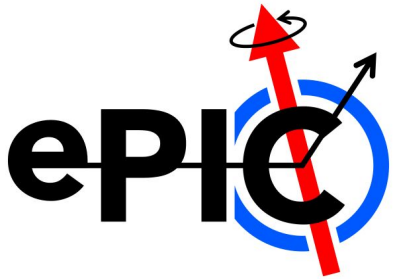
Document Management for ePIC

M.Potekhin on behalf of the “ad-hoc” Committee on Collaborative Tools:

J. Lajoie,
S. Dalla Torre,
M. Diefenthaler,
M. Potekhin,
E. Sichtermann,
P. Steinberg,
B. Surrow,
T. Ullrich,
M. Zurek



The ePIC Collaboration General Meeting
05/17/2024

Overview

- ePIC has an urgent need of an effective document management system, which requires technology selection and policy decisions.
- To that end, the Collaboration leadership formed the so called “ad-hoc” Collaborative Tools Committee in April 2024, tasked with formulating the optimal approach and making a recommendation to ePIC.
- The Committee’s activities included studies and testing of the technical aspects of the available platforms and consideration of future policy directions.
- *The recommendation is now ready and it consists of adoption of zenodo.org for immediate and medium term use in ePIC.*

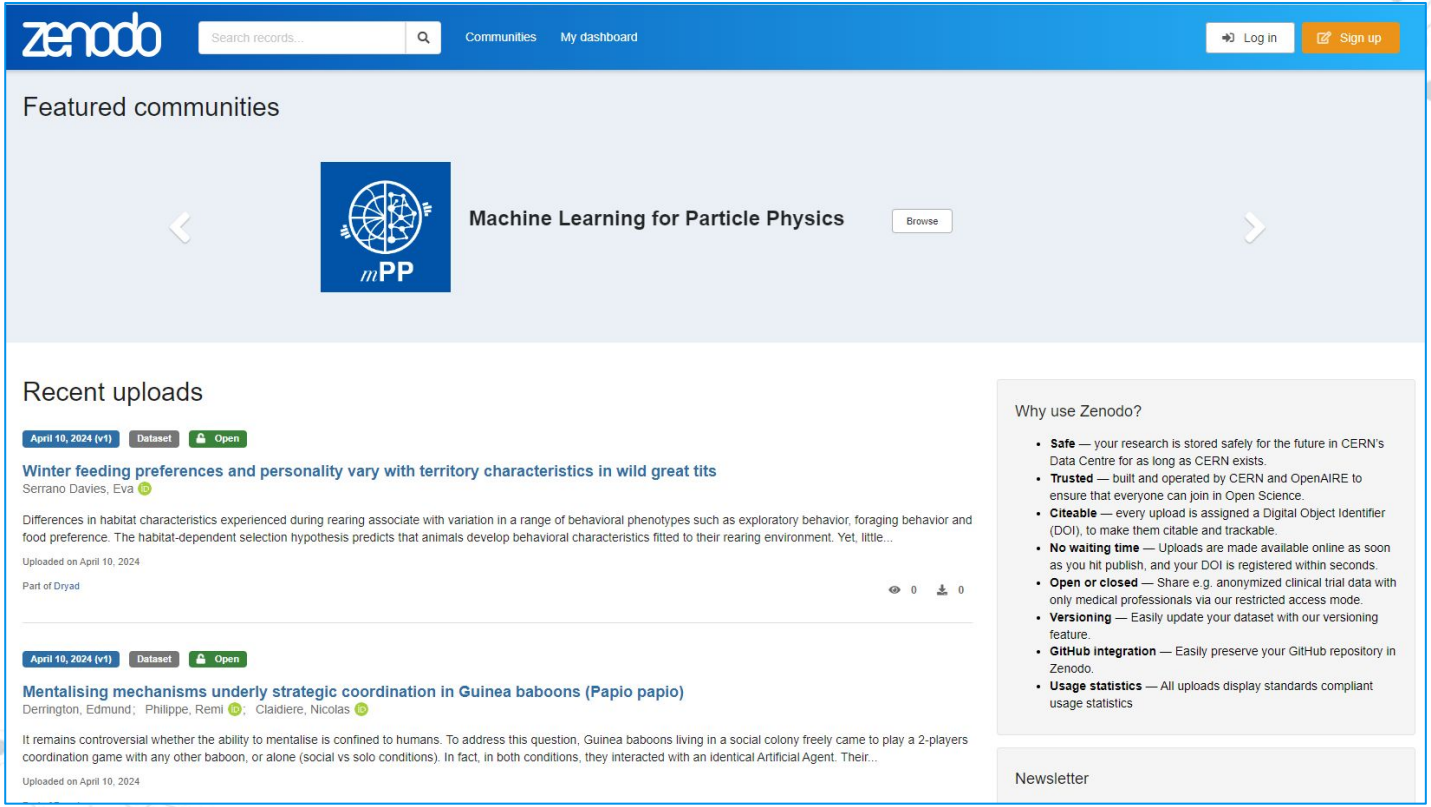
Definitions

- In the context of this discussion, by "document management" we mean an effective repository of research materials, with strong search capabilities, versioning, access control and other useful features.
- By contrast, the "document development" includes the functionality to manage the collaborative process of creation, development and improvement of papers and other materials (cf. GitHub issues, Overleaf, comments on Google Docs, CERN CDS etc).
- The most immediate need is for the "document management" since the development functionality can be covered satisfactorily by Overleaf, GitHub etc until we may eventually arrive to an integrated solution.

Invenio, Invenio RDM, Zenodo… What are the differences?

- Invenio is a software framework developed at CERN and used in a considerable number of systems (HEPData, OpenData, zenodo.org etc)
- Over the years Invenio underwent evolution and became portable and usable outside of CERN, culminating in the creation of “Invenio RDM” .
- RDM stands for “Research Data Management” — *Invenio RDM* is a specific Open Source Application built on top of Invenio
- zenodo.org is a service: an application using most of the mechanics of Invenio RDM, hosted at CERN. *It has many similarities and some differences with Invenio RDM, while currently sharing a lot of the underlying framework*.

https://zenodo.org/ – landing page
Named after **Ζηνόδοτος**, the inventor of metadata (280 BC)



