

# Activities in support for digital document repositories -Invenio and Zenodo-



**[Carlos Fernando Gamboa \(cgamboa@bnl.gov\)](mailto:cgamboa@bnl.gov)**

April 30th 2020

# Invenio

An Open Source framework for large scale digital repositories developed at CERN

The screenshot shows the Zenodo website. At the top, there is a search bar and navigation links for 'Upload' and 'Communities'. Below the header, there are several sections: 'Recent uploads' with a list of items including 'Historical climate model output of ECHAM5-wiso from 1871-2011 at T106 resolution' and 'Reproducibility Package for "Reproducible research and GIScience: an evaluation using AGILE conference papers"'. There are also sections for 'Using GitHub?', 'Zenodo in a nutshell' (listing features like Research Shared, Citable/Discoverable, and Communities), and 'Software Citation and the Wikidata Ecosystem'.

**Zenodo**  
~100 TBs / 2M records

General purpose research data repository

The screenshot shows the CDS Video Platform interface. At the top, there is a search bar and navigation links for 'Upload' and 'Log In'. The main content area features a large video player with the title 'LINAQ4 joins the CERN accelerator chain'. Below the video player, there is a 'RECENT' section titled 'MOST RECENTLY ADDED VIDEOS' with three video thumbnails: 'CMS Virtual Visit from Portugal, 2018.04.11', 'CMS Virtual Visit from Hungary, 2018.04.17', and 'CMS Virtual Visit from Italy, 2018.05.04'.

**CERN Video Platform**  
~35 TBs / >5000 videos

Digital Assets Management system with video encoding support

The screenshot shows the CERN Open Data Portal interface. At the top, there is a search bar and navigation links for 'About'. The main content area features a large heading 'Explore more than 1 petabyte of open data from particle physics!' and a search bar with the text 'Start typing...'. Below the search bar, there are sections for 'Explore' (listing datasets, software, environments, and documentation) and 'Focus on' (listing ATLAS, ALICE, CMS, LHCb, and OPERA). There are also sections for 'Learn' (Discover the world of open data from particle physics) and 'Analyse' (Run your own physics analyses, start virtual machines). At the bottom, there are links for 'Welcome to our updated portal' and 'CMS Guide to research use of CMS Open Data'.

**CERN Open Data Portal**  
1.7 PBs / >620K files

Courtesy of B. Gonzalez.

Examples of digital repositories on Invenio 3 hosted at CERN

# Invenio 3 framework

**Integrated in a scalable software architecture which addresses issues observed on previous releases, for example Invenio 1 used by CDS.**

- Invenio 1 simple installation, difficult to customize and scale.
- Invenio 3 allows communities to develop applications in a scalable and supported framework.
  - **Zenodo is a digital repository built on Invenio 3**

## **Flexible metadata**

- Flexible record and persistent identifier store.
- Record can use custom or standard metadata formats like JSON-LD, MARC21, Datacite.
- Invenio can manage bibliographic records, authority records, grants among others.
- Elasticsearch is leveraged by Invenio to provide scalable and complex searching capability.
- DOI (Digital Object Identifier) support to records to be properly citable.

## **Accessibility enabled for web UI or programmatically via a REST API**

- Implemented for metadata and files
- Invenio support S3, XRootD, WebDAV, among others.

## **State of the art authentication/authorization implementation**

- Authentication SSO via Github, Orchid out of the box.

**Invenio's modular framework allows interaction with Github, ORCID and expand its functionality**

# Background

- In December of 2017 we started to investigate digital repository options for BNL's science programs.
- Zenodo came as the natural choice.
- After evaluation and testing, it was as implemented as a R&Ds testbed for NNSD, CSI users.
- Now Invenio it is a service supported as part of SDCC mission to provide digital repositories.

To implement this strategy BNL has been an active partner of InvenioRDM since its creation.

# Evolution of Invenio related projects in the SDCC

## Zenodo test evaluation instance

Zenodo software not delivered as a product but as a service (CERN centric code), there is not support for software upgrades.

