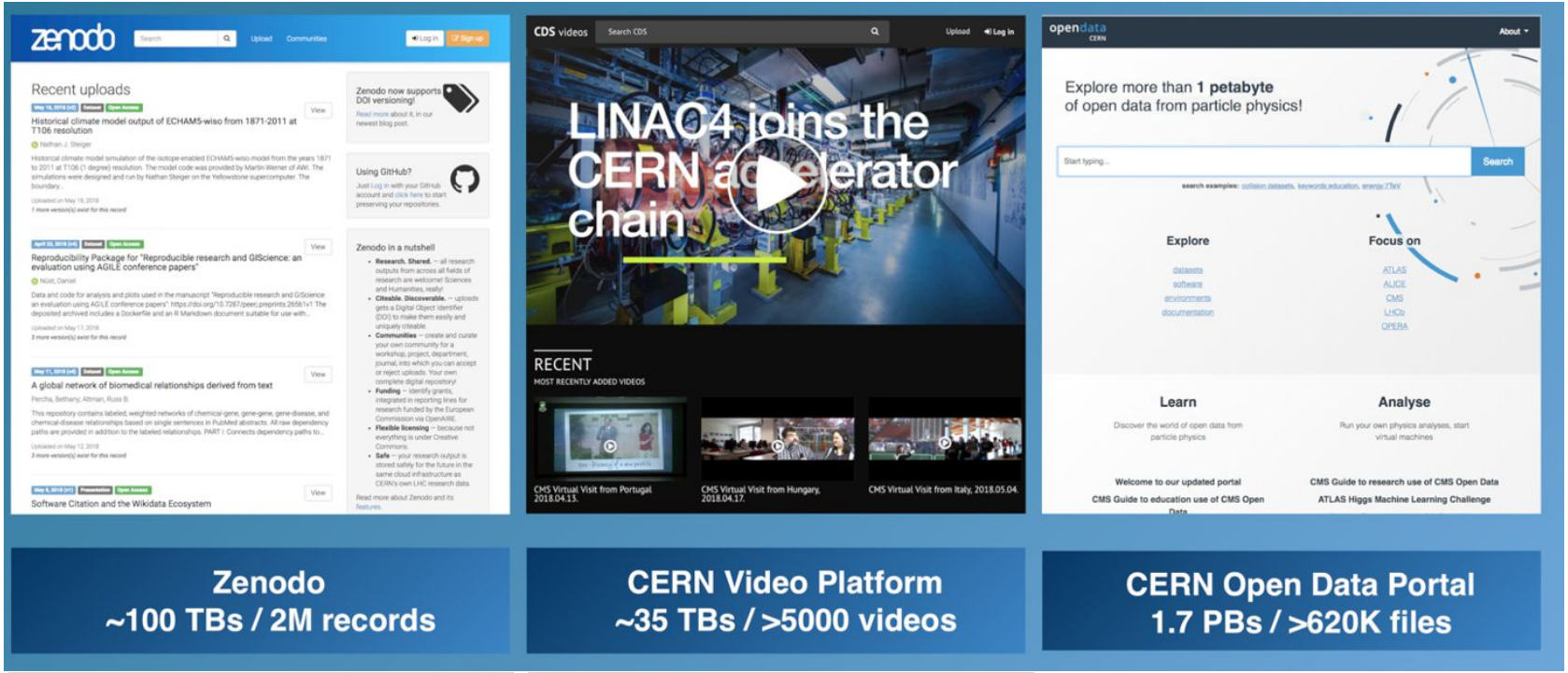
Activities in support for digital document repositories
-Invenio and Zenodo-

[Carlos Fernando Gamboa (cgamboa@bnl.gov)]

April 30th 2020

Invenio

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



Zenodo ~100 TBs / 2M records	CERN Video Platform ~35 TBs / >5000 videos	CERN Open Data Portal 1.7 PBs / >620K files
General purpose research data repository	Digital Assets Management system with video encoding support	

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.

