PHENIX Experience with Zenodo and the EIC perspective

Maxim Potekhin
BNL
Nuclear and Particle Physics Software Group
04/30/2020

Overview

- Will keep this short (see backup slides for some detail)
- We observe synergy and common requirements in a few projects/experiments
  - I’m involved in PHENIX and EIC so have some perspective in these areas
  - Previous experience in DUNE
- Will briefly consider history, current situation and outlook
- EIC, PHENIX and other projects need to define action items for short and medium term
  - Not in a good position to wait for a solution for much longer

2019

- PHENIX starts its Data and Analysis Preservation (DAP) effort
  - Taking stock - fragmented and hard to navigate websites, overgrown Wikis, obsolete documentation in software and other areas, loss of records of know-how necessary for analysis, diminishing manpower
  - Custom apps (PHP) used to keep notes, papers and other publications
  - **Knowledge management** recognized as the key work area for both medium and long-term goals of PHENIX - maintaining quality of existing analyses as well as fulfilling the DAP mandate by the Lab and the funding agency
  - Participation in the DAP workshop at CERN in Fall'19
- BNL EIC group
  - White papers and other such materials uploaded to the Wiki, Drupal and even Indico
  - An instance of DocDB, older version, not much traffic
  - The need for a document management system recognized but not considered as pressing
  - Change of landscape with the start of the Yellow Report process

2020

- PHENIX
  - Development of a permanent DAP Website with community involvement, started with run configuration, detector information etc, software tutorials and intros are in the works
    - Not a document management system
  - Explored Invenio “a la sPHENIX” (a custom app), held meetings - TBD
  - Converged on the decision to use the Zenodo instance at CERN, created a “PHENIX Community” on Zenodo, started with a low-hanging fruit - centralized storage and tagging of the PhD theses produced by the Collaboration (totaling 194)
  - Decision to make the analysis software and analysis notes public deferred for later
- EIC
  - An EIC GitHub organization established with a few dozen repositories and growing
  - An EIC Software Group Website under construction (using GitHub pages)
  - General use of collaborative tools has greatly expanded
  - Yellow Report as a top-priority deliverable with more complex documents on the horizon
  - The need to generate and systematize complex information has increased

EIC vs PHENIX

- While PHENIX finished taking data a while ago and the EIC is still pretty far in future, they both have deadlines that drive document management requirements
  - Yellow Report and CDR for EIC
  - PHENIX working at reduced personnel levels and needing to preserve know-how both in near term (analysis) and long term (DAP) - this is not optional

Platforms

- While helpful, the Wikis, Drupal, Dropbox etc not the right solutions
- DocDB has been popular in the past, heavily used in DUNE and still provides a baseline for evaluating features in other solutions
- In zeroth approximation, a document management application is a union of
  - Storage
  - Metadata
  - User Interface
- In that approximation, Zenodo is a drop-in replacement for DocDB
  - Plus many additional features and API
  - Minus agenda management which is no longer used anyway since taken over by Indico
- Policies: public vs restricted vs private

First experience with Zenodo

- Meets the definition of a good system:
  - Simple things are “easy”
  - Complex things are “possible”
- Initial learning curve is quick and painless, the keyword function transparent
- Versioning
- ORCID and DOI capability off the bat
- Tiers of access
  - Private (locked in)
  - Restricted (by individual request)
  - Public
- “Communities”
- GitHub integration
- Search capabilities (many!)

DOI, versions, keywords, conference-awareness

EXAMPLE

March 2, 2020

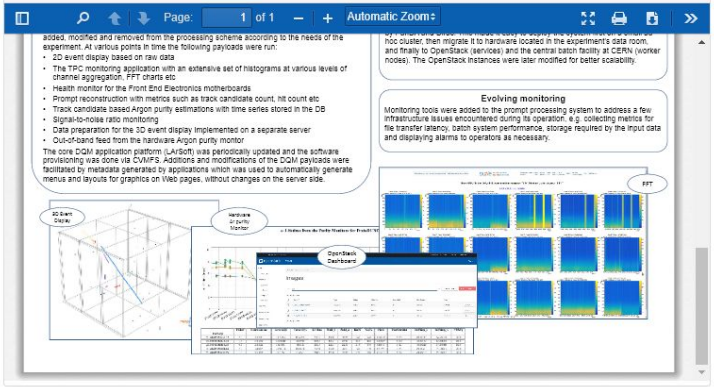
Poster | Open Access

Evolution of the Data Quality Monitoring and Prompt Processing System in the protoDUNE-SP experiment

Maxim Potekhin

The DUNE Collaboration has successfully implemented and currently operates an experimental program based at CERN which includes a beam test and an extended cosmic ray run of two large-scale prototypes of the DUNE Far Detector. The volume of data already collected by the protoDUNE-SP (the single-phase Liquid Argon TPC prototype) amounts to approximately 3PB and the sustained rate of data sent to mass storage is of the order of O(100) MB/s. In addition to this data being committed to mass storage and processed in the Grid environment a small fraction of it is captured by the Prompt Processing System which is optimized for continuous low-latency calculation of the vital detector metrics and parameters as well as the output rendered as event display images. This system is the platform for Data Quality Monitoring in protoDUNE-SP and has served a crucial role starting from the commissioning of the apparatus and throughout its operation in 2018-2019, which continues at the time of writing. We present our experience in operating the system in the CERN environment, as well as work currently underway to make the system more scalable, resilient and to simplify system recovery procedures in preparation for the second run of protoDUNE-SP foreseen after the Long Shutdown of the LHC in the Fall of 2019.