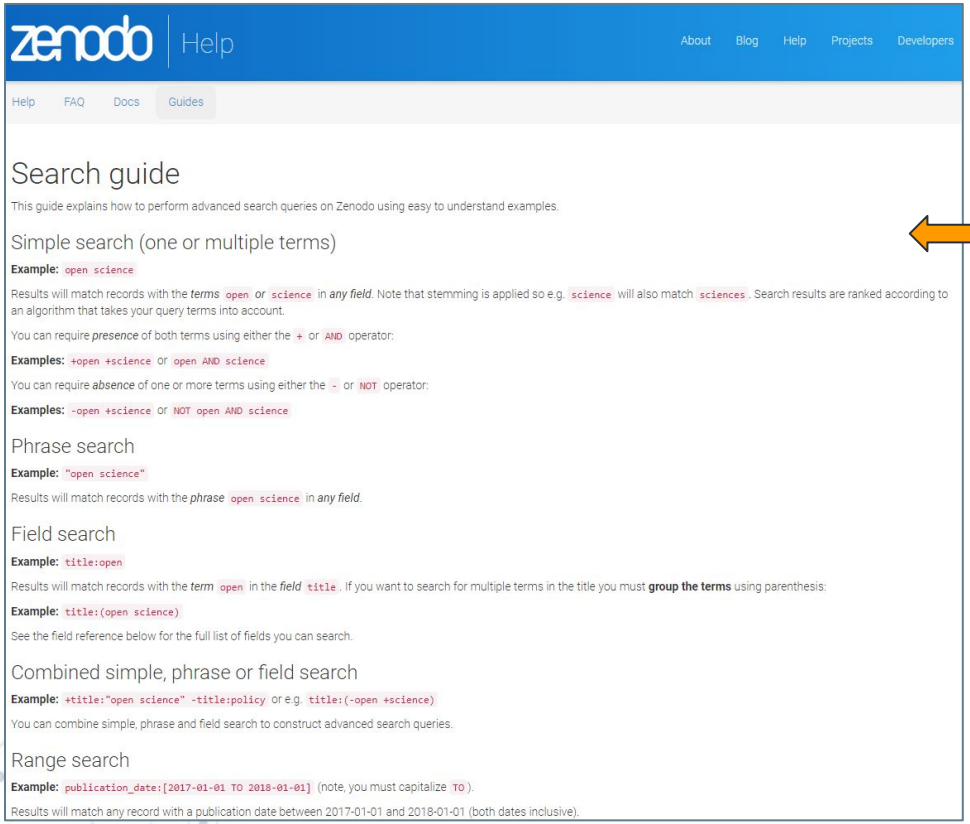
zenodo.org: creating an account

- zenodo.org is an open access service, accounts are free
- It supports the creation of “local” accounts, i.e. one can sign up simply using their e-mail address and create a password.
- In addition, the website also provides the possibility of using the user’s GitHub or ORCID credentials.
- *ePIC members are encouraged to create accounts using one of these methods and explore the system.*

Zenodo and Invenio RDM: key features

- *Due to the technology convergence, many (but not all) features are shared or similar*
- **Search**: powerful search capabilities including “elasticsearch”, detailed search on the metadata and an unlimited number of optional keywords added to committed items.
- **Data Format Flexibility**: data in any format can be handled, with preview capability for PDF, graphics etc
- **Curated Communities**: an organization can create a managed collection of documents, and enhance its brand and findability
- **Access control**: a few access levels managed by the organization

Search functionality, extensively documented
https://help.zenodo.org/guides/search/



zenodo | Help

About Blog Help Projects Developers

Help FAQ Docs Guides

Search guide

This guide explains how to perform advanced search queries on Zenodo using easy to understand examples.

Simple search (one or multiple terms)

Example: `open science`

Results will match records with the *terms* `open` *or* `science` in *any field*. Note that stemming is applied so e.g. `science` will also match `sciences`. Search results are ranked according to an algorithm that takes your query terms into account.

You can require *presence* of both terms using either the `+` or `AND` operator:

Examples: `+open +science` or `open AND science`

You can require *absence* of one or more terms using either the `-` or `NOT` operator:

Examples: `-open +science` or `NOT open AND science`

Phrase search

Example: `"open science"`

Results will match records with the *phrase* `open science` in *any field*.

Field search

Example: `title:open`

Results will match records with the *term* `open` in the *field* `title`. If you want to search for multiple terms in the title you must **group the terms** using parenthesis:

Example: `title:(open science)`

See the field reference below for the full list of fields you can search.

Combined simple, phrase or field search

Example: `+title:"open science" -title:policy` or e.g. `title:(-open +science)`

You can combine simple, phrase and field search to construct advanced search queries.

Range search

Example: `publication_date:[2017-01-01 TO 2018-01-01]` (note, you must capitalize `TO`).

Results will match any record with a publication date between 2017-01-01 and 2018-01-01 (both dates inclusive).

Just a fragment of the help page – lots of information there

Rich, flexible functionality

+Elasticsearch

zenodo.org: bonus features

- **Conference Awareness**: Zenodo organically incorporates optional information about conferences. This is very helpful.
- **GitHub Integration**: software releases tagged in GitHub can be committed to Zenodo in a straightforward manner, and assigned DOIs. This results in *citeable software* which is increasingly considered an important capability.
- **DOI**: one of the factors to ensure durability of the materials (see the DOI slide below for more info) – this is not provided by RDM out of the box
- Account creation and auth/auth **ORCID and/or GitHub**: this can be very useful if ePIC chooses to rely on ORCID for uniquely identifying its members, which is one of current considerations. This feature has been tested by ePIC members.

