Invenio Based Digital Repositories at BNL

Carlos Fernando Gamboa (cgamboa@bnl.gov)

Scientific Data and Computer Center (SDCC), BNL

HEPiX Autumn 2020 online workshop, October 12th-16th 2020

BROOKHAVEN NATIONAL LABORATORY

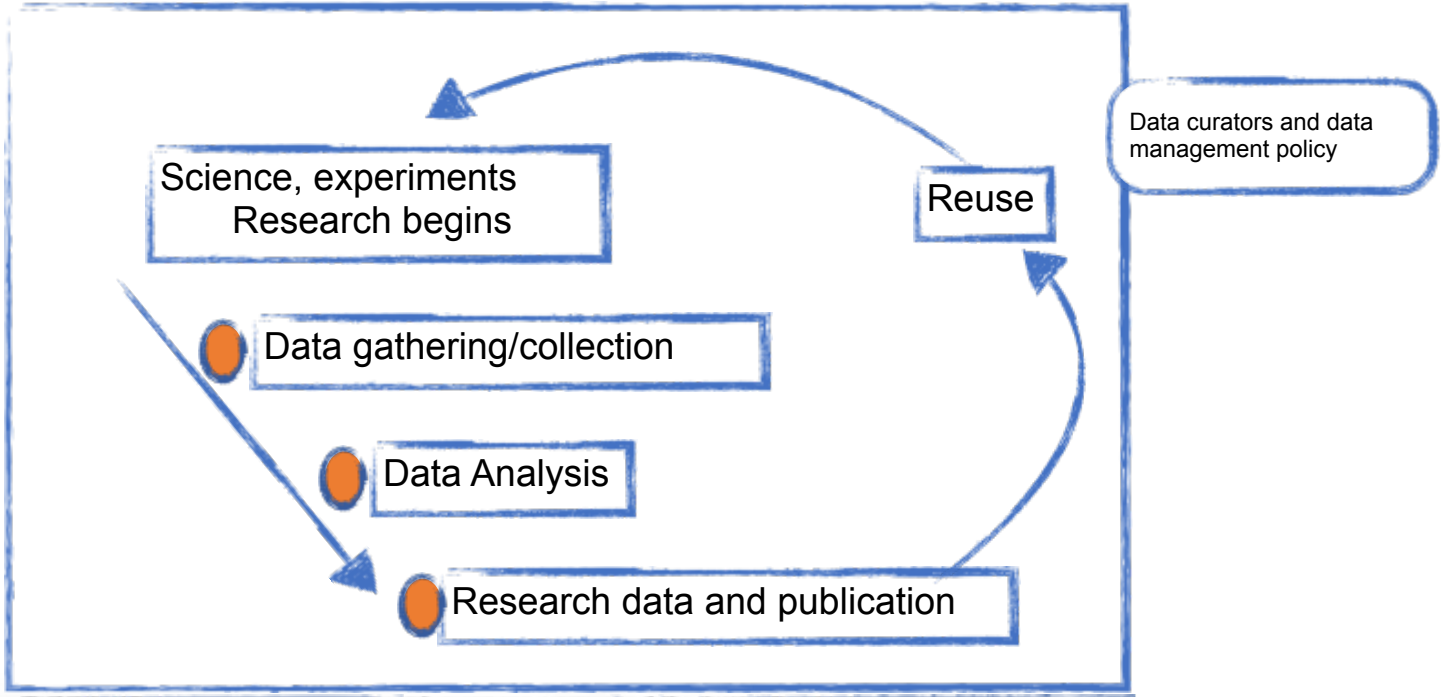
U.S. DEPARTMENT OF ENERGY

BROOKHAVEN SCIENCE ASSOCIATES

A scientific data workflow

- **Research data work**

  Research Digital Management repositories are key elements of infrastructure



Institutional infrastructure supports the repositories

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 hosted at CERN)

https://zenodo.org

zenodo | Search | Upload | Communities | Log in | Sign up

COVID-19 related communities

Need help uploading? Contact us

Chicago COVID-19 Response

Browse | New upload

This repository community collects research outputs and information objects relevant to the COVID-19 / SARS-CoV-2 efforts in Chicago. Users are encouraged to upload their research objects in this collection to facilitate sharing and discovery of information. Although Open Access articles and...