Preview



Edit

New version

18 views

11 downloads

See more details...

Indexed in OpenAIRE

Publication date:
March 2, 2020

DOI:
DOI 10.5281/zenodo.3693788

Keyword(s):
dqm | prompt processing | data quality monitoring | DUNE | neutrino | Liquid Argon

Meeting:
24th International Conference on Computing in High Energy and Nuclear Physics (CHEP 2019), Adelaide, Australia, 4-8 November 2019

License (files):
Creative Commons Attribution 4.0 International

Versions

Version 1 — Mar 2, 2020
10.5281/zenodo.3693788

Cite all versions? You can cite all versions by using the DOI 10.5281/zenodo.3693787. This DOI represents all versions, and will always resolve to the latest one. Read more.

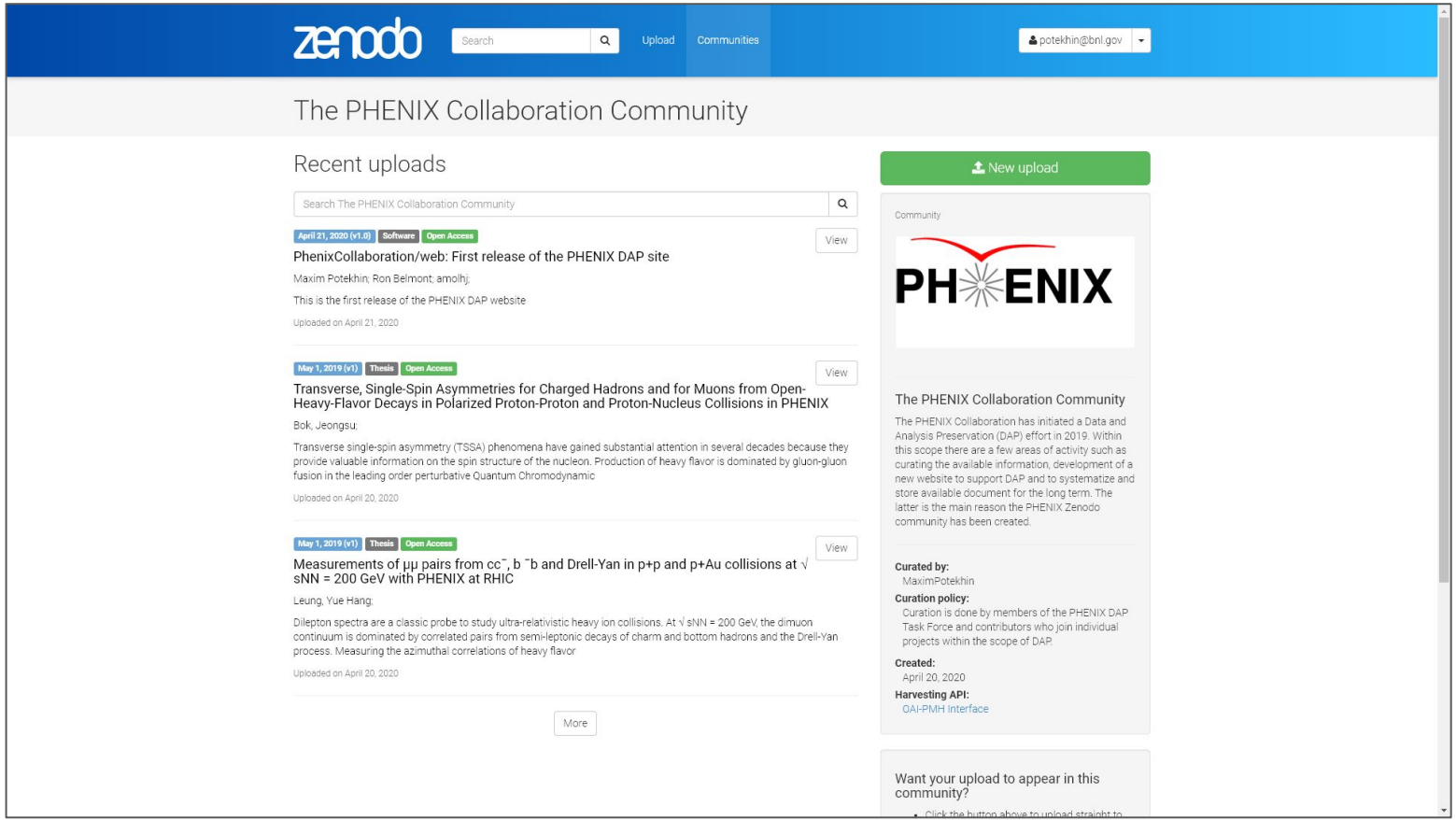
Recent Zenodo activities in PHENIX

