Invenio: A Digital Document Repository Framework

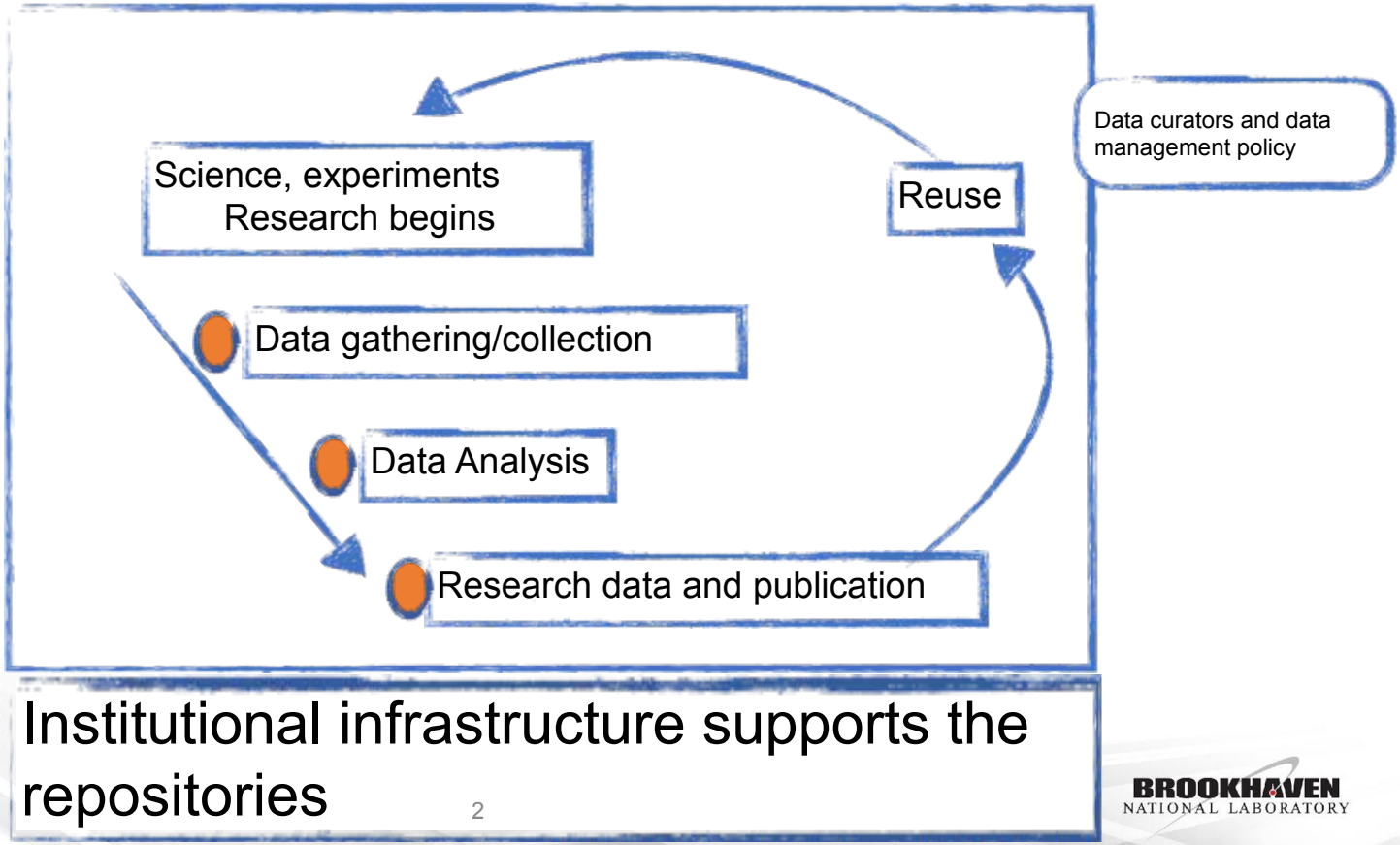
Carlos Fernando Gamboa (cgamboa@bnl.gov),
Scientific Data and Computer Center, August 25 2020
SDCC Technical talks series

BROOKHAVEN NATIONAL LABORATORY
U.S. DEPARTMENT OF ENERGY
BROOKHAVEN SCIENCE ASSOCIATES

A scientific data workflow

- **Research data work**

  Research Digital Management repositories are key elements of infrastructure



Research Digital Management (RDM) repository

A web based service that provides a scientific community a means to share and preserve their scientific results enable reproducibility and empower reuse of datasets

In recent years, RDMs have been adopting ***Open Science*** and ***FAIR*** data policies. In general terms:

- **Open Science:** is the movement to make scientific research and data accessible to all

- **FAIR** refers to a digital record that is:

  Findable, metadata are assigned a globally unique and persistent identifier

  Accessible, metadata are retrievable by their identifier using a standardized communications protocol. Open or by providing Authentication/Authorization schemes

  Interoperable, data need to interoperate with applications or workflows for analysis, storage, and processing

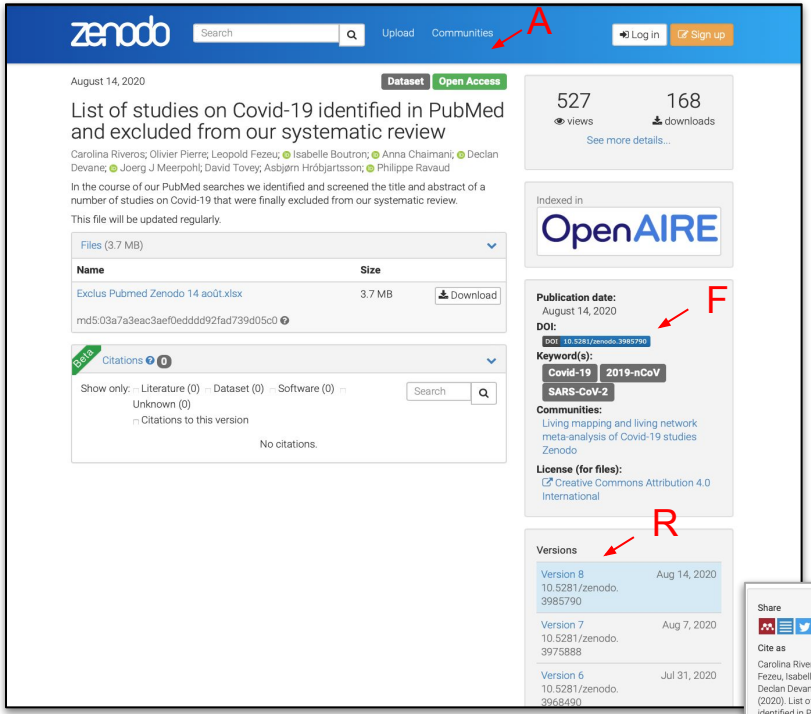
  Reusable, metadata and data should be well-described so that they can be replicated and/or combined in different settings