Experience in PHENIX: move to Zenodo in 2020

- Zenodo plays a crucial role in the successful Data and Analysis preservation effort in PHENIX. It replaced a legacy PHP/DB-based document catalog which had a limited feature set, limited visibility and was increasingly difficult to maintain.
- The learning curve turned out to be quite simple.
- At the time of writing, 670 items produced by PHENIX have been committed and indexed with curated keywords, including more than 160 PhD thesis documents and a large number of the conference presentations, and some tutorials and technical write-ups.
- Please see the backup slides for more detail.

DOI (Digital Object Identifier)

- Please see https://www.doi.org/ for details. Quote:
  - *The DOI Foundation is a not-for-profit organization. We govern the Digital Object Identifier (DOI) system on behalf of the agencies who manage DOI registries and provide services to their respective communities. We are the registration authority for the ISO standard (ISO 26324) for the DOI system and we are governed by our **Registration Agencies**.*
- Using the DOIs facilitates the long-term persistence of the materials committed to various repositories as it is conducive to designing and implementing future migrations of the repository platforms. A DOI (included in a link) will be *resolved* to the same document, for posterity — if everything is done right. Example: 10.5281/zenodo.7052131
- …but minting the DOIs is done by external agencies and costs money. It's not very cheap. *That is not our concern* if we elect to use the CERN instance, in which case it's free.