- Approved for PHENIX Data and Analysis Preservation (EC, conveners)
  - Location of the host not considered an issue
- “Zenodo Communities” - see next slide - functional testing started
  - A “PHENIX Collaboration” community created, **started populating it with materials**
  - DAP site links to Zenodo
- Communication with the developers, looking for guidance regarding
  - Possible future data migration from Zenodo to Invenio RDM
  - Feature requests for community management
  - Storage allocation and use pattern discussion
- GitHub integration - “nice to have” but not core - initial testing done
  - Additional cloud replica of your GitHub release tagged with arbitrary metadata (discoverability)
  - Citeable via DOI

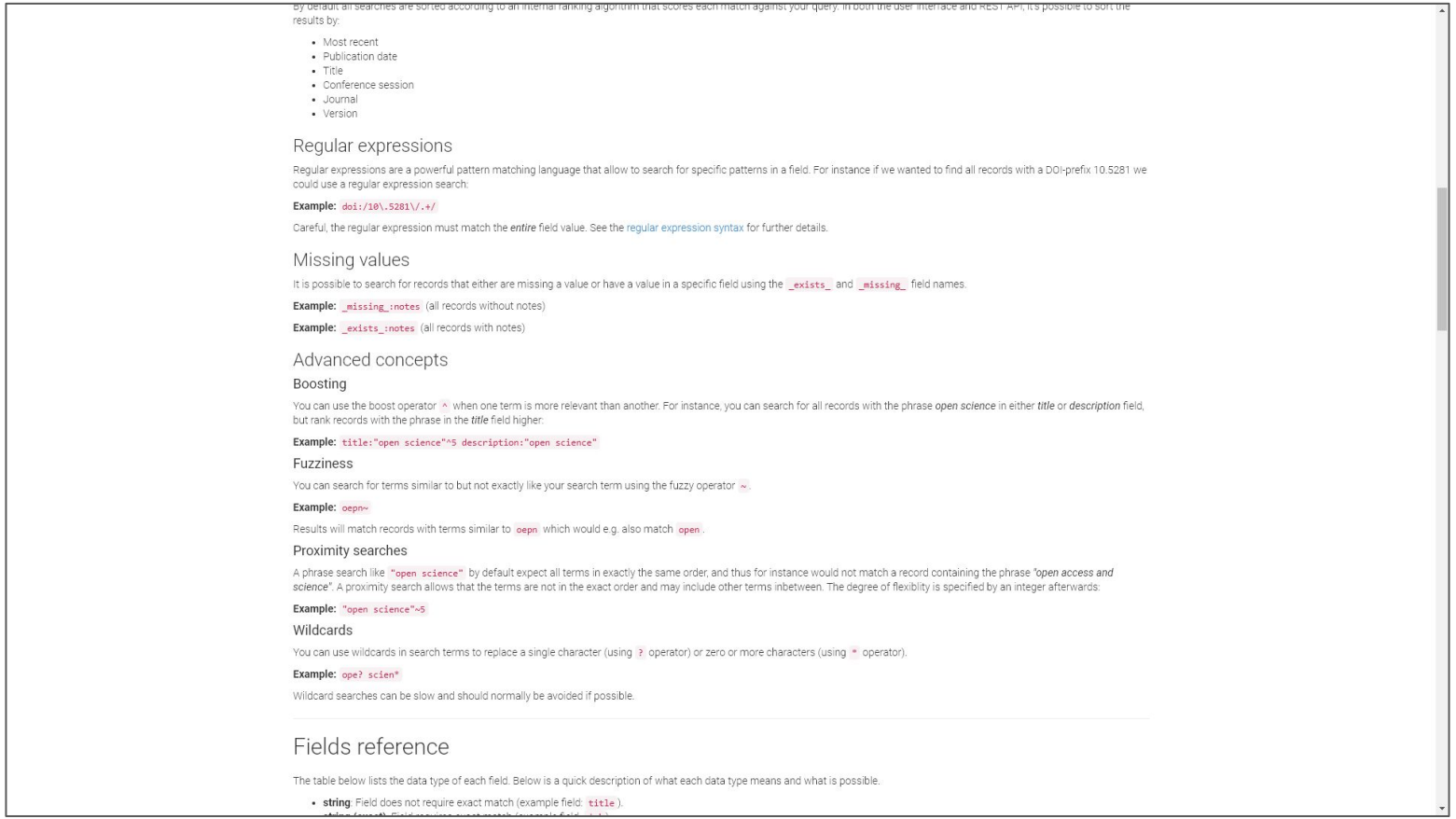
Zenodo Community (another way to tag material)