See backup slide for FAIR data policy

Zenodo (example of a Research Data Management Repository)

https://zenodo.org

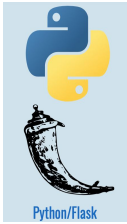
Repository landing page

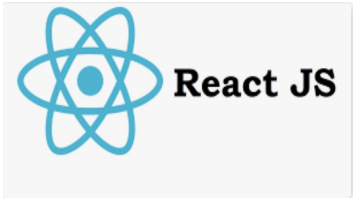


Record landing page

Zenodo is built using **Invenio 3** framework

Invenio 3 is a open source framework to build scalable digital repositories





Integrated in a scalable software architecture

- 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
- DOI (Digital Object Identifier) to support records to be properly citable
- Elasticsearch is leveraged by Invenio to provide scalable and complex searching capability

Accessibility enabled for web UI or programmatically via a REST API

- Implemented for metadata and files
- Invenio supports different data transmission and storage protocols (e.g S3, XRootD, WebDAV, among others)

State of the art authentication/authorization implementation

- Single Sign On and Authentication OAuth allows integration with Github, ORCID out of the box

Invenio software is distributed as modular framework

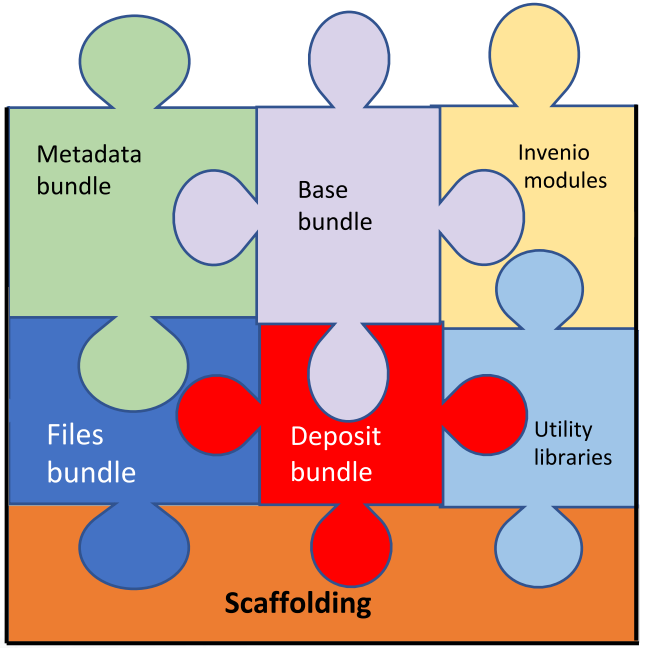
Code is released in modules distributed in ***bundles***
The Base bundle aggregates modules needed to create a generic web application, i.e:

- **Invenio-config:** Configuration loading pattern responsible for loading configuration from Python modules, instance folder and environment variables
- **Invenio-app:** Flask, WSGI, Celery and CLI applications for Invenio including security-related headers and rate limiting
- **invenio-admin:** Administration interface for Invenio based on Flask-Admin

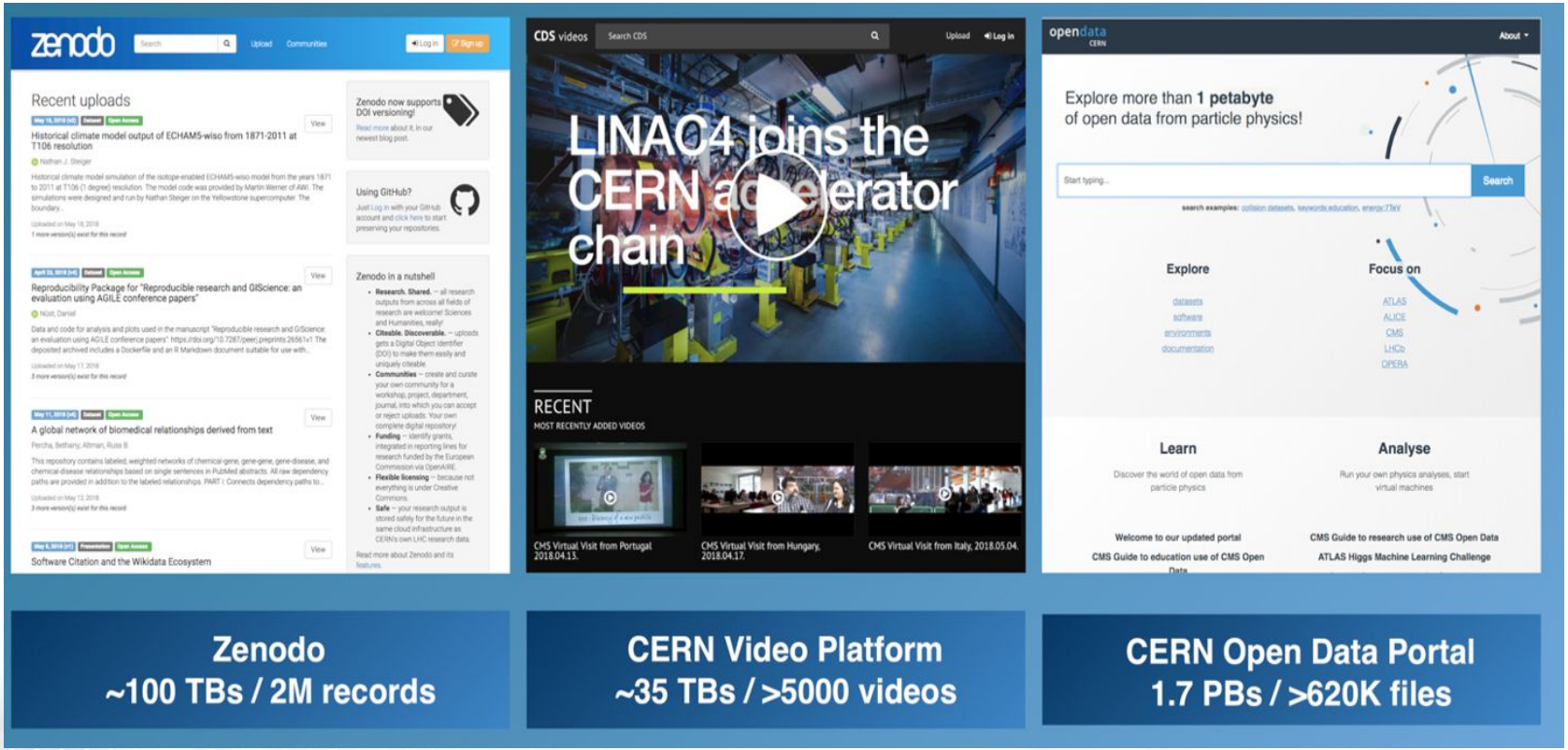
Other bundles and sample modules are :

