Recent experience with software documentation for RHIC and EIC

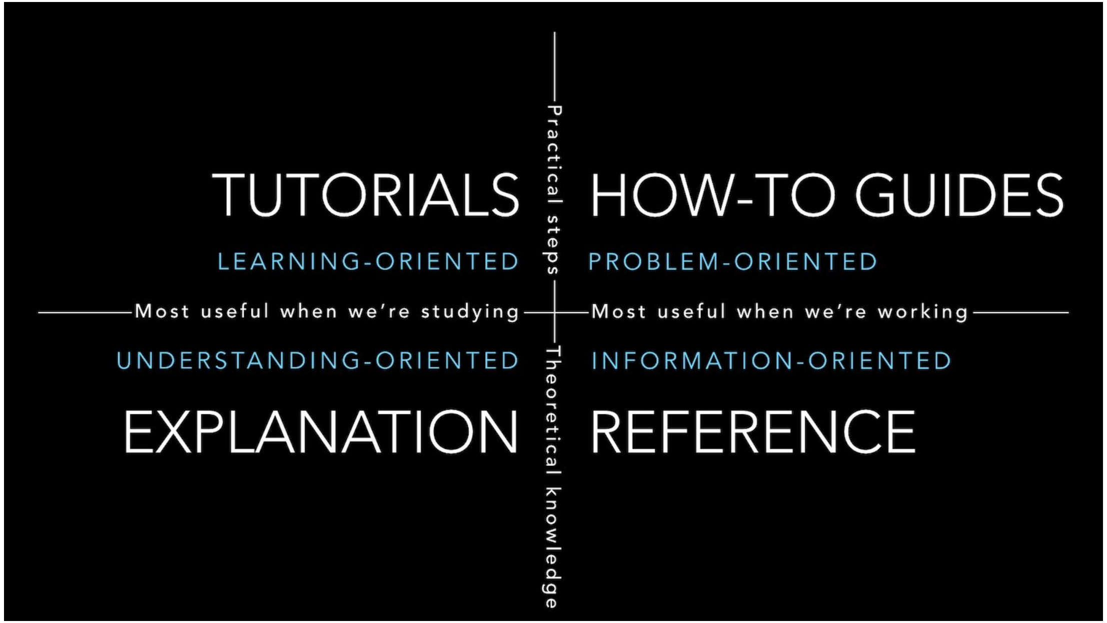
M.Potekhin (BNL NPPS)

EIC CompSW meeting
07/20/2022

The goal

- To see what can be learned from our recent experience with documentation
- To ensure efficiency of the CompSW operations…
- …and ultimately make our lives easier ☺

“The quadrant diagram” from the decision doc – very useful…



Practical steps

TUTORIALS

LEARNING-ORIENTED

HOW-TO GUIDES

PROBLEM-ORIENTED

Most useful when we’re studying

Most useful when we’re working

UNDERSTANDING-ORIENTED

EXPLANATION

INFORMATION-ORIENTED

REFERENCE

Theoretical knowledge

The quadrants projected into the platform space

- There is no definitive mapping of any of the quadrants to some set of optimal tools
- The medium is less important than the design of the respective documents (in a pinch a text file will do the job)
- However some platforms are more optimal than others for specific applications, and that needs to be discussed

One version of the potential map – comments welcome…

Tutorials

Web*

Sphinx**

YouTube

How-to

Web

Sphinx

OpenData

REANA

Jupyter**

READMEs

Explanation

Web

Zenodo

YouTube

Reference

Web

Sphinx

Doxygen**

Zenodo***

*Web = Static sites, Wiki etc

**Not in the scope of this presentation

***Or an equivalent repo

Web platforms – team of editors

- Regardless of the chosen platform, it is important to have a well defined team of editors with sufficient control over the content – without that, documentation is all but guaranteed to eventually become unusable as evidenced by many Wikis etc
- This also implies that “open to all” editing access such as often used in Wikis is problematic – it is conducive to quickly filling the site with initially useful content, however lack of defined responsibility and coordination, people moving on etc will eventually become detrimental to the quality of the web materials. Clutter and bloat.

Web platforms – technologies – a note on Drupal

- Recently, Drupal has had a history of disruptive version upgrades which start from PHP (and thus are made mandatory for security reasons) and then percolate into versions of Drupal itself, and further into Drupal modules, creating a compatibility tensor
- Disruptions are very real, from EICUG experience on BlueHost
- Some migrations require considerable expertise, and it's not clear whether a Drupal professional will be available to EIC at any given time; if customized modules are used, this is becomes a liability
- For these reasons the EICUG has recently migrated – with success – from Drupal to a static web page generation process (Jekyll) – http://www.eicug.org/

Web platforms – experience with static websites (Jekyll)