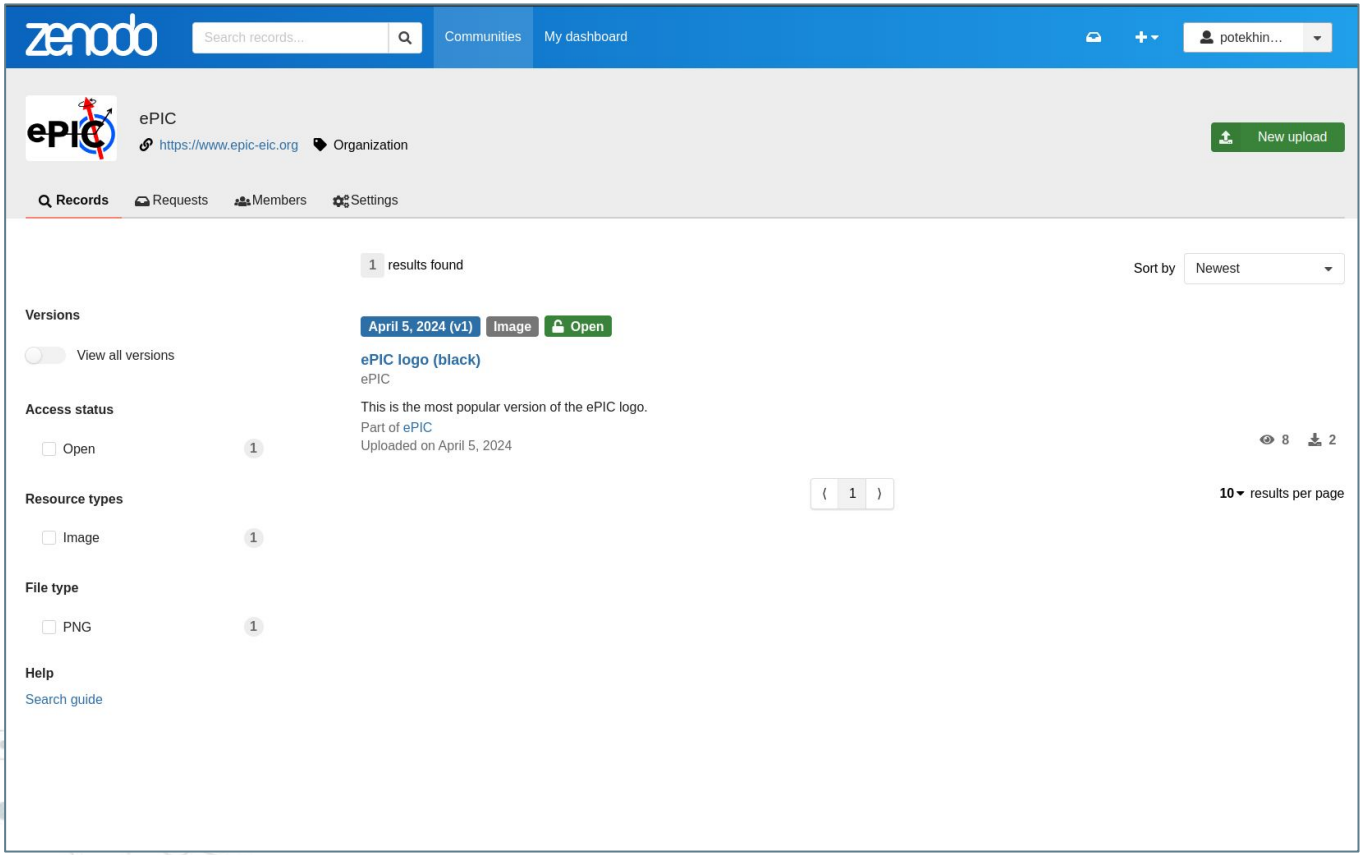
Community management

- The Zenodo "community" is a *collection of documents* managed by a research organization. There are many communities on the Zenodo platform. It adds visibility, findability, brand recognition.
- Levels of management and access:
  - Owner: has all powers including appointing managers
  - Manager: manage members, curate records, view restricted records
  - Curator: curate records, view restricted records
  - Reader: view restricted records
- Curation — a material is submitted to a community for a review. It is then reviewed (with comments if needed) and a decision is made whether to include it in the official collection. Automatic e-mail messages are helping to manage the process, and it's well implemented on the website.

Access control

- Having the ability to control access to specific materials is a critical requirement for most document management systems and it certainly is in ePIC. In the current version of Zenodo deployed at CERN, records can be either *public or restricted*.
- To access a restricted record, a person can be either registered as the "reader" in the ePIC community, or create an access request to be handled by the community managers.
- Note that there is no such thing as a completely invisible record once it's published. It may still be restricted (i.e. content unavailable to general public) but it would be still listed.

A prototype of the ePIC Community hosted on zenodo.org



Pragmatic considerations

- zenodo.org does not require any investment or much preparation to be immediately put into use by the Collaboration. Its advantages were mentioned in the “bonus features” section and notably include minting of the DOIs, which has an additional importance due to the possibility of future data migration (see next slides). Additionally, Zenodo enjoys a high level of support from CERN personnel and principal developers who are located at CERN.