- **Auth bundle:** invenio-oauth
- **Files bundle:** invenio-files-rest
- **Statistics bundle** (beta): invenio-stats
- **Deposit bundle** (alpha): invenio-deposit
- **Invenio modules** (alpha): invenio-github
- **Utility libraries:** Datacite
- **Scaffolding:** cookiecutter to create base application template

An invenio base digital repository



Examples of digital repositories on Invenio 3 hosted at CERN



Courtesy Jose Benito Gonzalez

Invenio 3 at BNL

Initially we investigated digital repository options for BNL's science programs

- After evaluation and testing Zenodo was implemented as a R&Ds testbed for NNSD and CSI users
- By interacting with the Invenio framework and testing its capabilities, these communities built their own digital repositories to meet their specific needs (SET and GENESIS)

Most recently new BNL scientific communities (NPP) and DOE Medical Therapeutics are interested in a RDM repository like Zenodo

Now Invenio based repositories is a service supported as part of SDCC mission

Digital repositories hosted at BNL

SDCC supports *custom* data repositories based on invenio for different scientific communities:

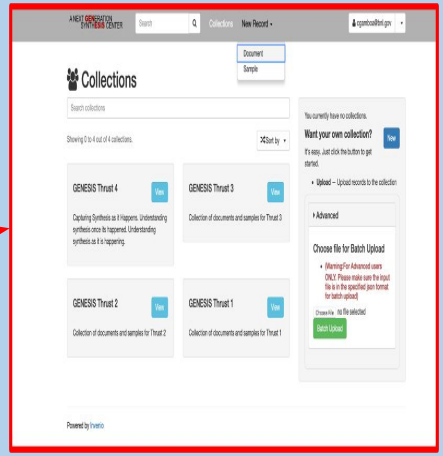
National Nuclear Security Administration (NNSA)

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

Materials Science community

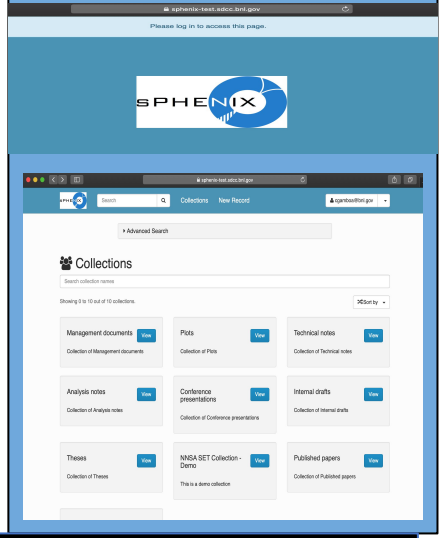
Application **GENESIS**, Next-Generation Synthesis Center

Repositories in operation



Repository in development and testing

Nuclear and Particle Physics (NPP)



SDCC supports infrastructure for Invenio based applications, along with customized network, storage and Authentication infrastructure enabled to host services (production, testing and developing)

Invenio 3 at BNL: Zenodo based repositories EIC-Zenodo

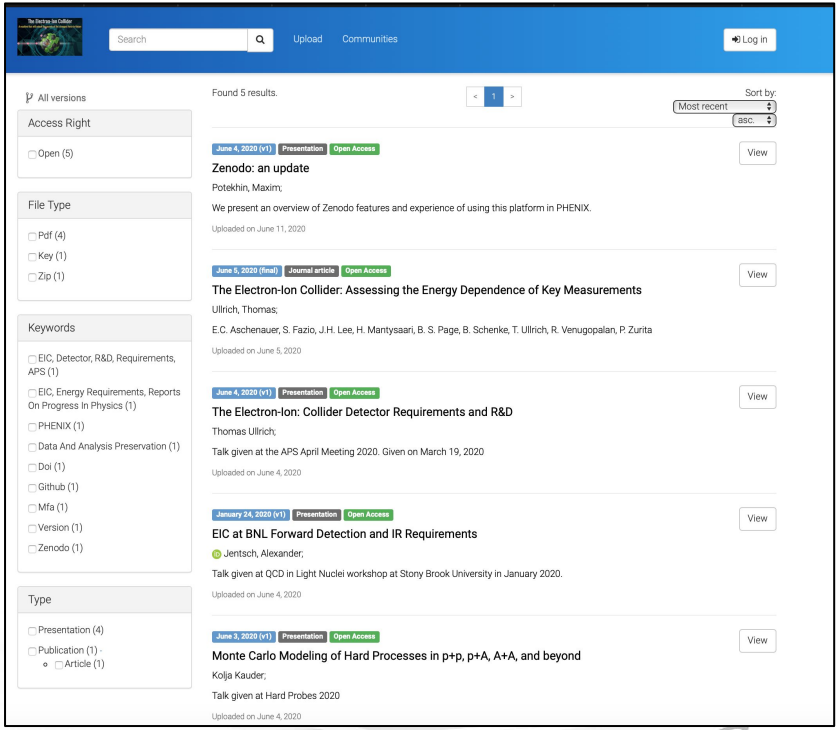
Learning about Zenodo:

- Community was able to experiment its features using a test instance
    - Helped identify requirements

- CILogon, Federated ID (InCommon / COMange) used for authorization (allows to use institutional credentials to login into the web application)

- Based on this experience the community requested a production instance

A production EIC-Zenodo instance being commissioned and is accessible to BNL



Covid-19-archive, a BNL custom digital repository based on Zenodo

Being commissioned to host COVID-19 related digital documents as a part of DOE COVID - Medical Therapeutics project based on Zenodo software DOE project for which BNL receives funding

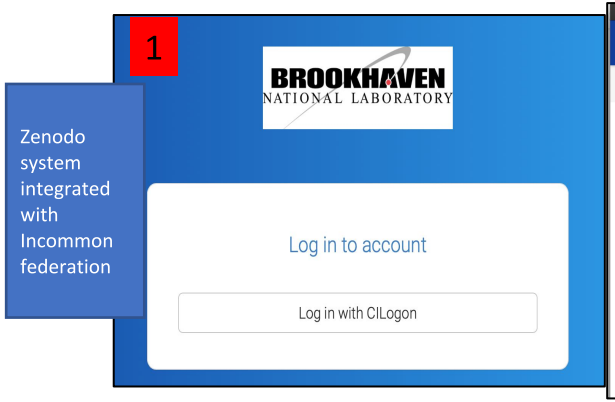
A selected group of researchers uploads and curates the documents in the repository:

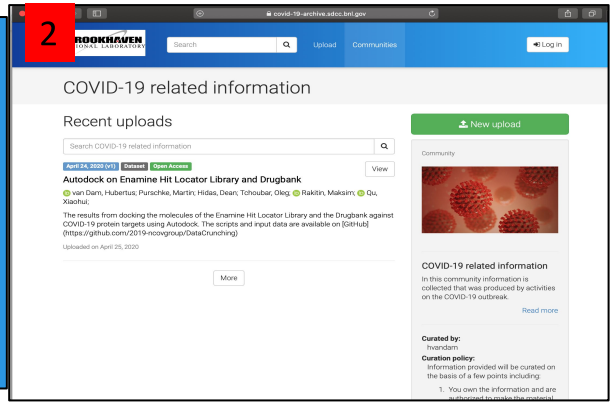
1. The selected researches will be able to use their institution's (ANL, ORNL ,.., BNL) login and passwords to authenticate to the system