Curated by: saragon

Featured uploads related to COVID-19

Want your dataset featured? Contact us

August 5, 2020 | Dataset | Open Access
Japanese Sample Tweets, COVID-19 Keywords and Emotions from 2020-01-01 to 2020-06-30 (88,495,817 tweets and 47,539,139 retweets)
Yoshida, Mitsuo
Data Tweets_YYYY-MM.tsv.gz: The first column is the tweet id, the second column is the date and time (JST) when the tweet was posted, the third column is the tweet id of the mention destination,...
Uploaded on August 14, 2020

August 7, 2020 (v1) | Software | Open Access
jadebc/covid19-infections: Nature Communications final version
Benjamin-Chung, Jade; Wu, Sean
This study estimated the total number of SARS-CoV-2 infections in the U.S. from February 28 to April 18, 2020 correcting for incomplete testing and imperfect test accuracy. This study used a...
Uploaded on August 14, 2020

August 6, 2020 (v4) | Dataset | Open Access
Geostatistical Analysis of SARS-CoV-2 Positive Cases in the United States
Peter K. Rogan
Geostatistics analyzes and predicts the values associated with spatial or spatial-temporal phenomena. It incorporates the spatial (and in some cases temporal) coordinates of the data within the...
Uploaded on August 14, 2020

Browse COVID-19 related research

Recent uploads

August 14, 2020 (v8) | Dataset | Open Access
List of studies on Covid-19 identified in PubMed and excluded from our systematic review
Carolina Riveros; Olivier Pierre; Leopold Fezeu; Isabelle Boutron; Anna Chaimani; Declan Devane; Joerg J Meerpohl; David Tovey; Asbjørn Hróbjartsson; Philippe Ravaud
In the course of our PubMed searches we identified and screened the title and abstract of a number of studies on Covid-19 that were finally excluded from our systematic review. This file will be updated regularly.
Uploaded on August 14, 2020
7 more version(s) exist for this record

View

August 14, 2020 (v1) | Dataset | Open Access
Electron microscopy of SARS-CoV-2 particles - Dataset 03

View

Need help?

Contact us

Zenodo prioritizes all requested related to the COVID-19 outbreak.

We can help with:

- Uploading your research data, software, preprints, etc.
- One-on-one with Zenodo supporters.
- Quota increases beyond our default policy.
- Scripts for automated uploading of larger datasets.

Repository landing page

zenodo | Search | Upload | Communities | Log in | Sign up

A

August 14, 2020 | Dataset | Open Access

List of studies on Covid-19 identified in PubMed and excluded from our systematic review

Carolina Riveros; Olivier Pierre; Leopold Fezeu; Isabelle Boutron; Anna Chaimani; Declan Devane; Joerg J Meerpohl; David Tovey; Asbjørn Hróbjartsson; Philippe Ravaud

In the course of our PubMed searches we identified and screened the title and abstract of a number of studies on Covid-19 that were finally excluded from our systematic review.

This file will be updated regularly.

Files (3.7 MB)

Name	Size	
Exclus Pubmed Zenodo 14 août.xlsx	3.7 MB	Download

md5:03a7a3eac3aef0eddd92fad739d05c0

Beta Citations 0

Show only: Literature (0) Dataset (0) Software (0) Unknown (0) Citations to this version

Search

No citations.

527 views | 168 downloads

See more details...

Indexed in OpenAIRE

F

Publication date: August 14, 2020

DOI: DOI 10.5281/zenodo.3985790

Keyword(s): Covid-19 | 2019-nCoV | SARS-CoV-2

Communities: Living mapping and living network meta-analysis of Covid-19 studies Zenodo

License (for files): Creative Commons Attribution 4.0 International

R

Versions

Version 8 — 10.5281/zenodo.3985790 — Aug 14, 2020

Version 7 — 10.5281/zenodo.3975888 — Aug 7, 2020

Version 6 — 10.5281/zenodo.3968490 — Jul 31, 2020

Share

Cite as

Carolina Riveros, Olivier Pierre, Leopold Fezeu, Isabelle Boutron, Anna Chaimani, Declan Devane, ... Philippe Ravaud. (2020). List of studies on Covid-19 identified in PubMed and excluded from our systematic review [Data set]. Zenodo. http://doi.org/10.5281/zenodo.3985790

Start typing a citation style...

I

Export

BibTeX CSL DataCite Dublin Core DCAT JSON JSON-LD GeoJSON MARCXML Mendeley

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 3 at BNL

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

- After evaluation and testing Zenodo was implemented as a R&Ds testbed used by different BNL communities
- By interacting with the Invenio framework and testing its capabilities, these communities built their own digital repositories to meet their specific needs

Most recently new BNL scientific communities (Nuclear and Particle Physics) 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:

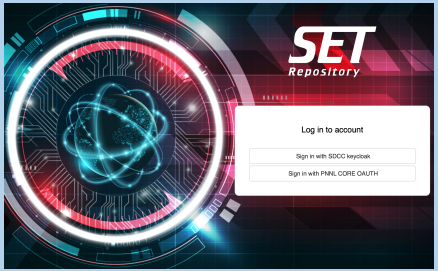
National Nuclear Security Administration

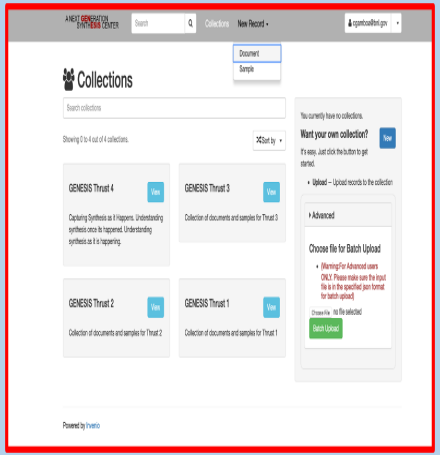
Application **SET,** Smuggling Detection and Deterrence Science and Engineering Team