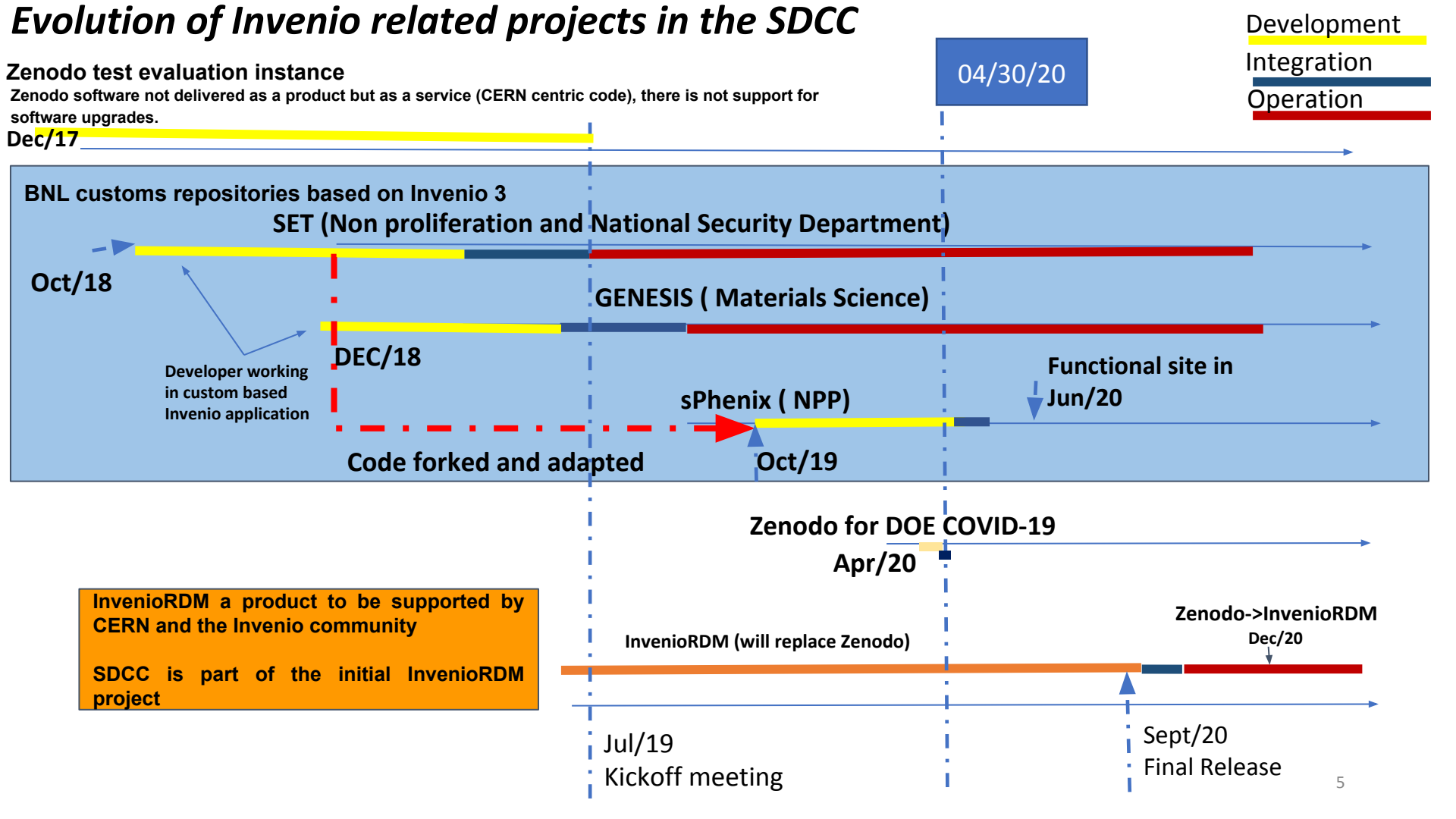
Invenio's modular framework allows interaction with Github, ORCID and expand its functionality

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



Digital repositories hosted at SDCC

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

Invenio based custom applications:

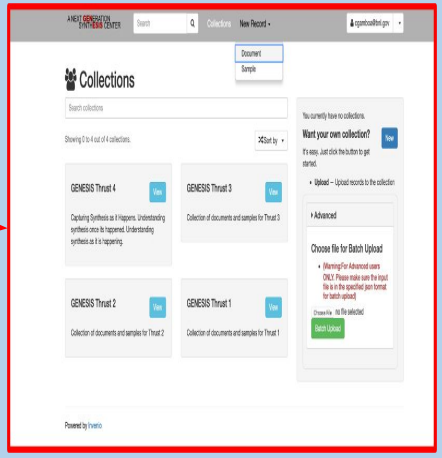
National Nuclear Security Administration (NNSA)

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

Materials Science community

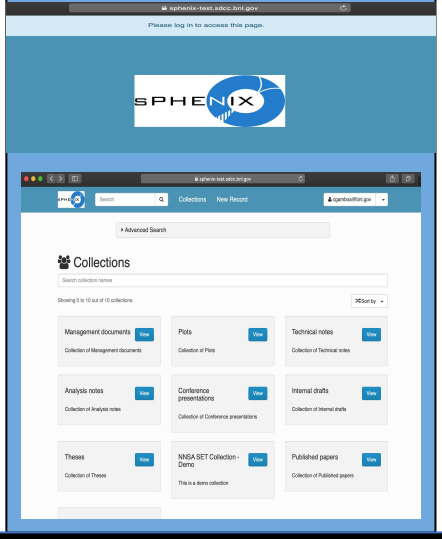
Application **GENESIS**, Next-Generation Synthesis Center

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.