Dec/17

04/30/20

Development  
Integration  
Operation

## BNL customs repositories based on Invenio 3

### SET (Non proliferation and National Security Department)

Oct/18

Developer working in custom based Invenio application

DEC/18

### GENESIS ( Materials Science)

### sPhenix ( NPP)

Functional site in Jun/20

Code forked and adapted

Oct/19

## Zenodo for DOE COVID-19

Apr/20

InvenioRDM a product to be supported by CERN and the Invenio community

SDCC is part of the initial InvenioRDM project

InvenioRDM (will replace Zenodo)

Zenodo->InvenioRDM  
Dec/20

Jul/19

Kickoff meeting

Sept/20

Final Release

# Digital repositories hosted at SDCC

SDCC supports custom data repositories based on INVENIO for different scientific communities

Invenio based custom applications:

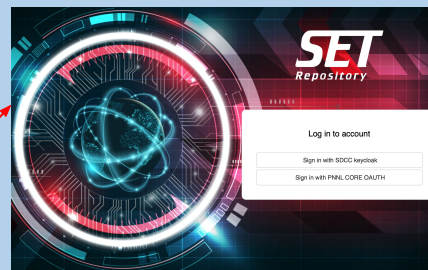
**National Nuclear Security Administration (NNSA)**

Application **SET**, Smuggling Detection and Deterrence Science and Engineering Team

**Materials Science community**

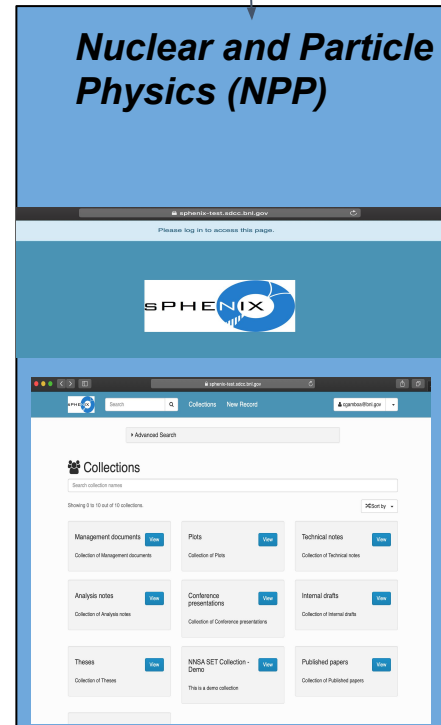
Application **GENESIS**, Next-Generation Synthesis Center

Repositories in operation



Repository in development and testing

**Nuclear and Particle Physics (NPP)**



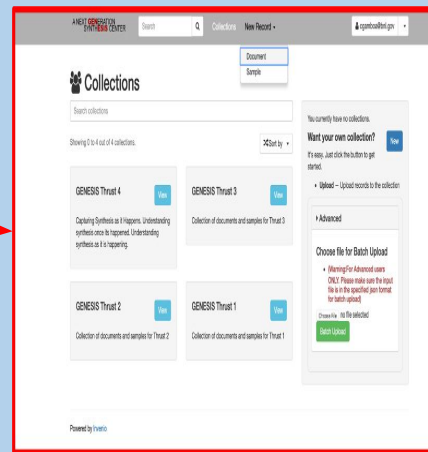
SDCC contributing to Invenio's international community



**Invenio Research Data Management (RDM) platform**

SDCC is working to build a research data management platform called [InvenioRDM](#) along with CERN and other multidisciplinary and commercial institutions.

InvenioRDM is a platform allows researchers to share and preserve scientific results. **InvenioRDM will replace Zenodo.**

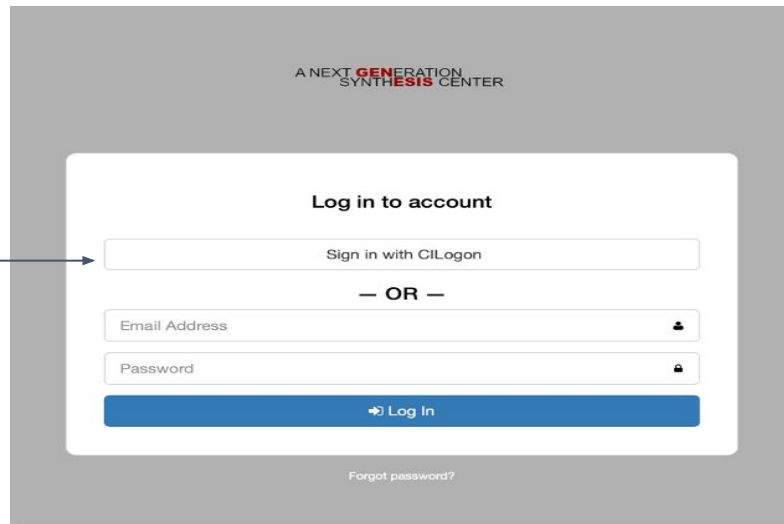


SDCC supports infrastructure for Invenio based applications, along with customized network, storage and Authentication infrastructure enabled to host services.