Materials Science community

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

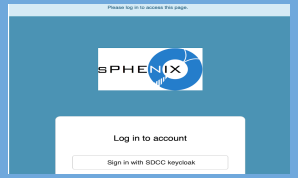
Repositories in operation

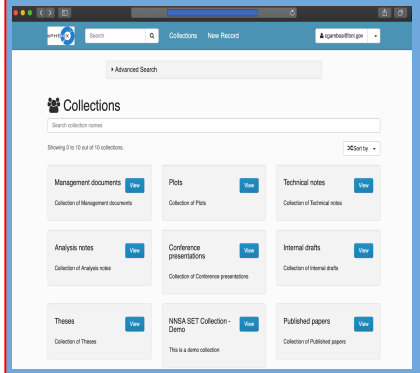




Repository in development and testing

Nuclear and Particle Physics (NPP)

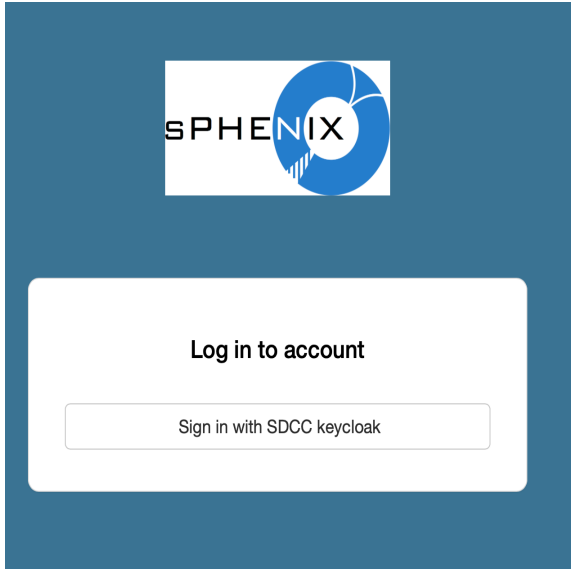


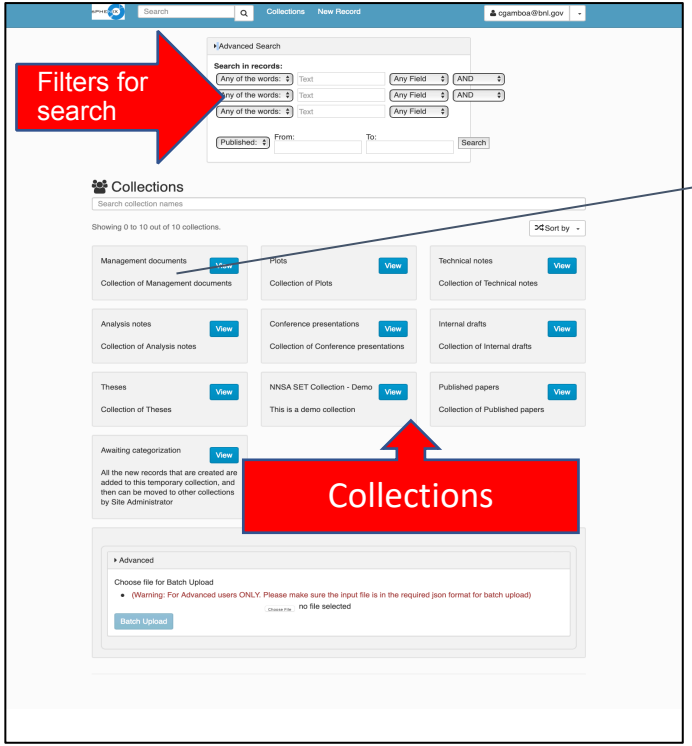


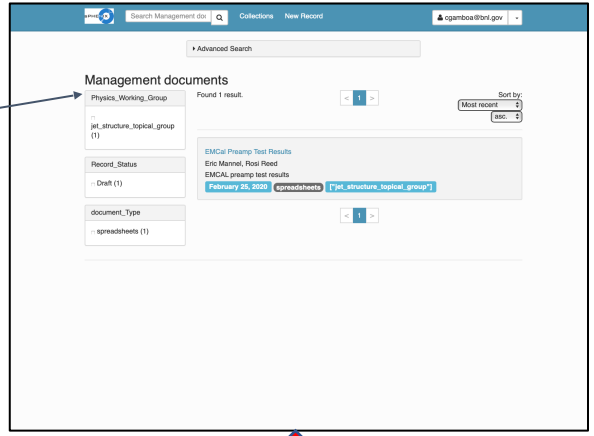
SDCC supports infrastructure for Invenio based applications, along with customized network, storage and Authentication infrastructure enabled to host services (production, testing and developing)

