataverse/issues/2739
- (Phil) From the Dataverse perspective, we want robust metadata for software. We're looking at CodeMeta in https://github.com/IQSS/dataverse/issues/3736 and https://groups.google.com/d/msg/dataverse-community/nDMbMv4fKf4/P5YxHJzDBgAJ
- (Phil) Projects that are similar to Sid (Jupyter Notebooks, etc.): https://docs.google.com/spreadsheets/d/1bvNyn4wwE1eAoBJz8HyXYT3xbtDWhDZBixg5wgRnA2U/edit?usp=sharing
- (Rimas) Next steps?
- (Ana) Dataverse is a good place to preserve data and code. We don't know what the future of GitHub is. Dataverse has a commitment toward long term preservation.
- (Phil) I would *love* a GitHub Actions template for "deposit to Dataverse". :)
- (Ana) Immediate benefits in economics, etc. They've put work into CI. Can we automate getting the data from GitHub into Dataverse?
- (Phil) Can we nudge people toward adding code and workflows (CI) into GitHub for eventual deposit into Dataverse? If there's only a README in the GitHub repo, that's not a great dataset. Add some data. Better. Then add some code, if you can. Great. You're a

researcher who knows what CI? Great! Add a workflow using GitHub Actions. Now we have a Dockerfile. Your dataset is looking really good now. :)

- (Rimas) Are people willing to adopt and learn new tools, including command line tools?
- (Phil) There's a whole variety of researchers. Some are early adopters and love the command line. Most would prefer a tool in the browser.
- (Phil) Since you two are ok with my SLOPI communication style, I'll link to these notes from the "Computational Reproducibility in Dataverse" doc above and the two open issues about code deposit: https://github.com/IQSS/dataverse/issues/2739 and https://github.com/IQSS/dataverse/issues/5372