- Integration with ORCID and/or GitHub for user authorization, offered by Zenodo, is a potentially important feature.

Will Future Data Migration be possible?

- Both Invenio RDM and Zenodo feature a complete RESTful API, which implies that any function is accessible via HTTP (for example, a Python or a shell script, in addition to the Web UI).
- Both platforms use the same, well designed and complete JSON schema in their interfaces.
- We have conducted a study and a test of this interface to ensure that migration of data out of Zenodo to a different platform and/or location (e.g. an Invenio RDM instance at BNL or elsewhere) is possible.
- The answer is positive, i.e. future migration will be possible and there is no “vendor lock-in” embedded in our present choice of the document management platform.

Next steps

- ePIC can start using zenodo.org immediately. Effective utilization of this resource will necessitate the creation of optimal policies for both users and working groups. Before we consider such policies, we would like to solicit your feedback via this Google Form

https://docs.google.com/forms/d/e/1FAIpQLSfW2o0JxkkSC8m3CD7cExHpH2YJ1QA_llJcG-kwrTnS42kj8g/viewform?usp=pp_url

Summary

- zenodo.org is a service for the international science community and is widely used around the globe.
- The committee formed in ePIC with the purpose of evaluating the available options for document management completed its due diligence and arrived to the conclusion that immediate use of zenodo.org is the optimal way forward.