- A way to organize material, and to consistently attribute materials to a collaboration/project/experiment - keeping a consistent brand
  - No need for multiple Zenodo instances?
- An improvement in visibility/discoverability/PR
  - An addition to the already existing metadata query aids in discovery of materials
- Anyone can upload a material to the community which is subject to **curation**
  - The curator gets notified and inspects the submission
    - If accepted, it becomes posted under the community umbrella
    - If rejected, it still remains on Zenodo site but is not officially owned/acknowledged by the community, this is an accordance to the “open access” platform
  - *There is currently one curator per community and there is no easy way to transfer this duty to a different account* (something few people expected) but a fix is on the way according to the lead developer and other team members. Unofficial ETA is late 2020.

PHENIX Community on Zenodo



Advanced search capabilities



By default all searches are sorted according to an internal ranking algorithm that scores each match against your query. In both the user interface and REST API, it's possible to sort the results by:

- Most recent
- Publication date
- Title
- Conference session
- Journal
- Version

Regular expressions

Regular expressions are a powerful pattern matching language that allow to search for specific patterns in a field. For instance if we wanted to find all records with a DOI-prefix 10.5281 we could use a regular expression search:

Example: `doi:/10\.5281\/.+/`

Careful, the regular expression must match the ***entire*** field value. See the regular expression syntax for further details.

Missing values

It is possible to search for records that either are missing a value or have a value in a specific field using the `_exists_` and `_missing_` field names.

Example: `_missing_:notes` (all records without notes)

Example: `_exists_:notes` (all records with notes)

Advanced concepts

Boosting

You can use the boost operator `^` when one term is more relevant than another. For instance, you can search for all records with the phrase *open science* in either *title* or *description* field, but rank records with the phrase in the *title* field higher:

Example: `title:"open science"^5 description:"open science"`

Fuzziness

You can search for terms similar to but not exactly like your search term using the fuzzy operator `~`.

Example: `oepn~`

Results will match records with terms similar to `oepn` which would e.g. also match `open`.

Proximity searches

A phrase search like `"open science"` by default expect all terms in exactly the same order, and thus for instance would not match a record containing the phrase *"open access and science"*. A proximity search allows that the terms are not in the exact order and may include other terms inbetween. The degree of flexiblity is specified by an integer afterwards:

Example: `"open science"~5`

Wildcards

You can use wildcards in search terms to replace a single character (using `?` operator) or zero or more characters (using `*` operator).

Example: `ope? scien*`

Wildcard searches can be slow and should normally be avoided if possible.

Fields reference

The table below lists the data type of each field. Below is a quick description of what each data type means and what is possible.

- **string**: Field does not require exact match (example field: `title`).

Policy issues

- Zenodo defines itself as an open science platform i.e. for the most part public
- It does have access tiers: private, restricted and public
  - “Restricted” means that a request for access is forwarded to the owner
  - Not designed to handle “roles” for large groups of people
- Consider the fact that DocDB instances are often protected
  - In reality I would say 95%+ of materials don’t need to be protected
  - However in some cases (preliminary analysis etc) this may be important

GitHub/Zenodo mechanics (see backup slides)

- A snapshot of a GitHub repo can be included in Zenodo organically+DOI
  - Integration/app link is in place: prepares and preserves tarballs of your releases
  - Makes your code easy to find (using the metadata) and to reference by a unique ID
  - Nice GUI
  - DOI reference to the code - becomes citeable
- Easy to use
  - Well-developed interface, I tested this functionality and it was quite simple
  - DOIs take some time O(10min) to propagate to the DOI.org system, but this is not a problem