sPHENIX Document store
Invenio custom application

Beta release





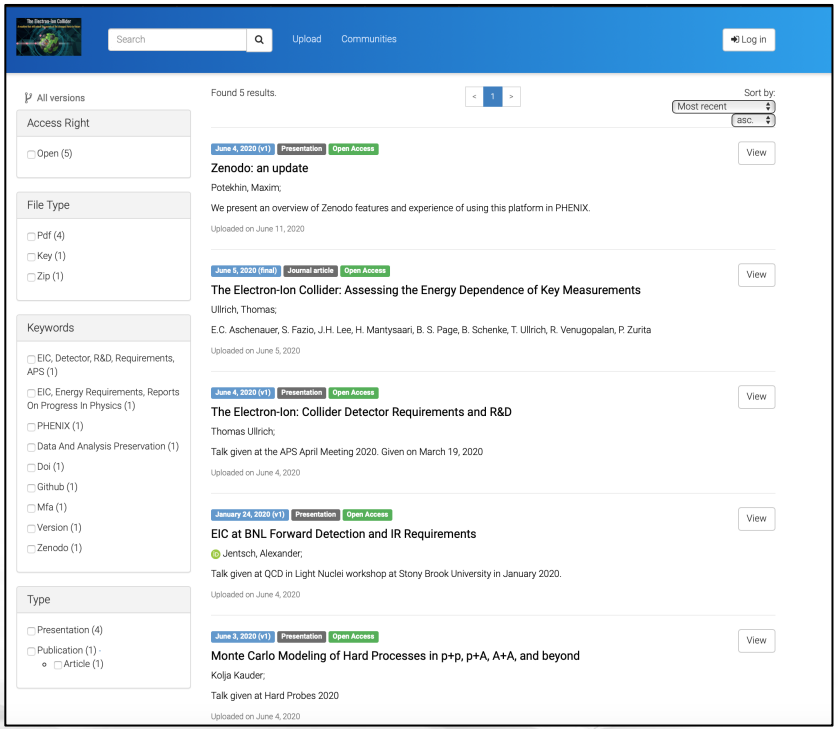


Integrated with SSO using SDCC keycloak infrastructure

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

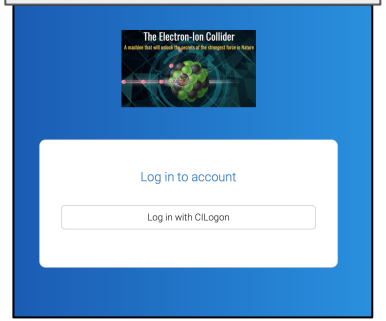


A production EIC-Zenodo instance recently commissioned

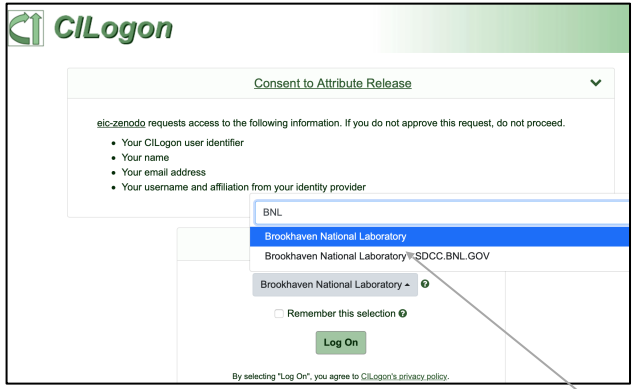
Invenio 3 authentication flow using Incommon Federated Id/COManage

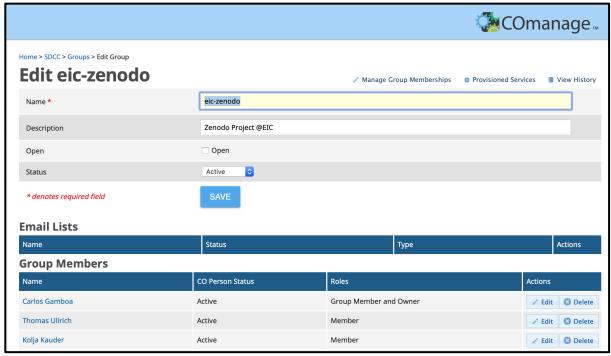
1

Zenodo login page



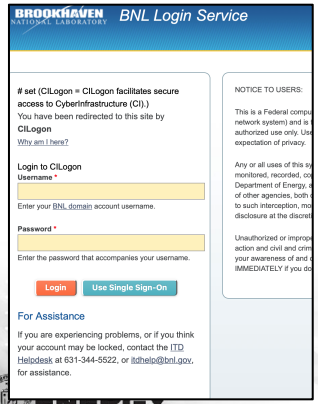
2



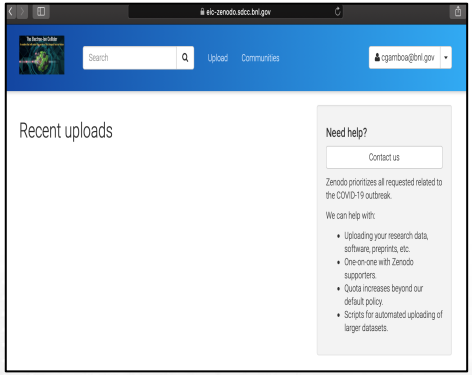


Incommon Fed ID/COManage integrated and used to restrict write access to the Zenodo instance