- The experience is quite positive
- A few sites have been developed recently for the RHIC, EIC and other communities, including the EICUG Software Group site: https://eic.github.io/
- Speed and security (good chances of approval from Cybersecurity)
- Versatility of deployment options: not only GitHub pages, but straightforward Apache and nginx deployment – can build the site on any machine and quickly place it on any server (simple copy)
- Password-based protection of pages is possible (utilizing the server capability) and has been tested with nginx

Web platforms – Markdown/Jekyll/Liquid features

- Markdown is an intuitive/simple way to create and maintain content, provides welcome compatibility with the various README.md
- If one advantage of Jekyll were to be listed here, it would be transparent storage of structured data in YAML and/or CSV format
  - Human-readable, version controlled
  - Is used to create data structures with functionality approaching that of a database e.g. select, join, filter etc
- Can use pretty much any web page template, with Bootstrap being one popular option; a lot of flexibility in defining layout and navigation; CSS/styling is very approachable
- A feature-rich macro language (Liquid) facilitates site building

Jekyll development/maintenance workflow

- Just a normal GitHub workflow with branches or forks, and PRs, with the “GitHub issues” mechanism used to track development
- Works well with the idea of designated “editors” who in this case are the managers of the GitHub repo; access can be controlled, without making contributions from the collaborators and teams difficult; validated in the recent commissioning of the EICUG website
- As of last year VS Code is integrated into GitHub so it is possible to do productive work even without a proper workstation – the web functionality of VS Code is identical to the desktop, and the site is rendered within minutes of the git commit
- Zero (or close to zero) routine maintenance is required

Jekyll – hosting documents

- Pretty easy to do, hosting from the “static” area of the web server, hence speed
- Can use keywords to aggregate relevant links and other data on the pages of the website
- However, it doesn’t scale well due to sheer volume (has been tried in PHENIX) – including various limits set by GitHub
- Hence, Zenodo…

Zenodo in PHENIX

- Zenodo has become the cornerstone of Data and Analysis Preservation in PHENIX
- PHENIX is using the CERN instance: great support, generous storage, ease of access – and there is currently a test instance at BNL
- Prior to migration, PHENIX was using a homegrown PHP/DB solution which had limited discoverability and didn't age well, with substantial maintenance effort required at times
- Close to 600 documents have been migrated or committed to Zenodo, indexed with keywords maintained on the PHENIX website – in addition to the efficient elasticsearch

Zenodo – salient features

- Excellent search (incl. elasticsearch), keywords, visibility
- Works for any material (including data and software, images, video)
- Has a few privacy options
- Features persistent unique identifiers, by association with the DOI foundation: https://www.doi.org/
- Having DOI identifiers is instrumental for long-term durability of the resource since it effectively enforces storage platform migration policies and technologies that guarantee that an item committed to a repo remains accessible by the same ID/URL, indefinitely (in theory)
- GitHub integration (DOIs minted for software releases, citable)

Jekyll + Zenodo

- In practice, this turns out to be a great combination
- Document storage is offloaded to Zenodo, so the site is not bloated, and discovery of materials is facilitated
- By keeping a managed list of keywords on the static site (Jekyll), it’s easy to create customized views into the Zenodo repository (e.g. “dirc+pid”, “emcal+geometry+gdml”, “ml+pid”)
- An excellent way to store and retrieve conference materials

A note on the static website functionality

- The site should only contain materials that cannot be optimally stored elsewhere – and there is still plenty of such material
- Consequently, a large part of the site functionality is being an aggregator and a hub providing access to various other platforms
  - GitHub
  - Zenodo
  - OpenData
  - YouTube
  - Doxygen
  - …

OpenData – the ultimate how-to platform

- OpenData is a portal hosted at CERN
- A point of synthesis for documentation, data and software, all accessible from the same page
- Facilitated by use of containers
- Complete use cases/walkthroughs can be created and hosted in an accessible manner
- Extensively used by the LHC experiments
- Created for LHC, but hosting for EIC can be negotiated – NB. this was successful for PHENIX which is now a member of OpenData
- Downside – publication process might take some time

REANA – another how-to instrument

- REANA provides a way to manage multiple elements of a specific work item as a coherent, self-contained package:
  - The software environment (in terms of a Docker image and environment variables, as well as subsidiary binaries)
  - Samples of input data
  - A well defined description of the actual workflow (YAML)
- With proper packaging, it's easy to provide consistency
- Jupyter: notebooks can be run over REANA workspaces, providing an accessible way for inspection and experimentation, enhancing the how-to aspect

A note on containers

- In the present context, containers are an extremely helpful tool, for example facilitating creation of tutorials
- Meshes well with Data and Analysis Preservation, including REANA
- Enhances the OpenData functionality

Summary

- Static website generation appears to work well and meets the goals of low maintenance and long-term durability; it is already used by the EIC community, following positive experience in HEP and NP
- OpenData is an intriguing and full-fledged platform for hosting the "how-to" type of materials
- A modern digital repository such is Zenodo is of essence and will play a crucial role across various classes of software-related documentation, with DOI capability being an important factor
- It is advisable to not develop a document handling system in-house, as it will create a huge technology debt