Backup

Data Migration – importance of the DOIs

- Consistent reliance on JSON schemas used in the REST API depends in a large part on proper unique identifiers i.e. DOIs.
- Note that the DOIs are used not just for external access to a particular article, slide or figure, but as a crucial database keys that binds together a substantial amount of records pertaining to each items, its versions, metadata and related records. In that regard, Zenodo is a self-sufficient system.

Location of the Data Storage

- ePIC is an international Collaboration with world-wide participation.
- It is not a part of the DOE or its agencies.
- There is no requirement to store documents produced by an international team of scientists on the US soil.

Zenodo integration with the PHENIX website

- Starting in 2019, the PHENIX Collaboration migrated its website to a new platform based on Jekyll and leveraging GitHub pages. It was designed and implemented by the NPPS group at BNL.
- Reliance on zenodo.org for material storage allowed us to create a lightweight, portable and easy to maintain website, since it by itself does not contain much data.
- The website is used to find a serve the content from Zenodo using a number of links and importantly, a *collection of curated keywords* also available as links from its web pages. This is optional but in practice turned out to be extremely useful.

PHENIX: keyword references to zenodo.org

PHENIX Keywords

Listed on this page are *recommended* keywords used for two purposes:

- to **tag materials** placed on this site so they can be consistently referenced across this web resource
- to enhance discoverability of the PHENIX materials committed to Zenodo

Consistent use of the keywords is strongly recommended. Please note that HEPData materials follow a different set of conventions.

The keywords are case-sensitive. We adopted lowercase convention for all keywords for the following reason: Zenodo is using a complex query n
so it's best to avoid ambiguity. Note that Zenodo keywords can actually be a combination of words and white space (i.e. phrases). Multiple such co

In the tables below, the keywords are grouped in categories. Each entry in the left column acts as a query link to *Zenodo*, for that specific keyword.

General (12 items)

Keyword	Description
alice	ALICE - an experiment at CERN
atlas	ATLAS - an experiment at CERN
bup	Beam Use Proposal
cms	CMS - an experiment at CERN
decadal plan	Two long-term PHENIX research proposals
phenix	PHENIX - an experiment at RHIC
phobos	PHOBOS - an experiment at RHIC
rhic	Relativistic Heavy Ion Collider (RHIC)
sphenix	sPHENIX - an experiment at RHIC
star	STAR - an experiment at RHIC
tutorial	a category of PHENIX documents on Zenodo
wa98	WA98 - an experiment at CERN

Conferences (117 items)

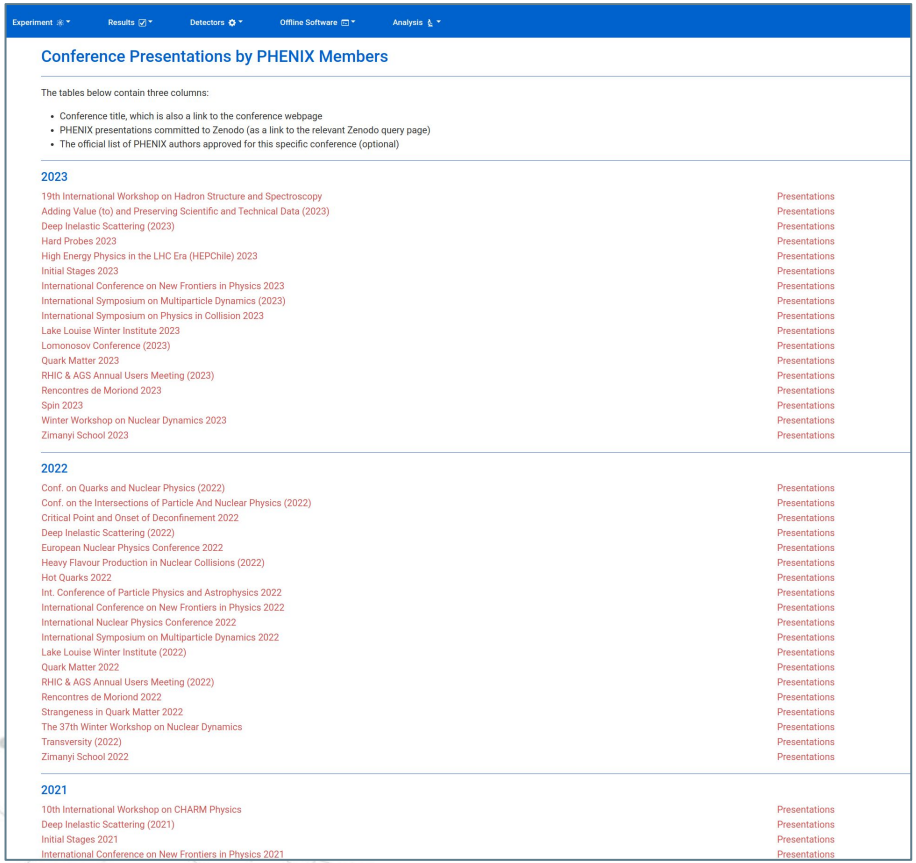
Keyword	Description
aum16	RHIC & AGS Annual Users Meeting (2016)
aum17	RHIC & AGS Annual Users Meeting (2017)
aum18	RHIC & AGS Annual Users Meeting (2018)