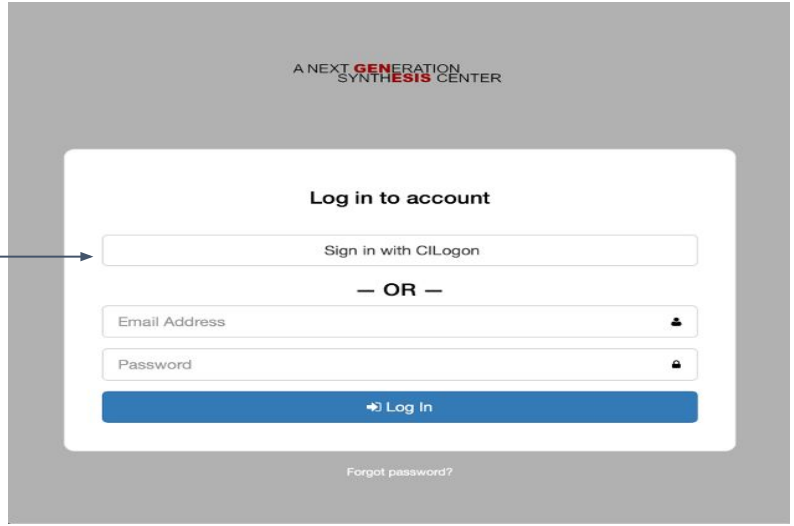
InvenioRDM is a platform allows researchers to share and preserve scientific results. **InvenioRDM will replace Zenodo.**

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.

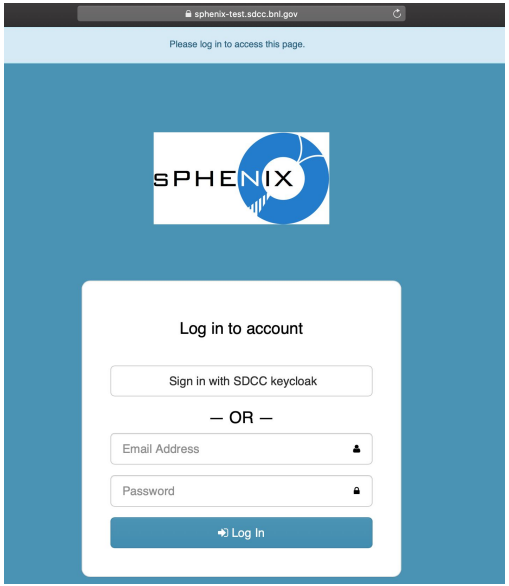
- 2FA DUO capable for authentication using Keycloak and Invenio's oauth native module. (Integrates ITD Active Directory + DUO)
- 2FA using PNNL's IDP
- DOE OneID federation to be integrated in May 2020