3



4



For now users allowed to login must:

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

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

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

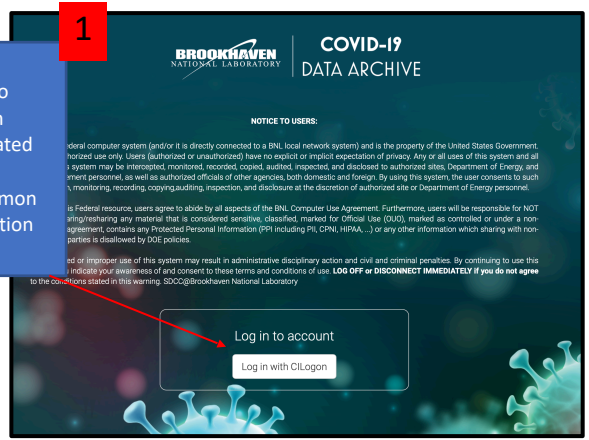
3 General users will be able to download data (files) from the repository based on **document status**:

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

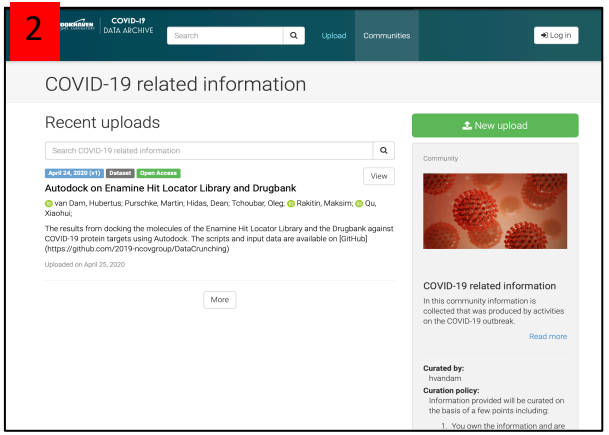
Zenodo will be migrated to invenioRDM

Zenodo system integrated with Incommon federation

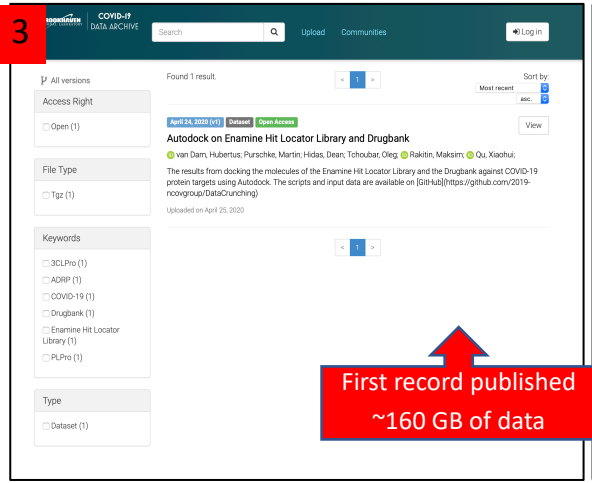
1



2

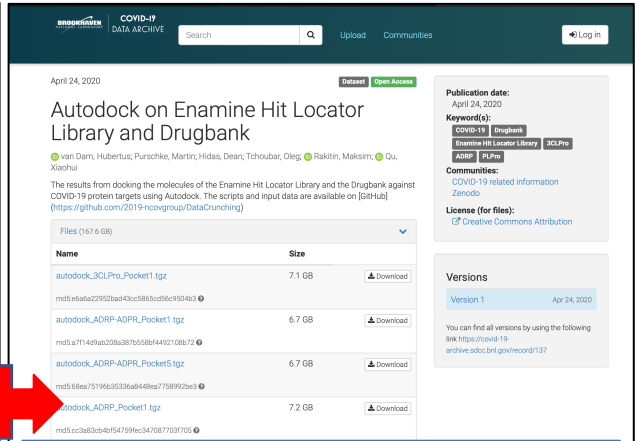


3



First record published

~160 GB of data



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

Why InvenioRDM?