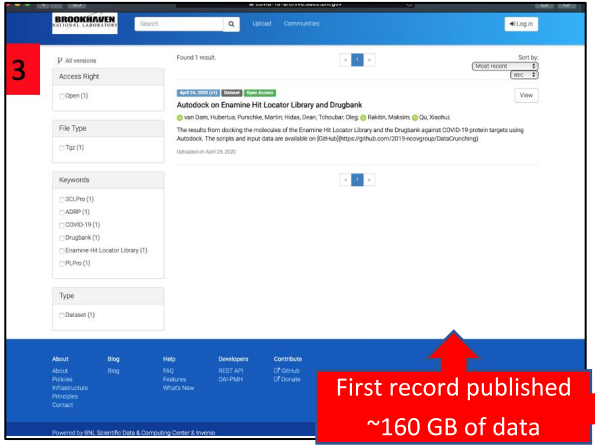
2. A *community* can be created to collect and curate topic/theme centric aggregation of documents

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

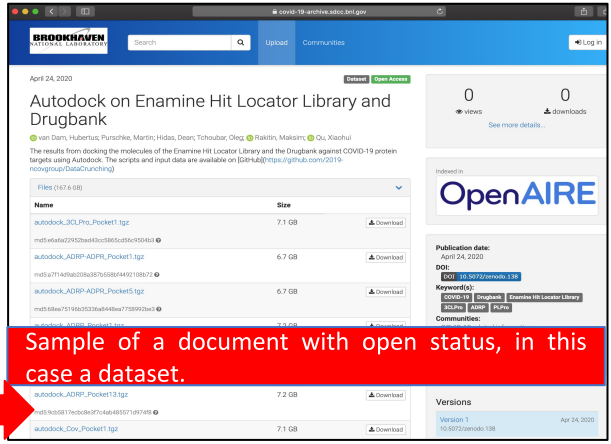
Zenodo will be migrated to invenioRDM







First record published
~160 GB of data



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

Why InvenioRDM?

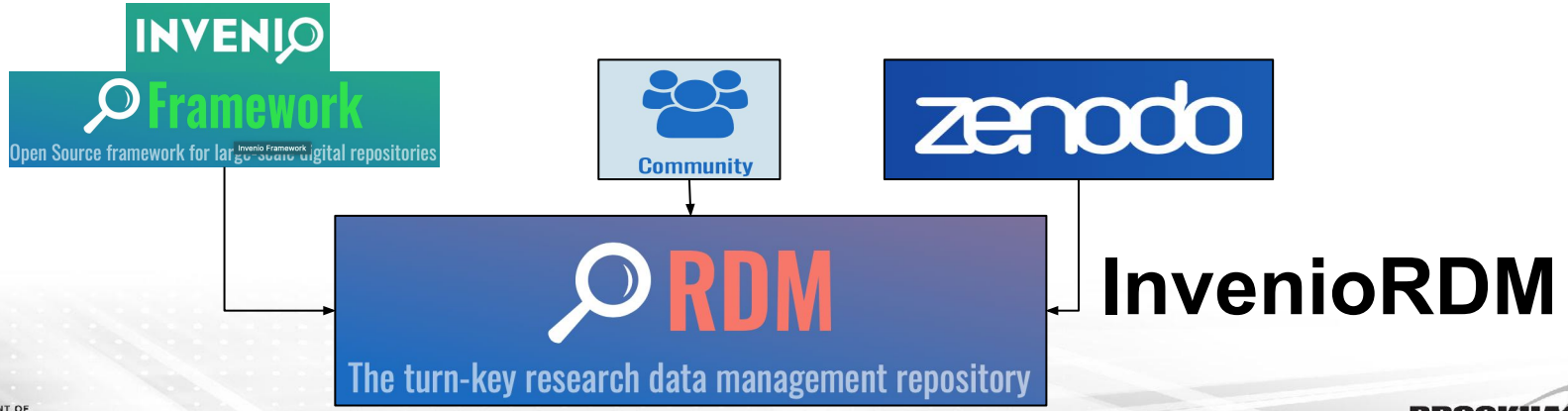
Different institutions were looking for a solution for a RDM for their communities

Zenodo was seen as a model to create their local RDM. However, while Zenodo code is released as open source it is not designed to be deployed outside of CERN

Other institutions tried to build a RDM using Invenio 3 framework. However, Invenio 3 ***is a code library*** used to build an application from scratch (e.g CERN Open Data or Zenodo)

Many organizations tried to share and reuse code modifications with not an easy portability

These interested multidisciplinary institutions gathered to create a collaborative open source (invenioRDM project) and grow a sustainable community. This project will provide a platform for institutions to be able to install their own RDM.



BNL has been active partner of invenioRDM project since its creation

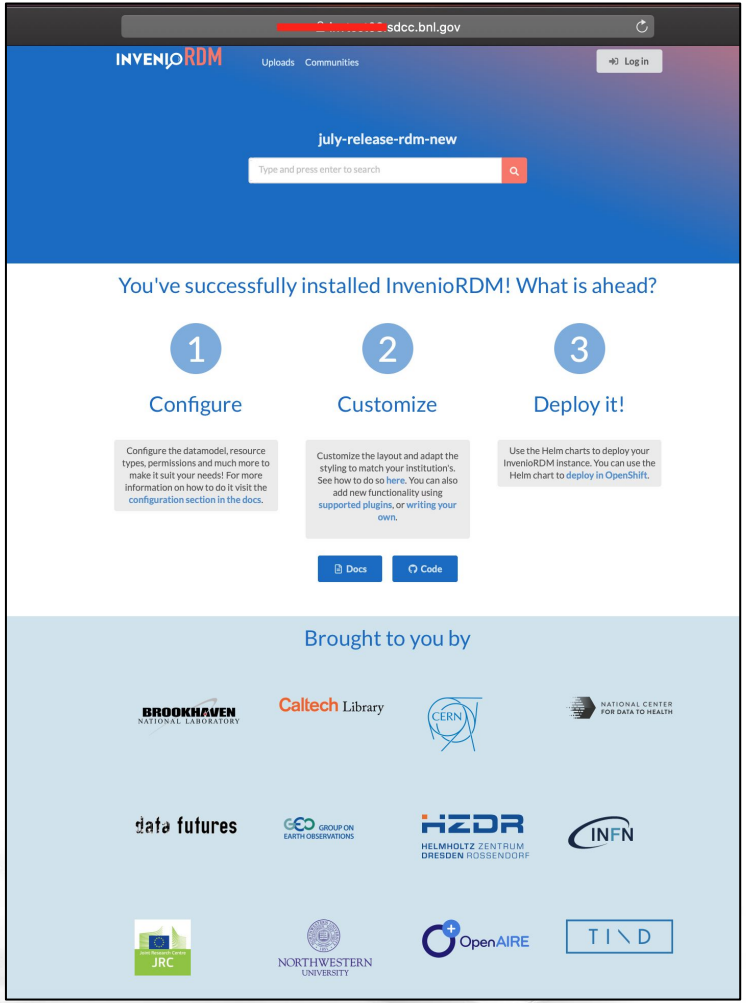
What is InvenioRDM?