## Versatile authentication mechanisms supported and integrated with the applications to provide SSO accessibility to users

### ***GENESIS, Materials Science community***

CILogon, Federated ID (InCommon / COMmange) support for authentication using Invenio's oauth native module.



### ***SET, National Nuclear Security Administration (NNSA)***

Due to the type of data been managed (OUO) provisioning the production environment required the deployment of isolated resources.

- 2FA DUO capable for authentication using Keycloak and Invenio's oauth native module. (Integrates ITD Active Directory + DUO)
- 2FA using PNNL's IDP
- DOE OneID federation to be integrated in May 2020

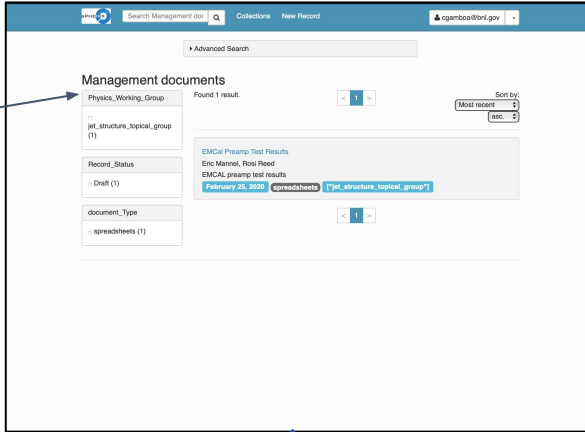
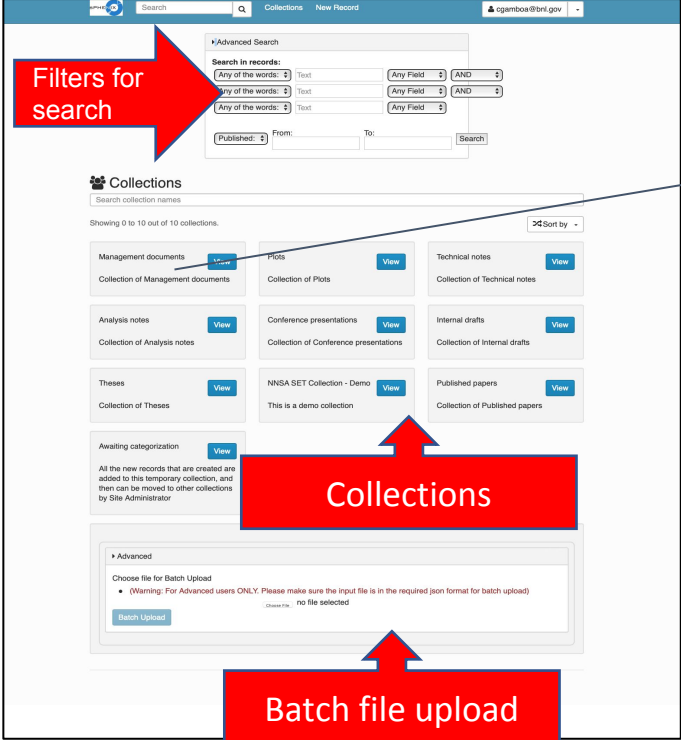
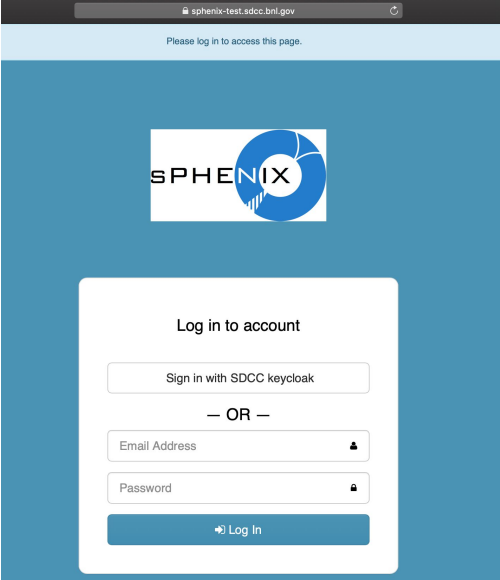