Summary

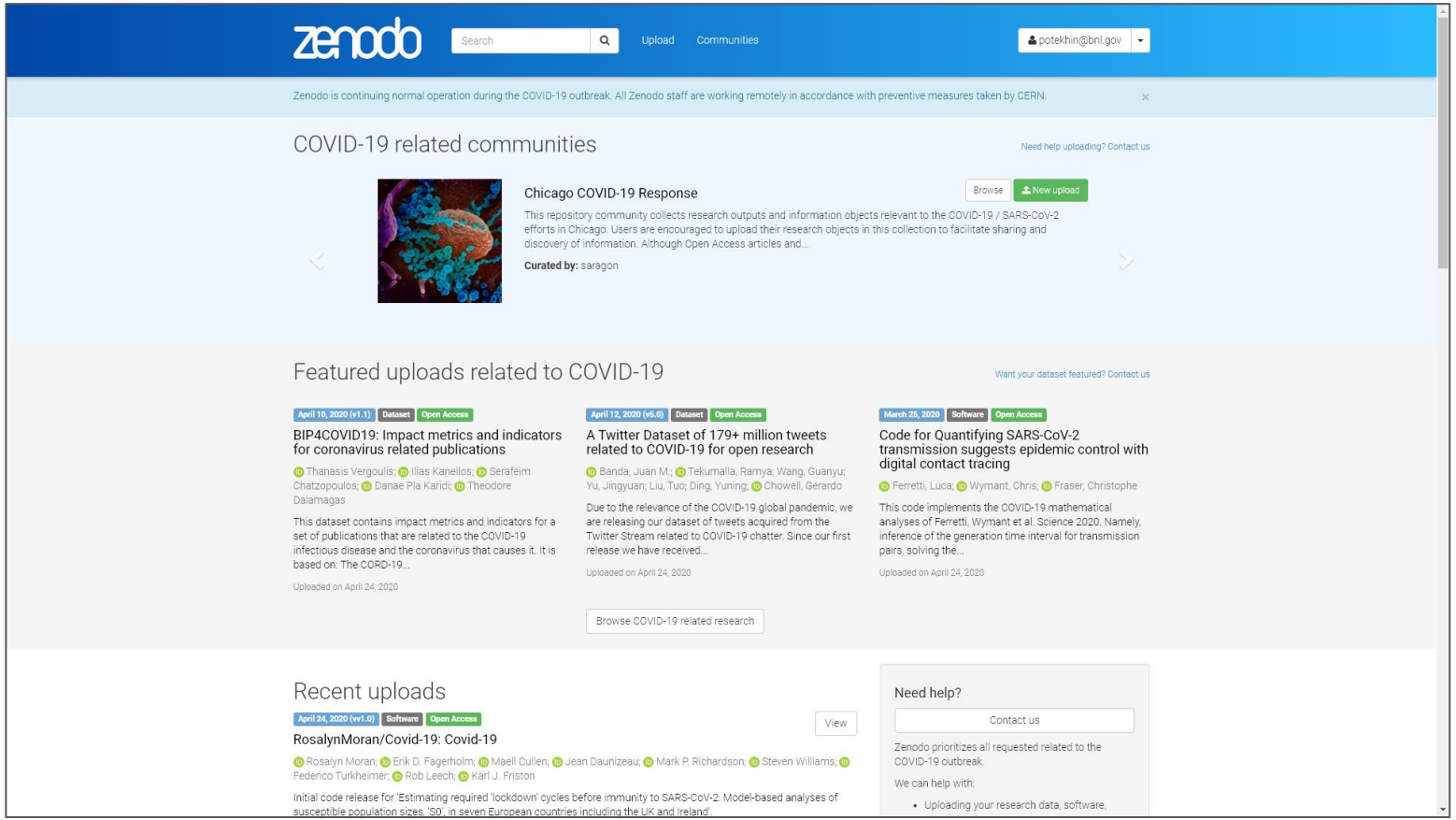
- Convergence of requirements in EIC, PHENIX (and potentially sPHENIX)
- Interest in the EIC Software Group (including GitHub integration)
- This is urgent (cf. PHENIX started using the CERN instance but it’s not too late to reconsider, and EIC needs this ASAP)
- Is Zenodo at BNL production-grade? Does it have full functionality?
- When can PHENIX migrate from CERN to BNL?
- Will there be multiple instances (and if yes, why)?
- Does it have same open policy wrt account creation and access?
- How much storage can we count on immediately, and in the long term?
- Quotas?
- What is the migration plan to Invenio RDM and do we need one?


BACKUP SLIDES

Terminology

- **Zenodo** is an open science data repository at CERN
  - In a nutshell, storage+metadata
  - Any data within the set limits
- **Invenio** is a toolkit used to in a number of CERN systems *including* Zenodo
  - A complex and capable framework.
  - Framework, not a system. *An application is needed to make use of its functionality*.
  - *cf. Zenodo is an Invenio-based application.*
- **Invenio RDM** (“research data management”) is a new product aiming to achieve
  - Portability (currently installing and configuring Invenio requires a high level of expertise)
  - Configurability i.e. eliminating the need for a custom app - a turnkey solution
  - ETA: late 2020

https://zenodo.org/ - named after Ζηνόδοτος, inventor of metadata in 280 BC



Zenodo “in a nutshell”