Keywords are classed as “general”, “conferences”, “detector”, “physics”, “software” etc, for ease of navigation.

They are rendered as functioning links to Zenodo queries.

Keywords are documented in the easy to read and maintain YAML files, which are rendered automatically on the website.

The conference page



Experiment | Results | Detectors | Offline Software | Analysis

Conference Presentations by PHENIX Members

The tables below contain three columns:

- Conference title, which is also a link to the conference webpage
- PHENIX presentations committed to Zenodo (as a link to the relevant Zenodo query page)
- The official list of PHENIX authors approved for this specific conference (optional)

2023

19th International Workshop on Hadron Structure and Spectroscopy	Presentations
Adding Value (to) and Preserving Scientific and Technical Data (2023)	Presentations
Deep Inelastic Scattering (2023)	Presentations
Hard Probes 2023	Presentations
High Energy Physics in the LHC Era (HEPChile) 2023	Presentations
Initial Stages 2023	Presentations
International Conference on New Frontiers in Physics 2023	Presentations
International Symposium on Multiparticle Dynamics (2023)	Presentations
International Symposium on Physics in Collision 2023	Presentations
Lake Louise Winter Institute 2023	Presentations
Lomonosov Conference (2023)	Presentations
Quark Matter 2023	Presentations
RHIC & AGS Annual Users Meeting (2023)	Presentations
Rencontres de Moriond 2023	Presentations
Spin 2023	Presentations
Winter Workshop on Nuclear Dynamics 2023	Presentations
Zimanyi School 2023	Presentations

2022

Conf. on Quarks and Nuclear Physics (2022)	Presentations
Conf. on the Intersections of Particle And Nuclear Physics (2022)	Presentations
Critical Point and Onset of Deconfinement 2022	Presentations
Deep Inelastic Scattering (2022)	Presentations
European Nuclear Physics Conference 2022	Presentations
Heavy Flavour Production in Nuclear Collisions (2022)	Presentations
Hot Quarks 2022	Presentations
Int. Conference of Particle Physics and Astrophysics 2022	Presentations
International Conference on New Frontiers in Physics 2022	Presentations
International Nuclear Physics Conference 2022	Presentations
International Symposium on Multiparticle Dynamics 2022	Presentations
Lake Louise Winter Institute (2022)	Presentations
Quark Matter 2022	Presentations
RHIC & AGS Annual Users Meeting (2022)	Presentations
Rencontres de Moriond 2022	Presentations
Strangeness in Quark Matter 2022	Presentations
The 37th Winter Workshop on Nuclear Dynamics	Presentations
Transversity (2022)	Presentations
Zimanyi School 2022	Presentations

