Management and Preservation of Analysis Notes: Issues and Options

Maxim Potekhin

Nuclear and Particle Physics Software Group

BROOKHAVEN NATIONAL LABORATORY

PHENIX EC Meeting

03/11/2021

PHENIX

Analysis notes: requirements

- Privacy/access control
- Version Control
- Search/discoverability

- Historically, all these functionalities were provided by the internal web server
  - Things are not the same after disruptions in 2020 due to evolving infrastructure and security requirements and departures of key personnel

The challenge

- The PHENIX DB and parts of the legacy web infrastructure are in crisis
  - Currently non-existent support
  - Direct impact on the remaining web services
  - Legacy web application code too expensive to maintain “ad hoc” (PHP, Postgres)
  - Ideally needs to be supported by SDCC
- Example of a solution: ongoing successful migration of many types of PHENIX materials to sustainable modern platforms e.g. Zenodo and GitHub
  - PHENIX Theses papers (done)
  - Conference presentations (good progress)
- The analysis notes case is different and more challenging
  - Large number
  - Privacy and protection, access managements

Analysis notes management going forward: some of the previously considered options

- BNLbox
  - BNL-based, conceptually similar to Dropbox, easy to use, Web UI and some CLI capability
  - No support for queries/keywords
  - Access can be restricted to a defined group of people
  - No issue tracking (see comments on GitHub in the next slides)
- Zenodo
  - A flagship CERN product, cross-disciplinary digital repository, DOI support
  - **Excellent** query mechanisms including *elastic search*
  - It is possible to make a particular entry “private” and grant access on request
    - This results in a matrix of N people vs M documents, so access control is a bit cumbersome
  - A viable long-term platform, already actively used by PHENIX for all sorts of materials
  - Transparent storage of folder hierarchies is impossible (however can store a ZIP file)
  - No issue tracking

Analysis notes management options: the shortlist

- Both contenders are Version Control Systems and based on *Git*
- GitHub (cloud) vs Gitea (BNL-hosted)
- GitHub
  - Considered because it's ubiquitous and industry-standard, has the right set of features
- Gitea
  - Because of expected commitment on the part of BNL for long-term support, features similar to GitHub

Analysis notes options: GitHub

- A leading cloud platform with substantial industry adoption and support
- Used by PHENIX since ~2yrs ago
- Can establish a private repository to host analysis notes
- Accessible to users on a managed list
- GitHub tags can be used for indexing (like keywords)
- Concerns about longevity of this resource (but should we be concerned?)
  - Will it be around in 10 years? My guess would be yes.
  - If it does go out of business we'll be in the company of world leading enterprises migrating to the next platform so all sorts of tools and options are guaranteed to appear
  - NB. Data won't be lost in any case - we'll have local repo clones, long term backup solutions such as HPSS and other BNL storage etc

"Today 52 percent of Fortune 50 companies use GitHub's Enterprise business tier, which costs $21 per user per month. Altogether, GitHub has more 23.1 million users in 200 countries and 1.5 million organizations."

Analysis notes options: Gitea at BNL

- Based on the same underlying repo technology (Git)
- A different product and not a cloud service like GitHub
- Another layer of access control (controlled by BNL)
  - This can be good or bad
- No free hosting of web site(s) which we enjoy with our updated PHENIX site - perhaps unimportant for the analysis note use case
- Some other differences
- We can *hope* for long-term commitment of our organization to this service

GitHub and Gitea commonalities

- Version control of analysis notes far superior to what we have now (~none)
- Natural support of directory trees
  - For example it is easy to handle folders with temporary/preliminary plots
  - Add any notes, materials, code snippets
  - ...the latter would actually be a step forward for quality of the information we store
- The "release" mechanism - having a reliable reference to a particular state of information for future reference
  - Making a note final is simple - done by cutting a release
- Tags - search capability - can maintain a curated set of k/w as we do with Zenodo
- The "issues" feature
  - Create and track comments, suggestions, bug reports and progress towards resolution
  - Can completely replace the email back-and-forth, +notification about repo activity
  - Might be ideal for developing analyses and analysis notes - we've never done this before but it is very likely to work

Preliminary plots

- Another example of a data product that ideally needs to be migrated to a more viable platform
- Many of the same characteristics as the analysis notes - access control, perhaps versioning

Decision

- My 0.5 cents: leaning towards GitHub
- Gitea is a close second as it is indeed likely that BNL will be willing to support this service for a long time
- However there aren’t convincing arguments about GitHub becoming unavailable on a O(10) year timescale
- Whatever decision is made, we’ll need a brief testing/validation period
- Need clarity on what to do with the notes committed to the legacy system