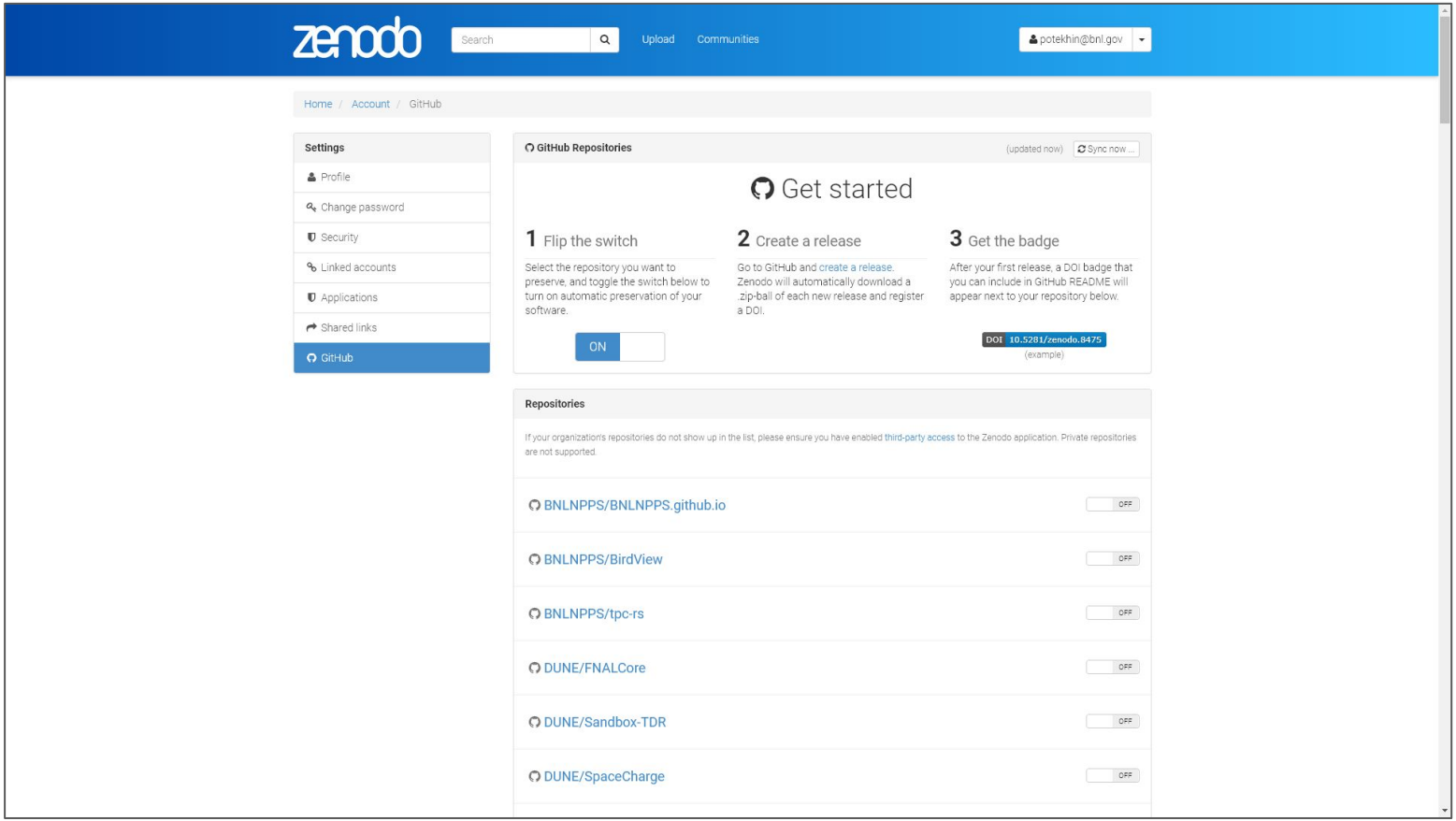
- General purpose digital repository
- Version control
- Data (storage space) + Metadata (DB)
- Extensive query capabilities
  - Full-text search is in the works
- DOI management (**doi.org** integration)
- ORCID-aware
- Gateway to other repositories
- GitHub integration (citeable code)
- Currently a service instance at CERN, being transformed into a more portable system under the “Invenio RDM” brand