InvenioRDM ***is an application*** built using invenio 3 aiming to be:

- A turn-key research data management repository: minimize the amount of developing work and code support to customize the RDM to the institution's needs

- Community supported: it is envisioned that a community of research institutions, private companies and individuals will be growing and help sustain this platform

It is planned that Zenodo will be migrated to InvenioRDM once it is released



InvenioRDM: Benefits and Features

Benefits

- **Research safely shared:** Permits share and preserve records with collaborators
- **Communities:** Allows user the creation and management of its community (e.g. journal, project, workshop).
- **Implements FAIR like policies for data deposited:** is achieved by a collection of a robust metadata in conjunction of an open API and powerful search index.
- **DOI persistent Identifier:** is available for citation and compliance with data sharing requirements
- **Simplicity:** Turn-key research data management platform can be installed in the local environment or by a service provider

Features

- **Class UX:** enhanced with user experience in mind
  - End-users, curators, sys admins and developers
- **Repository Profiles:** Comes with pre-configured repositories, Institutional Repositories (IRs), Research Data Management repository (RDM) and domain-specific repositories for health and biomedical sciences
- **Other features includes Resilience, Scalable and institutional integration**

InvenioRDM collaboration

Github is used to host:

Repository: https://github.com/inveniosoftware/invenio-app-rdm

Documentation: https://invenio-app-rdm.readthedocs.io/en/lates

Effort is coordinated by project boards:

- Priorities definition and its documentation
- Allows to identify or trace issues of the monthly developing sprints
- Example for this month https://github.com/orgs/inveniosoftware/projects/47

Invenio Request For Comments (RFC)

- Coordinate the design process
- Creates consensus with the parties involved
- Helps document invenioRDM development https://github.com/inveniosoftware/rfcs

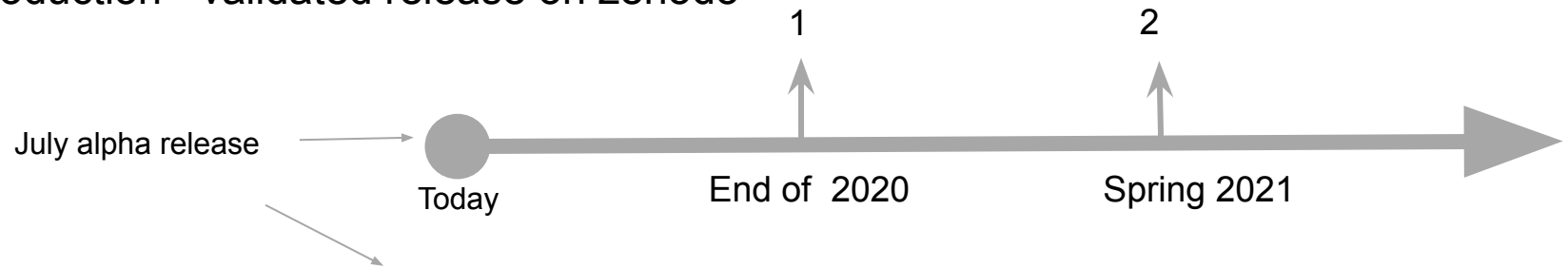
Brought to you by

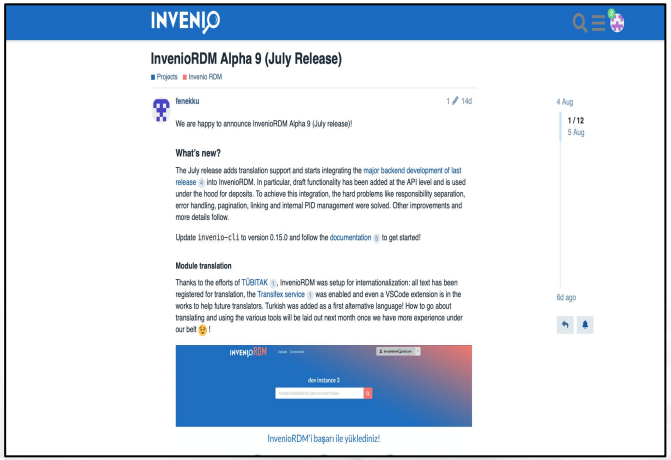
Collaborators

InvenioRDM (status)

Milestones

1) First minimal release - bare bones
2) Production - validated release on zenodo





invenioRDM at BNL

Electron Ion Collider (EIC), the DOE covid-19-archive Therapeutics and PHENIX interested in using it

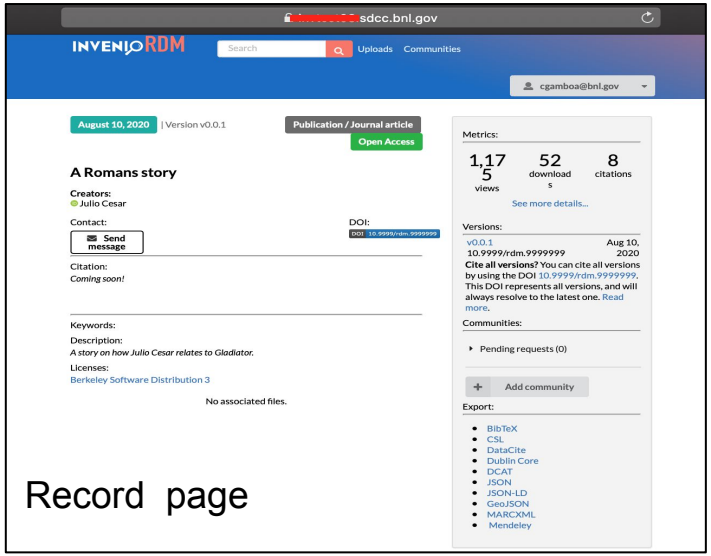
- Zenodo based deployments will be migrated to invenioRDM once released