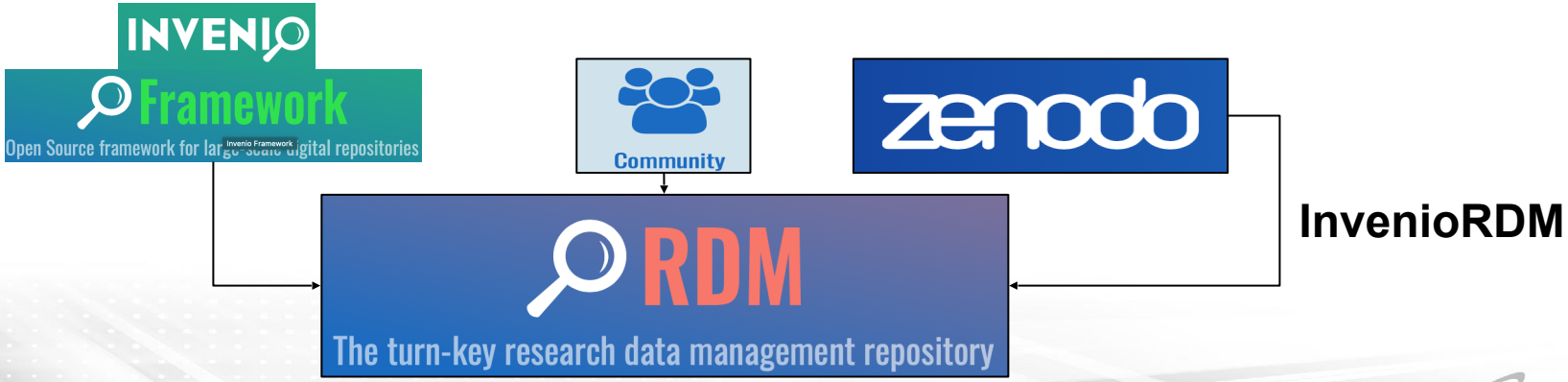
Different institutions were looking for a solution for a RDM for their communities

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

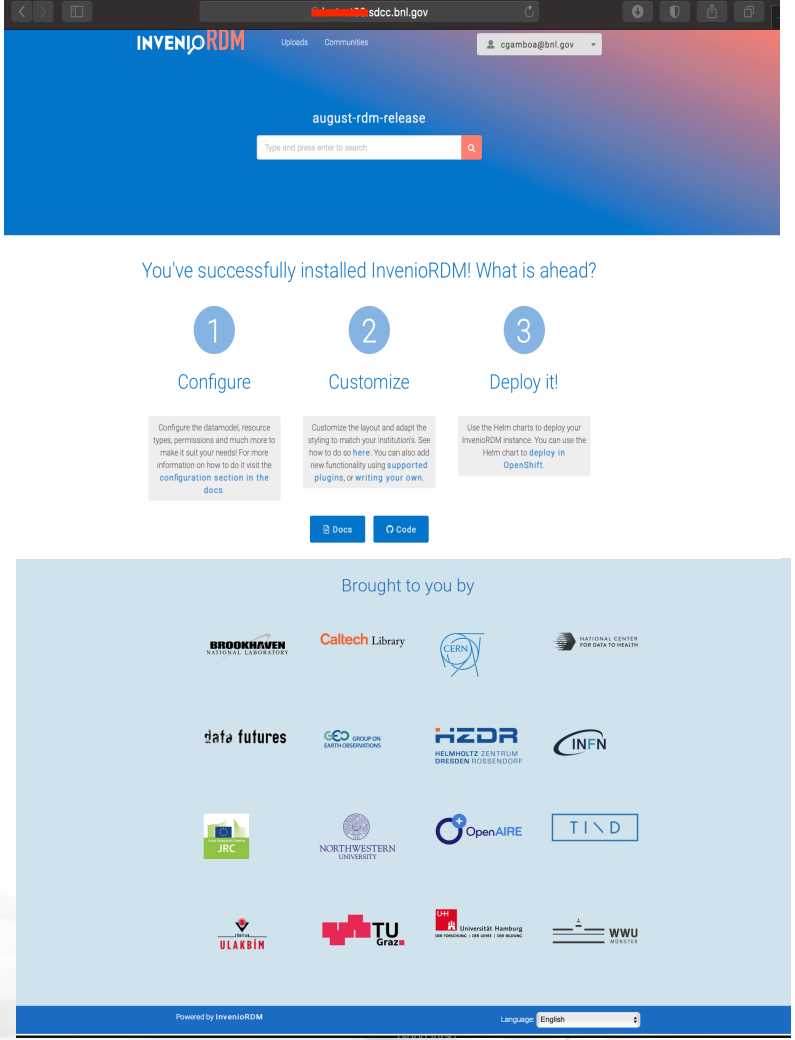
What is InvenioRDM?