Zenodo: durability

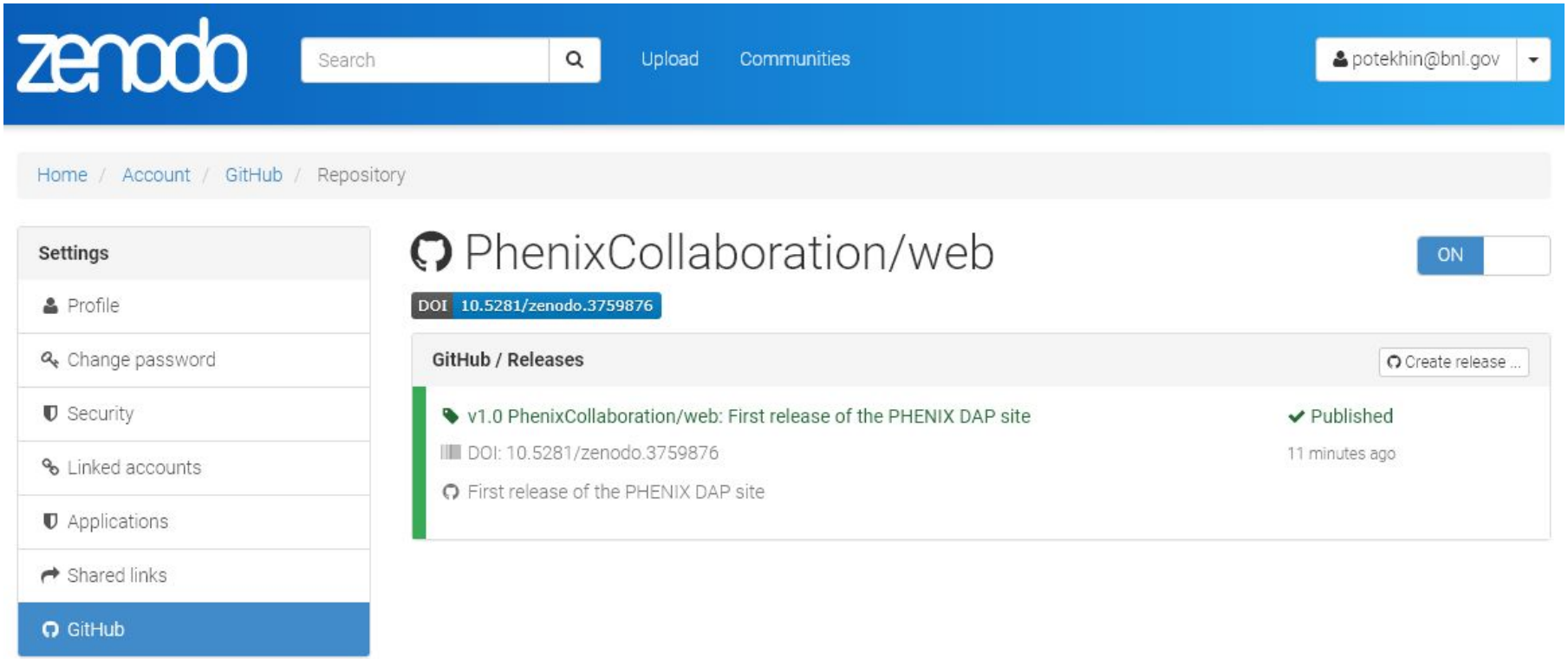
Motivations

- Managing documents and other materials is a universal necessity in the field
  - Consider the needs of the Yellow Report working groups (papers, presentations, tables etc)
  - Not a replacement of the Wiki (which is not a document handling system in the first place)
- Not too many products exist in that area
  - DocDB is used at FNAL, BNL and a few other places, it’s an aging product, no clear API
  - CERN CDS is not portable (NB shares the Invenio back-end with Zenodo)
- In EICUG here is currently not a single accepted solution or a policy
- The new EIC Software website is not designed as a general purpose document store (scalability, lack of proper metadata etc)
- Zenodo is an obvious contender