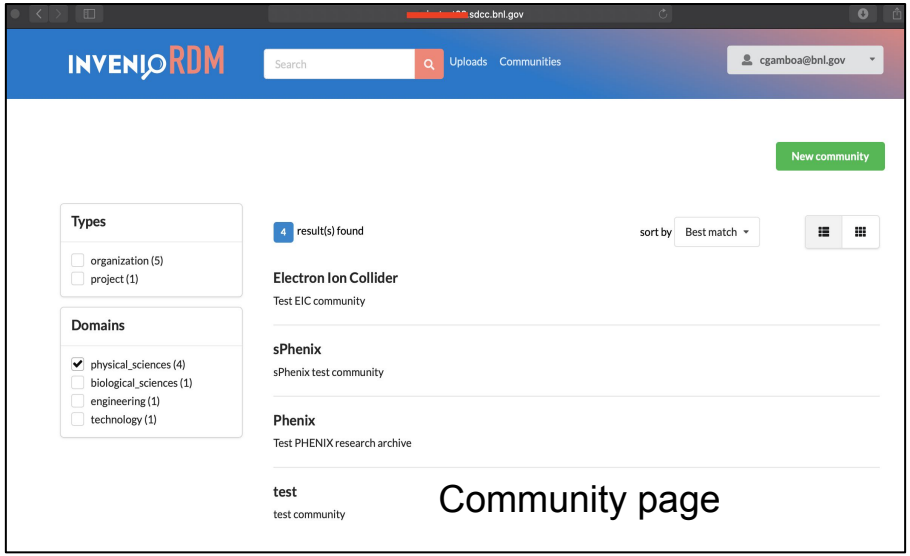
Other invenio based repositories (sPhenix) on development at BNL are looking at invenioRDM

OSTI (Office of Scientific and Technical Information) DOI integration with invenioRDM begun:

- DOE mandates that work funded by or performed at DOE labs should be registered with OSTI
  - OSTI DOIs are free
- Invenio-based repositories register DOIs directly through DataCite
- Project initiated by SULI summer student
  https://indico.bnl.gov/event/9091/contributions/40194/attachments/29865/46597/InvenioRDM_UI_and_DOI_Enhancement_Summary.pdf

invenioRDM as is (July alpha release)





Customization of different User Interfaces (repository landing, record and community pages) will need to be done to reflect BNL's communities needs.

To facilitate the customization of these components invenioRDM uses:

- Semantic UI: Web design framework for responsive layouts with intuitive/human-friendly HTML
- ReactJS: JavaScript framework for real-time rendering/updating of pages

invenioRDM at BNL near future work

A invenioRDM (“bare bones release”) testbed installation is expected to be available for interested BNL communities

- Definition of the record model
- Possibility to identify communities that can use a multi-disciplinary research repository

Expected to start transition of BNL invenio repositories to invenioRDM in spring 2021

Summary

An overview of invenio related projects at BNL was presented

SDCC has a solid expertise in hosting and operating digital document repositories based on Invenio

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

BNL scientific communities will use invenioRDM to host their digital research records

References

Invenio, https://invenio-software.org

InvenioRDM, https://invenio-software.org/products/rdm/

invenioRDM user docs, https://inveniordm.docs.cern.ch

DOE OSTI, https://www.osti.gov/

DataCite, https://datacite.org

Backup slides

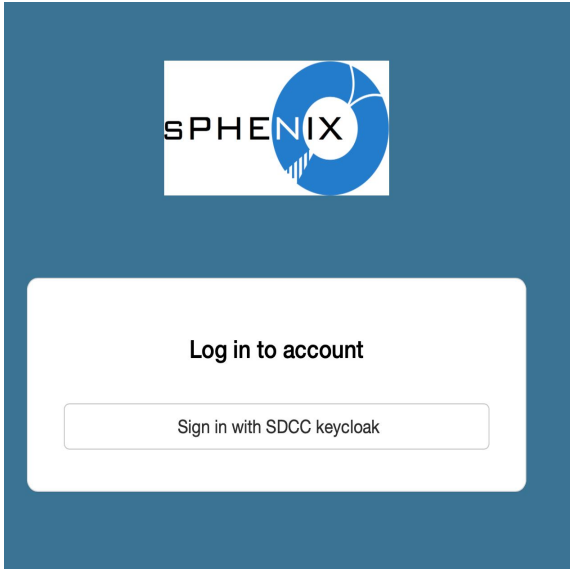
FAIR data policy

To be Findable	To be Accessible	To be Interoperable	To be Reusable
F1: (meta)data are assigned a globally unique and persistent identifier. A DOI is issued to every published record on InvenioRDM.	**A1: (meta)data are retrievable by their identifier using a standardized communications protocol** Metadata for individual records as well as record collections are harvestable using the OAI-PMH protocol by the record identifier and the collection name. Metadata is also retrievable through the public REST API.	**I1: (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.** InvenioRDM uses JSON Schema as internal representation of metadata and offers export to other popular formats such as Dublin Core or MARC-XML.	**R1: (meta)data are richly described with a plurality of accurate and relevant attributes** Each record contains a minimum of DataCite's mandatory terms, with optionally additional DataCite recommended terms and InvenioRDM's enrichments.
F2: data are described with rich metadata (defined by R1 below). InvenioRDM's metadata is compliant with DataCite's Metadata Schema minimum and recommended terms, with a few additional enrichments.	**A1.1: the protocol is open, free, and universally implementable** See point A1. OAI-PMH and REST are open, free and universal protocols for information retrieval on the web.	**I2: (meta)data use vocabularies that follow FAIR principles** For certain terms we refer to open, external vocabularies, e.g.: license (Open Definition), funders (FundRef) and grants (OpenAIRE).	**R1.1: (meta)data are released with a clear and accessible data usage license** License is one of the mandatory terms in InvenioRDM's metadata, and is referring to an Open Definition license. Data downloaded by the users is subject to the license specified in the metadata by the uploader.
F3: metadata clearly and explicitly include the identifier of the data it describes. The DOI is a top-level and a mandatory field in the metadata of each record.	**A1.2: the protocol allows for an authentication and authorization procedure, where necessary** Metadata are publicly accessible and licensed under public domain. No authorization is ever necessary to retrieve it.	**I3: (meta)data include qualified references to other (meta)data** Each referenced external piece of metadata is qualified by a resolvable URL.	**R1.2: (meta)data are associated with detailed provenance** All data and metadata uploaded is traceable to a registered InvenioRDM user. Metadata can optionally describe the original authors of the published work.
F4: (meta)data are registered or indexed in a searchable resource Metadata of each record is indexed and searchable directly in InvenioRDM's search engine immediately after publishing. Metadata of each record is sent to DataCite servers during DOI registration and indexed there.	**A2: metadata are accessible, even when the data are no longer available.** Data and metadata will be retained for the lifetime of the repository. Metadata are stored in high-availability database servers which are separate to the data itself. (note: recommendations for local implementations should be established here)		**R1.3: (meta)data meet domain-relevant community standards** InvenioRDM is not a domain-specific repository, yet through compliance with DataCite's Metadata Schema, metadata meets one of the broadest cross-domain standards available.