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

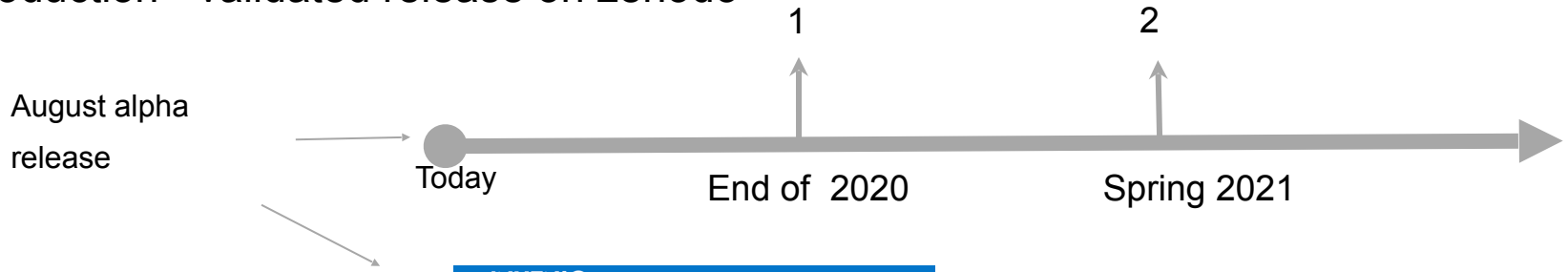
Brought to you by

Collaborators

InvenioRDM (status)

Milestones

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



INVENIO

InvenioRDM Alpha 10 (August Release)

Projects · Invenio RDM

fenekku 1 Sep

We are glad to announce InvenioRDM Alpha 10 (August release)!

What's new?

With less resources than usual, we made some limited changes this month. However, three areas *were* tackled.

- The concurrent Invenio Sprint brought the final Semantic-UI updates across all modules! If something does not show up right in your default RDM instance, then it's a legitimate bug now 🐞.
- These UI updates complement the updates to Search. The record search page now uses the new API endpoint from the last release. Customization of the search results is back (with documentation forthcoming). Different search pages are easier to setup.
- More library improvements across the board were merged.

Update `invenio-cli` to version 0.16.0 and follow the updated documentation 9 to get started.

Semantic-UI transition wrap-up

invenioRDM at BNL

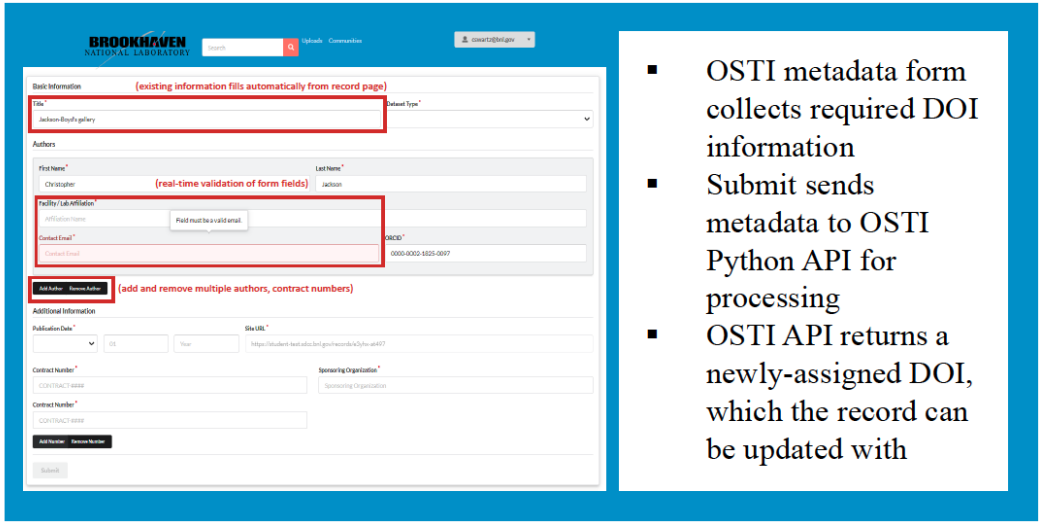
Electron Ion Collider (EIC), the DOE covid-19-archive Therapeutics and PHENIX interested in using it

- 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:

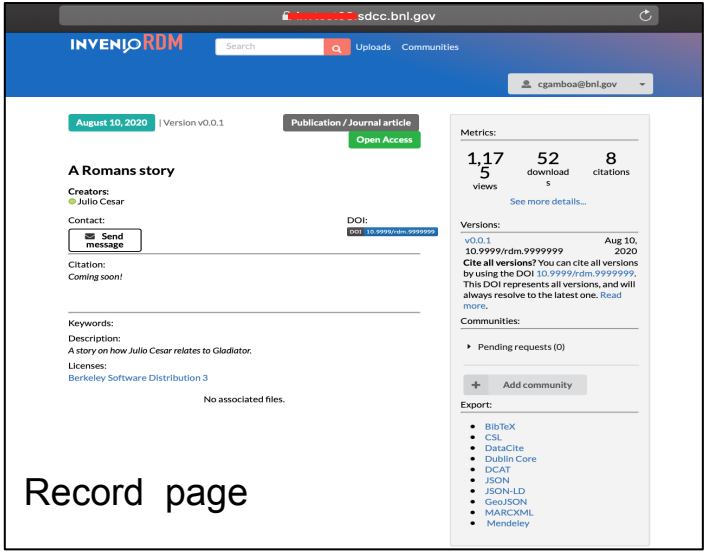
- DOE mandates that work funded by or performed at DOE labs should be registered with OSTI
  - OSTI DOIs are free
- Invenio-based repositories register DOIs directly through DataCite