# sPHENIX Document store

## Invenio custom application

in Beta release

Goal is to have a fully functional site by June 2020



Integrated with SSO using SDCC keycloak infrastructure

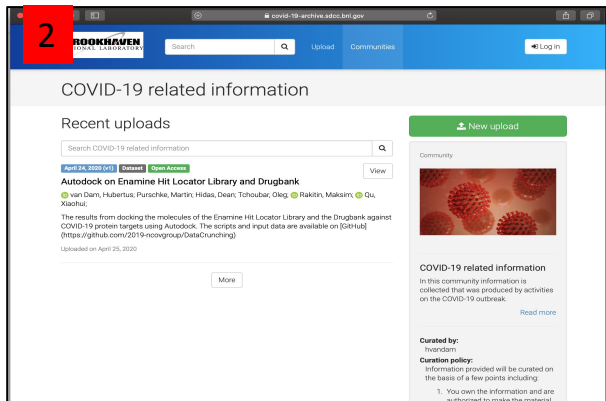
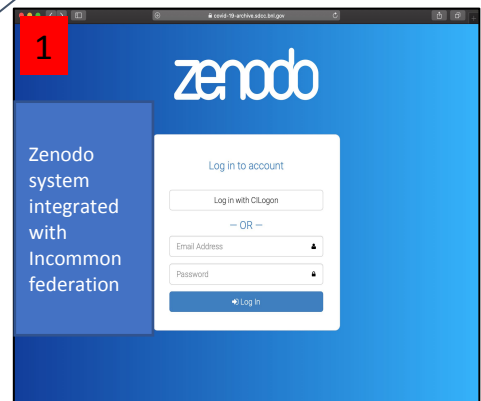


# Zenodo is an application based on Invenio 3

A BNL custom digital repository is being commissioned to host COVID-19 related digital documents as a part of DOE COVID - Medical Therapeutics project based on Zenodo software.

Installed and configured in a 1.5 week!

User requirements satisfied by Zenodo out of the box, due to the COVID-19 emergency minimization of development was achieved



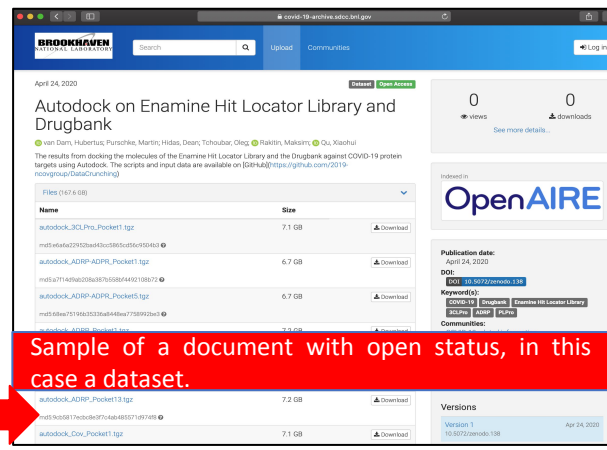
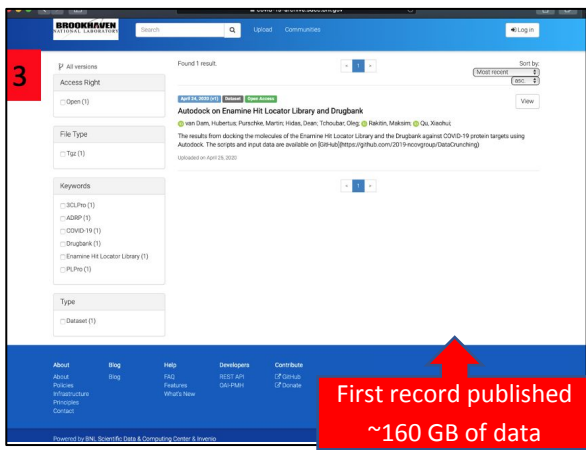
A selected group of researchers uploads and curates the documents in the repository:

1 The selected researches will be able to use their institution's (ANL, ORNL ..., BNL) login and passwords to authenticate to the system

2 A community can be created to collect and curate topic/theme centric aggregation of documents

3 General users will be able to download data (files) from the repository based on document status:

- **Open**, can read and download.
- **Restricted**, can request access.
- **Embargo**, once the embargo period ends the document is publicly available.
- **Closed**, not permitted.



# **In conclusion**

**SDCC has a solid expertise in hosting, developing and operating digital document repositories based on Invenio.**

**Lessons and experience gained while collaborating with different communities allows us to be prompt and agile while supporting new communities.**

**SDCC is supporting digital repositories for different scientific communities within BNL and DOE in the US.**

**SDCC is active within international Invenio community to ensure local community interests are well represented.**

# Contacts

sPHENIX, John Haggerty, Chris Pikenburg

SET, NNSD Warren Stern, Maia Gemmill, Yonggang Cui, Heather Orr (adjunct developer)

GENESIS, Line Pouchard

COVID-19, Kerstin Kleese Van Dam

CSI

Developer, Uma Ganapathy

# Links

Invenio, <https://invenio-software.org>

Invenio RDM, <https://invenio-software.org/products/rdm/>