Courtesy Sara Gonzales

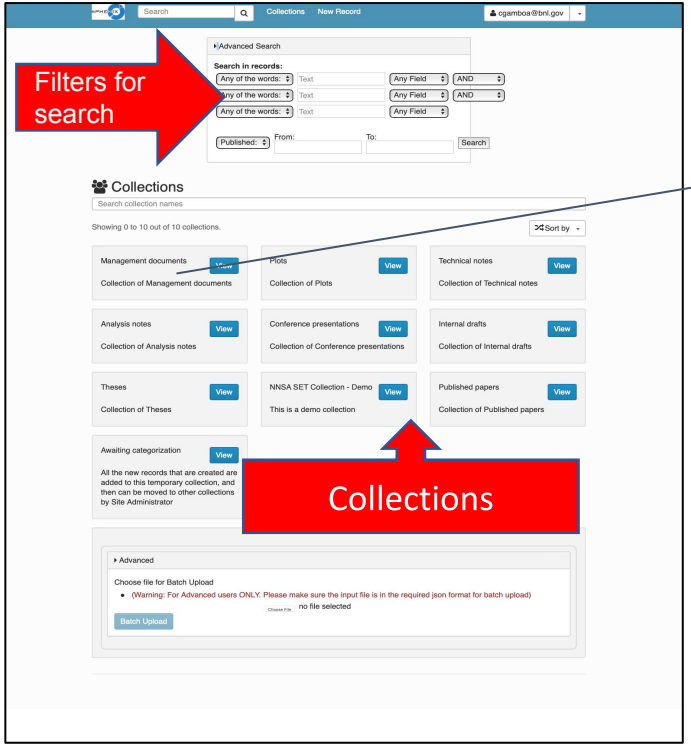
sPHENIX Document store

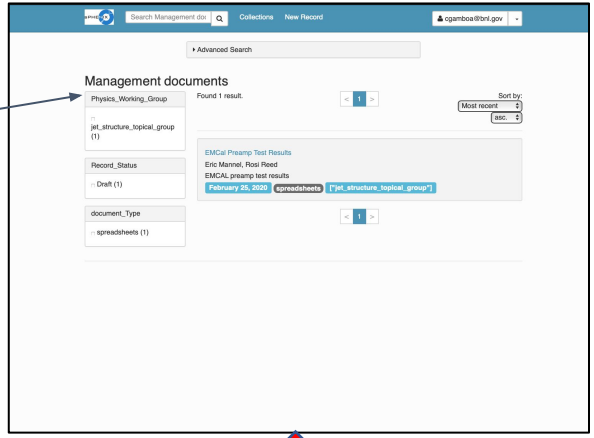
Invenio custom application



Integrated with SSO using SDCC keycloak infrastructure

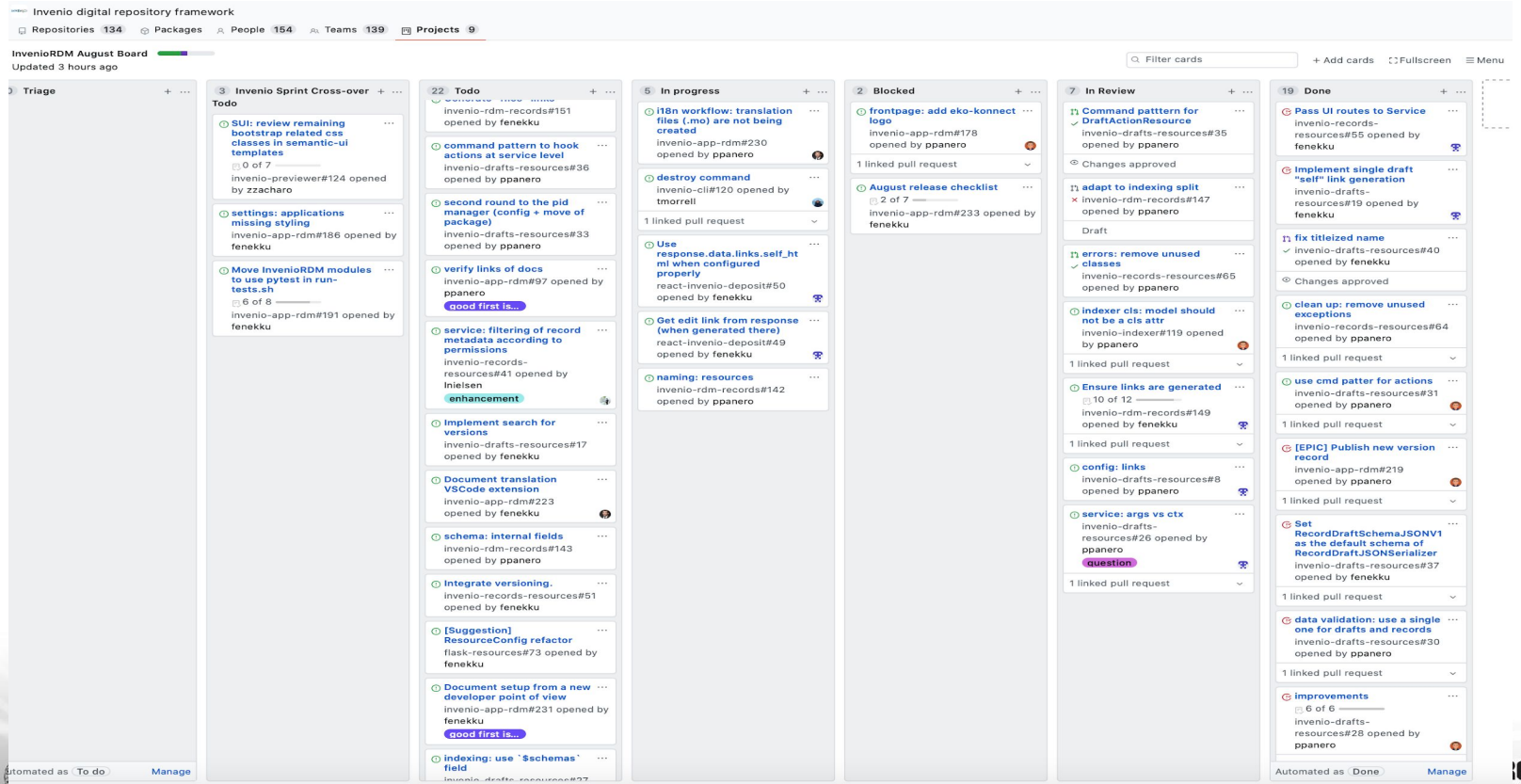
Beta release





InvenioRDM: August project board

https://github.com/orgs/inveniosoftware/projects/47



invenioRDM at BNL Integration with CILogon

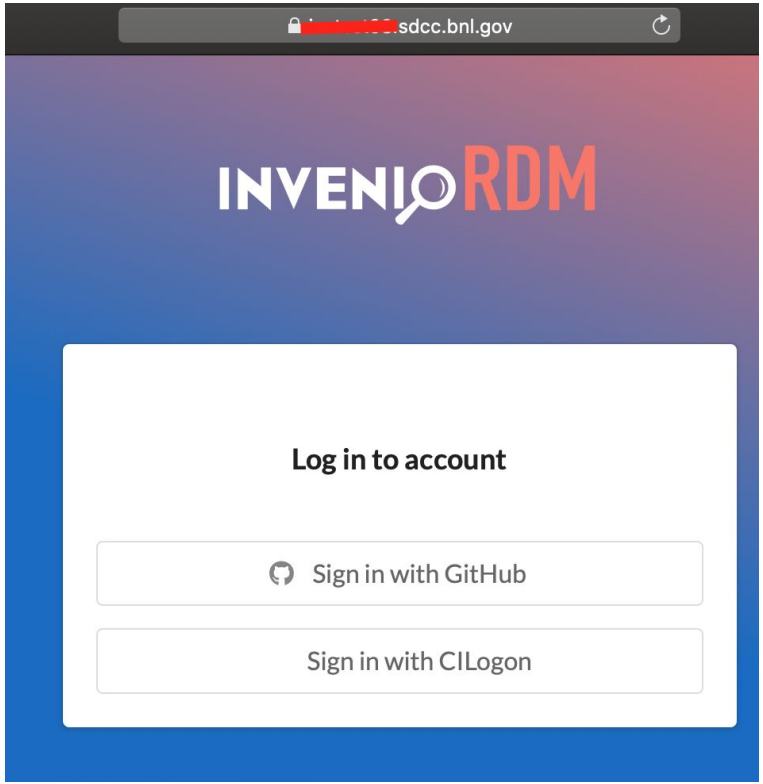
Out of box invenioRDM is integrated with Github OAuth authentication

A CILogon/Incommon OAuth plugging is being customized for invenioRDM

Module: invenio-oauthclient-1.4.0a1

CiLogon.py plugging locally supported

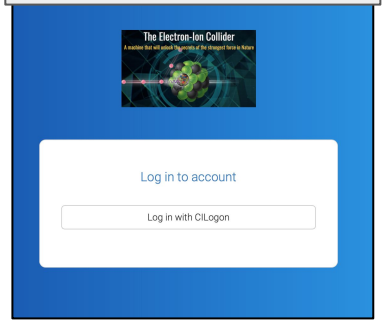
First known use case of CiLogon/Incommon for invenioRDM



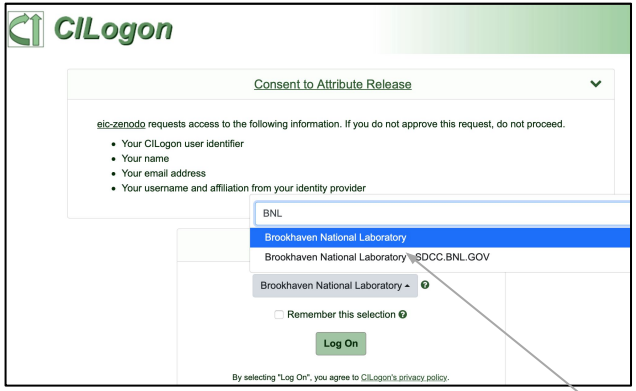