2021

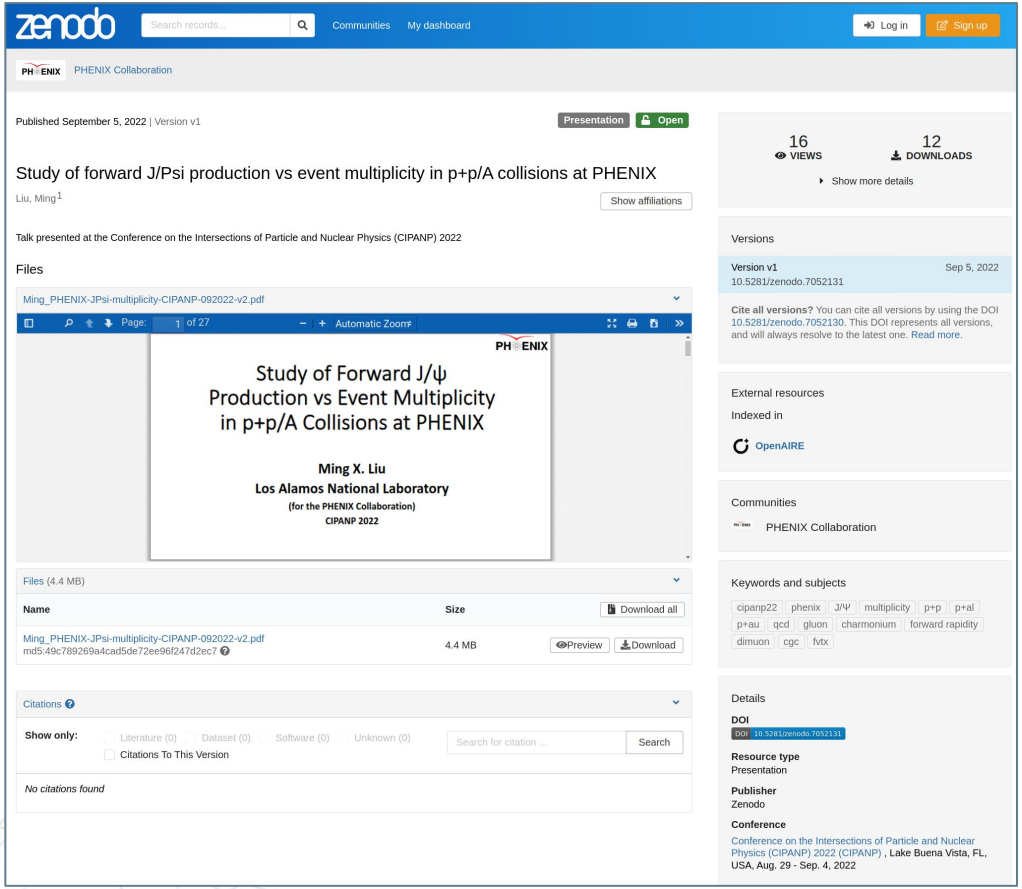
10th International Workshop on CHARM Physics	Presentations
Deep Inelastic Scattering (2021)	Presentations
Initial Stages 2021	Presentations
International Conference on New Frontiers in Physics 2021	Presentations

Using Zenodo, we preserved presentations by PHENIX members, made at 117 conferences (and counting), with modest effort.

The conference page contains links to collections of slides for specific conferences, as well as to the conference pages themselves. All rendering of the YAML content is 100% automatic.

Reliance on the Zenodo instance at CERN provides a degree of certainty of durability of the materials thus preserved, and transparent access for years to come.

An example of a PHENIX entry (conference slides)



Notable features:

- Custom keywords
- Slide viewer
- Community attribution
- Reference to the conference

zenodo.org: beyond PHENIX

- The EIC User Group (EICUG) solved its long-standing problem with document management (previously in ad-hoc storage in Drupal) by migrating to Zenodo in 2022.
- The LuSEE-Night project (astrophysics) which has considerable leadership and science contributions from BNL, has also started using zenodo.org in 2023.
- Apparently hosting reference astronomy data on zenodo.org is a common practice.