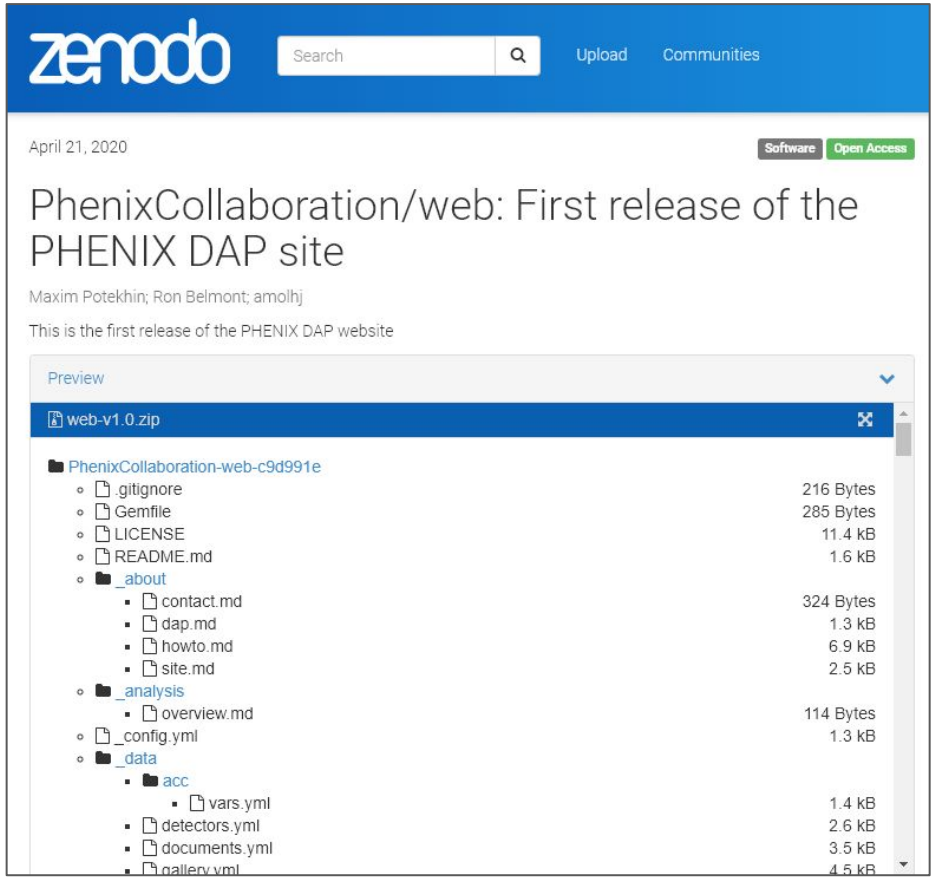
Zenodo - GitHub panel - repo selection



Zenodo - GitHub panel - published release

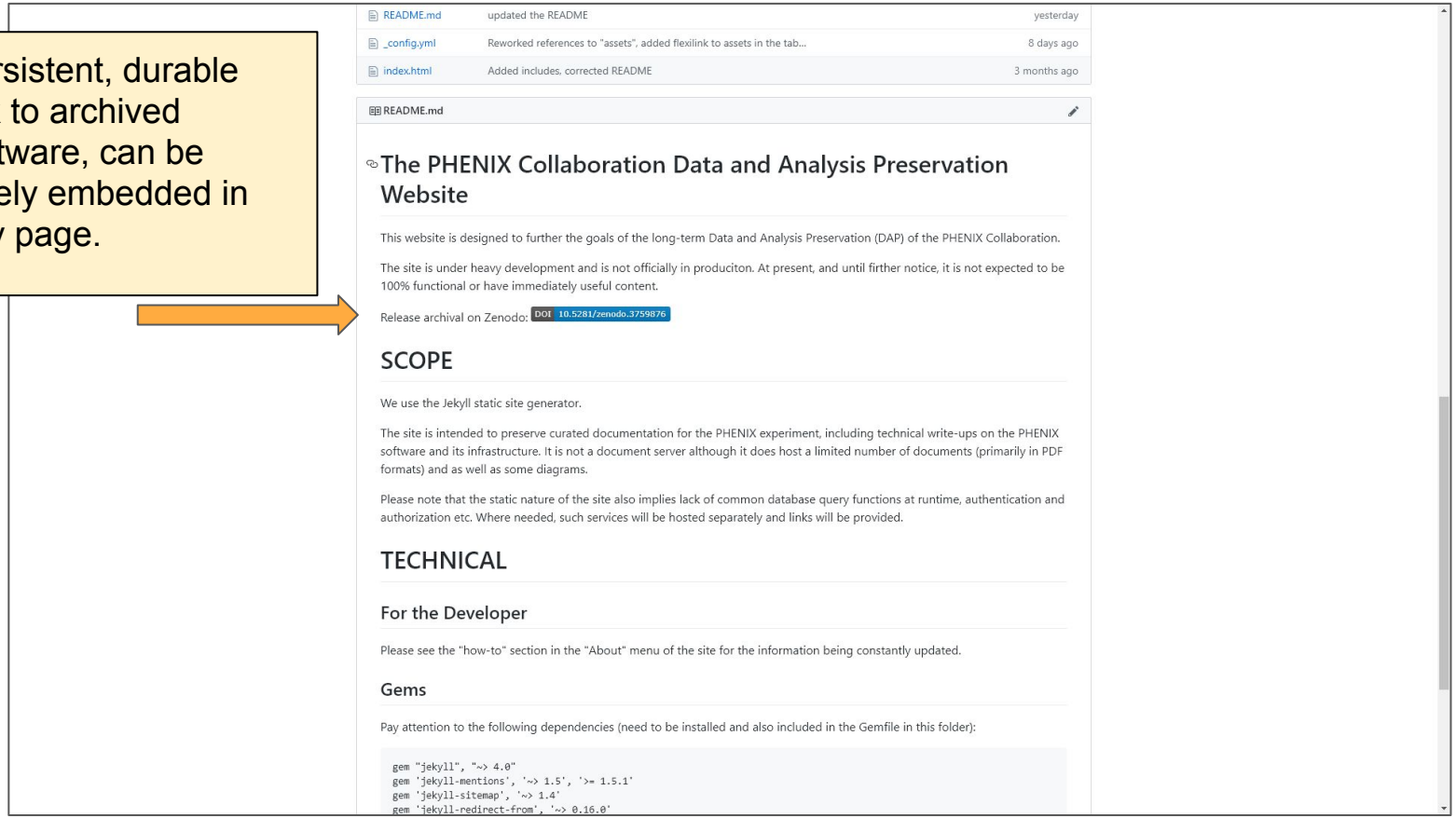


Zenodo - GitHub panel - published release browser



DOIs are an increasingly popular way to reference software

Persistent, durable link to archived software, can be nicely embedded in any page.



GitHub/Zenodo integration benefits

- Not a core functionality by a long shot, however...
- ...provides a uniform way to reference digital products using DOI
- ...metadata is a good thing to have - better discoverability!
- ...can leverage the Zenodo “community” feature to organize materials and increase visibility
  - Cf. simulated data and the code used to produce it can be kept under the same umbrella
- Longer term - Data and Analysis Preservation
- In general, an “EIC Software” community on Zenodo may be a useful thing to have (papers, conference presentations etc)