Invenio 3 authentication flow using Incommong Federated Id/COManage

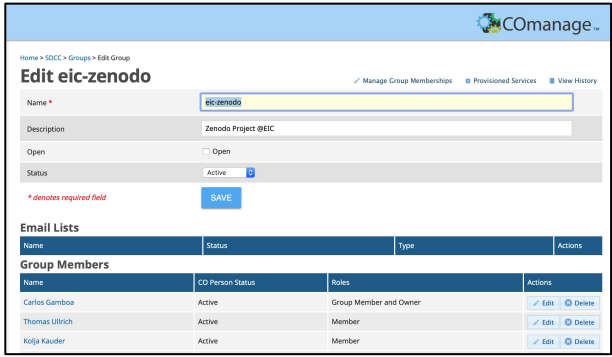
1

Zenodo login page



2



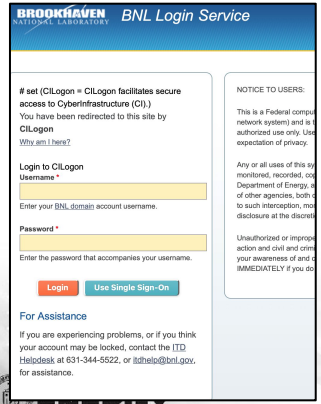


Incommon Fed ID/COManage integrated and used to restrict write access to the Zenodo instance

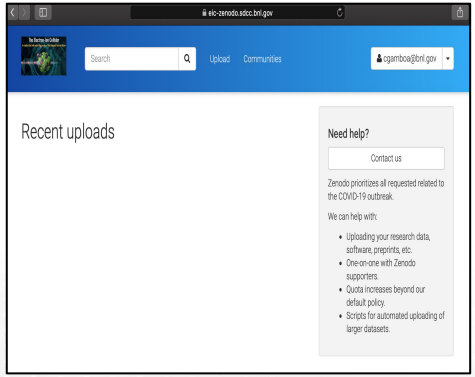
For now users allowed to login must:

- Belong to the COmage eic-zenodo group
- Use BNL Incommon IDPs to login

3



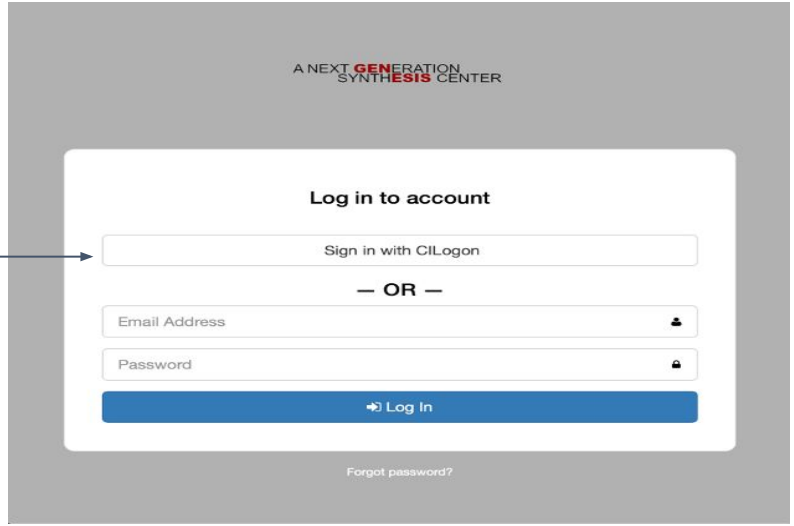
4



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

Contacts

SDCC, Eric Lancon, Tony Wong

EIC Zenodo, Jerome Lauret

Invenio, CERN, Jose Benito Gonzalez, Lars Holm Nielsen

sPHENIX, John Haggerty, Chris Pikenburg

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

GENESIS, Line Pouchard

COVID-19-ARCHIVE, Kerstin Kleese Van Dam

CSI (SET, GENESIS, sPhenix), Developer, Uma Ganapathy