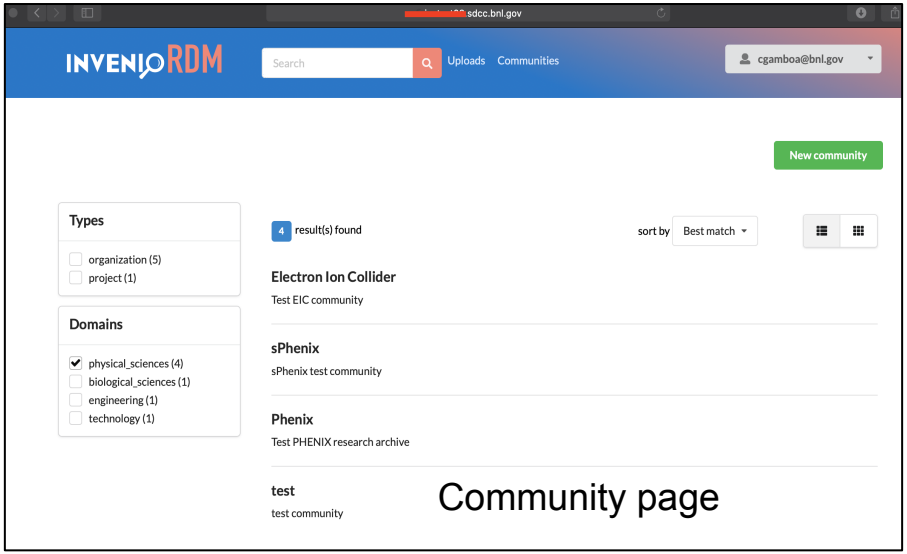
- OSTI metadata form collects required DOI information
- Submit sends metadata to OSTI Python API for processing
- OSTI API returns a newly-assigned DOI, which the record can be updated with

Project started by Cole Swartz
a Science Undergraduate Laboratory Internship (SULI) summer 2020 student

invenioRDM as is (August alpha release)



Record page



Community page

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: 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

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

SDCC is supporting digital repositories for BNL scientific communities and DOE in the US

BNL is a partner of the invenioRDM project and scientific communities will use it to host their digital research records

References

Invenio, https://invenio-software.org

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

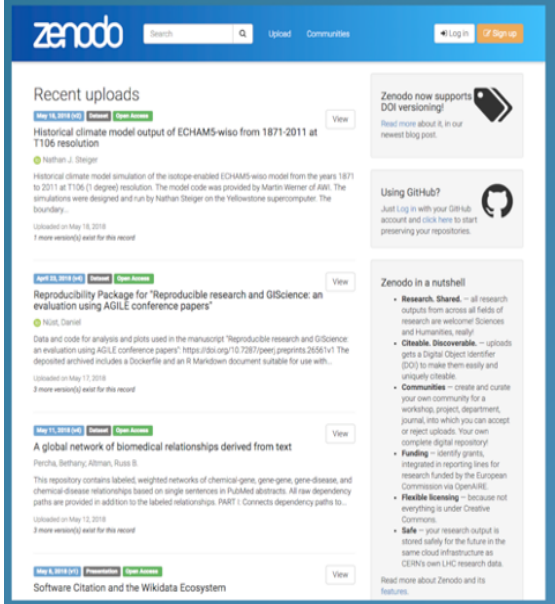
invenioRDM user docs, https://inveniordm.docs.cern.ch

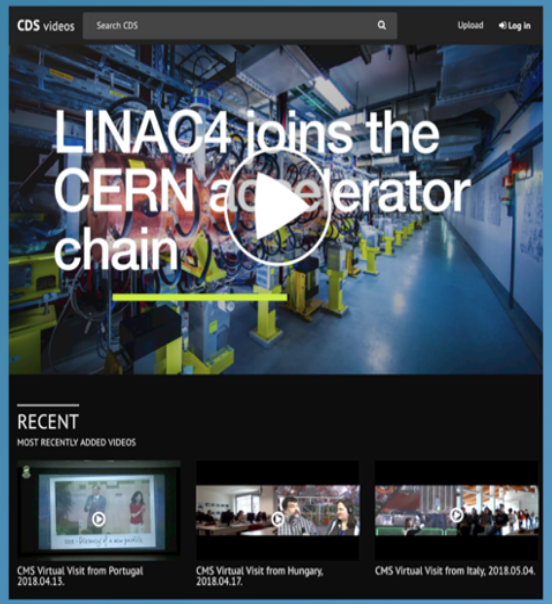
DOE OSTI, https://www.osti.gov/

DataCite, https://datacite.org

Backup slides

Examples of digital repositories on Invenio 3 hosted at CERN





Zenodo
~100 TBs / 2M records

CERN Video Platform
~35 TBs / >5000 videos

CERN Open Data Portal
1.7 PBs / >620K files

Courtesy Jose Benito Gonzalez

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

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