sPHENIX Document store

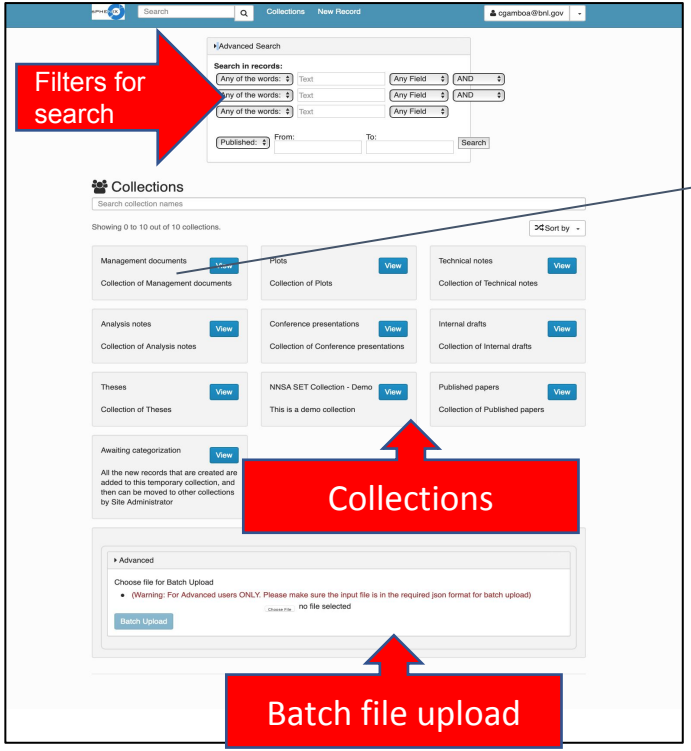
Invenio custom application

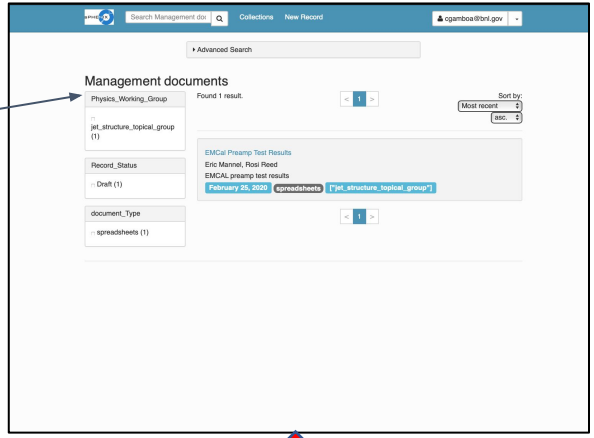
in Beta release

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!

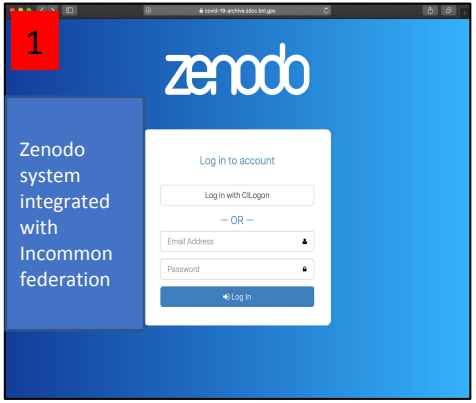
User requirements satisfied by Zenodo out of the box, due to the COVID-19 emergency minimization of development was achieved

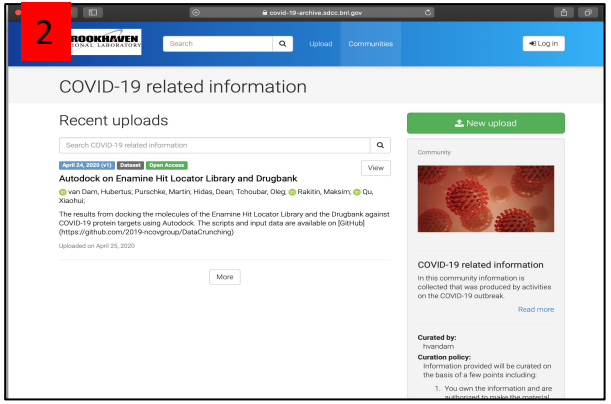
A selected group of researchers uploads and curates the documents in the repository:

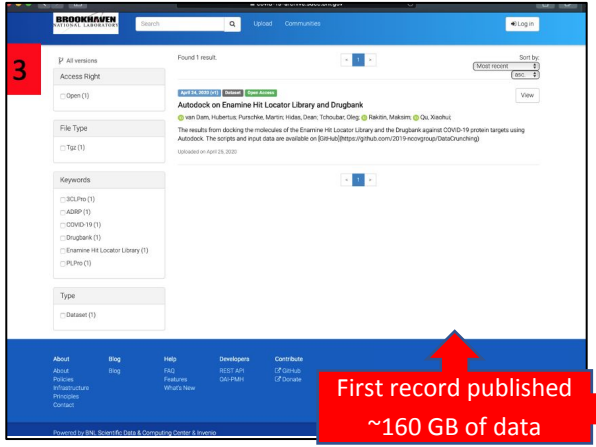
1. The selected researches will be able to use their institution's (ANL, ORNL ,.., BNL) login and passwords to authenticate to the system

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.

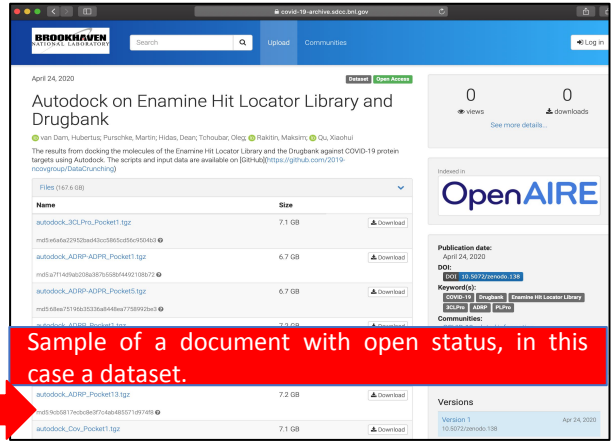






First record published

~160 GB of data



Sample of a document with open status, in this case a dataset.

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/