HELMHOLTZ PROJECT HERMES

HELMHOLTZ RICH METADATA SOFTWARE PUBLICATIONS

HERMES Kickoff Workshop 2021-11-12

INVENIO RDM

Powering open science and collaboration with Invenio

HERMES Kick-off Workshop

Northwestern University Invenio Team
November 12, 2021

Invenio software powers open science

CERN

zenodo

Free Open Source Software

Invenio is Free Open Source Software supported by a committed community of multidisciplinary institutions.

Code | Docs | Examples

Friendly and Responsive Community

Although Invenio was born at CERN, its community is growing bigger every day. Talk to the team now in our chatroom or forum.

Chatroom | Forum | Get Involved | Events

The "Safe bet"

Invenio community has been around for 20 years. Solid services have been built on top of it to ensure long-term confidence.

Live services | Products

How did this collaboration start (and what about Zenodo?!)

What motivated the InvenioRDM project?

- Some organizations tried to reuse the existing open source Zenodo source code
- Other orgs tried to use the Invenio Framework to build a RDM repository from scratch
- Several orgs tried to make the same modifications but had no easy way of sharing their changes

All these groups came together to create a collaborative open source project and grow a sustainable community.

Zenodo will also run on InvenioRDM by the end of the project period.

We're leveraging Invenio as a strong foundation. Here's why.

RDM platforms *are critical to help preserve and share research, enable reproducibility, and empower reuse of datasets, protocols, engagement or study materials, & a wide range of other research products.*

- **Research, shared.** Securely share and preserve data records and a wide range of research types with collaborators. Allows easy dissemination to the community.
- **Discoverable.** Leverages metadata standards and the powerful Elasticsearch full-text search engine retrieves, facets, sorts, and filters your searches with ease.
- **Scalable.** Invenio is fast. Designed to manage 100+ million records and petabytes of files. All data can be archived independently of the size.
- **Communities.** Create and curate your own community (e.g., workshop, project, lab, or journal).
- A robust community: Large team of developers & active open source community. A SAAS-model for service via TIND (CERN spinoff). Invenio is widely used by many organizations & underlying technology (Python, Flask) widely supported.
- **Next-Generation:** With InvenioRDM, any organization can launch a turn-key open source next-generation repository platform with world-class features to support open and FAIR science. http://ngr.coar-repositories.org/
- **Get credit & be cited.** Get a DOI to make records easily and uniquely citable. Pre-formatted citation text makes it easy to cite your work and be cited. Contributor roles allow you to recognize the whole team.
- **Metrics.** Industry standard usage statistics for record pages with all tracking completely anonymized.
- **FAIR.** Advanced features to make your research Findable, Accessible, Interoperable, & Reusable.
- **Compliance-friendly.** Comply with data sharing mandates* and acknowledge your funders.
- **Easy.** Turn-key research data management platform & index can be easily deployed in the local environment by your team or by a service provider, such as TIND. Customize the look and feel to your local environment.

RDM

***Draft NIH Policy for Data Management and Sharing** https://osp.od.nih.gov/wp-content/uploads/Draft_NIH_Policy_Data_Management_and_Sharing.pdf

The InvenioRDM project has two goals:

Repository Platform

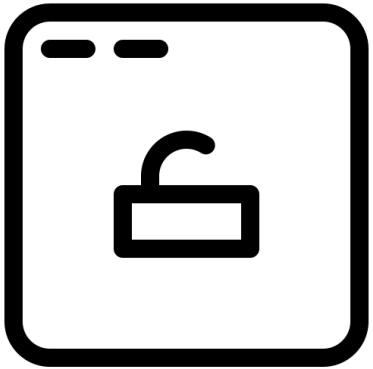
Build a turn-key research data management (RDM) repository platform based on **Invenio Framework** and **Zenodo**.

Community

Grow a community of research institutions, private companies and individuals to sustain the platform going forward.

The platform

A few highlights...

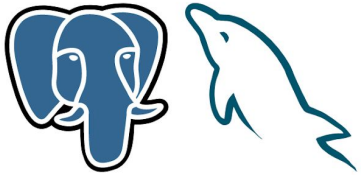


https://thenounproject.com/

InvenioRDM stack

Elasticsearch

Elasticsearch is an extremely fast JSON-native distributed search engine supporting anything from full-text to geospatial queries.



PostgreSQL or MySQL

PostgreSQL and **MySQL** are powerful relational databases with JSON-support as well as a strong reputation for reliability, robustness, and performance.

Python/Flask

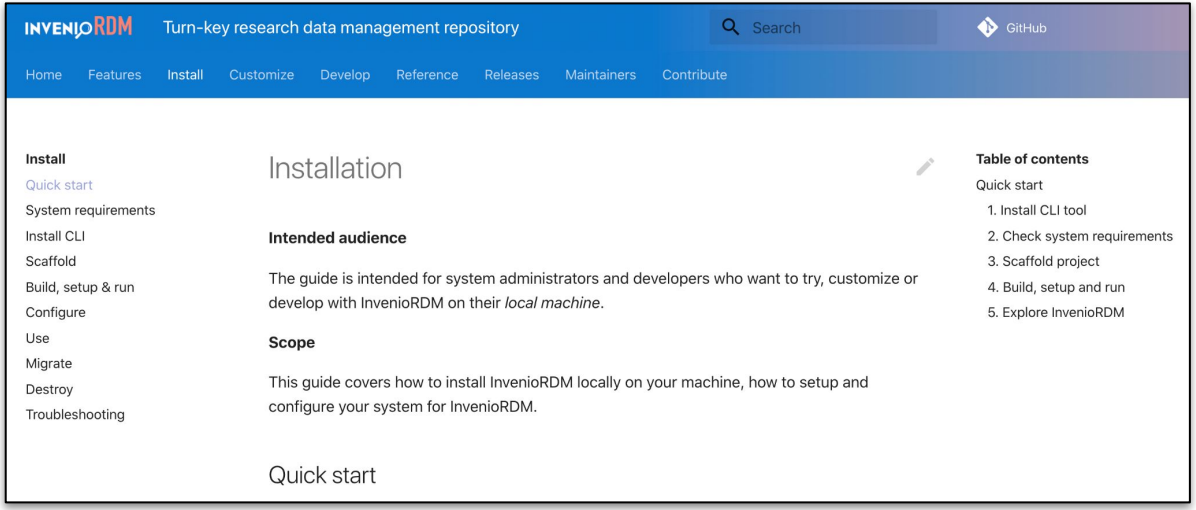
Invenio is built using **Python 3**, the **Flask** micro web framework and a suite of the best community-built Python libraries.

React

InvenioILS UIs are built using **React**, the well-known JavaScript library.

Invenio is JSON-native and provides RESTful APIs to make it easy to build apps on top of the framework

Standing up InvenioRDM



1- Install invenio-cli
pip install invenio-cli

2- Initialize your project
invenio-cli init
cd <project name>

3- Install it and run it
invenio-cli install
invenio-cli services setup
invenio-cli run

4- Visit https://localhost
firefox https://127.0.0.1:5000

System requirements

Invenio can run in Docker, on virtual machines, or on physical machines. Invenio can run on a single machine or a cluster of 100s of machines.

It all depends on exactly how much data you are handling and your performance requirements.

Small installation:

- Web/app/background servers and Redis: 1 node
- Database: 1 node
- Elasticsearch: 1 node

Medium installation:

- Load balancer: 1 node
- Web/app servers and background workers: 2 nodes
- Database: 1 node
- Elasticsearch: 3 nodes
- Redis/RabbitMQ: 1 node

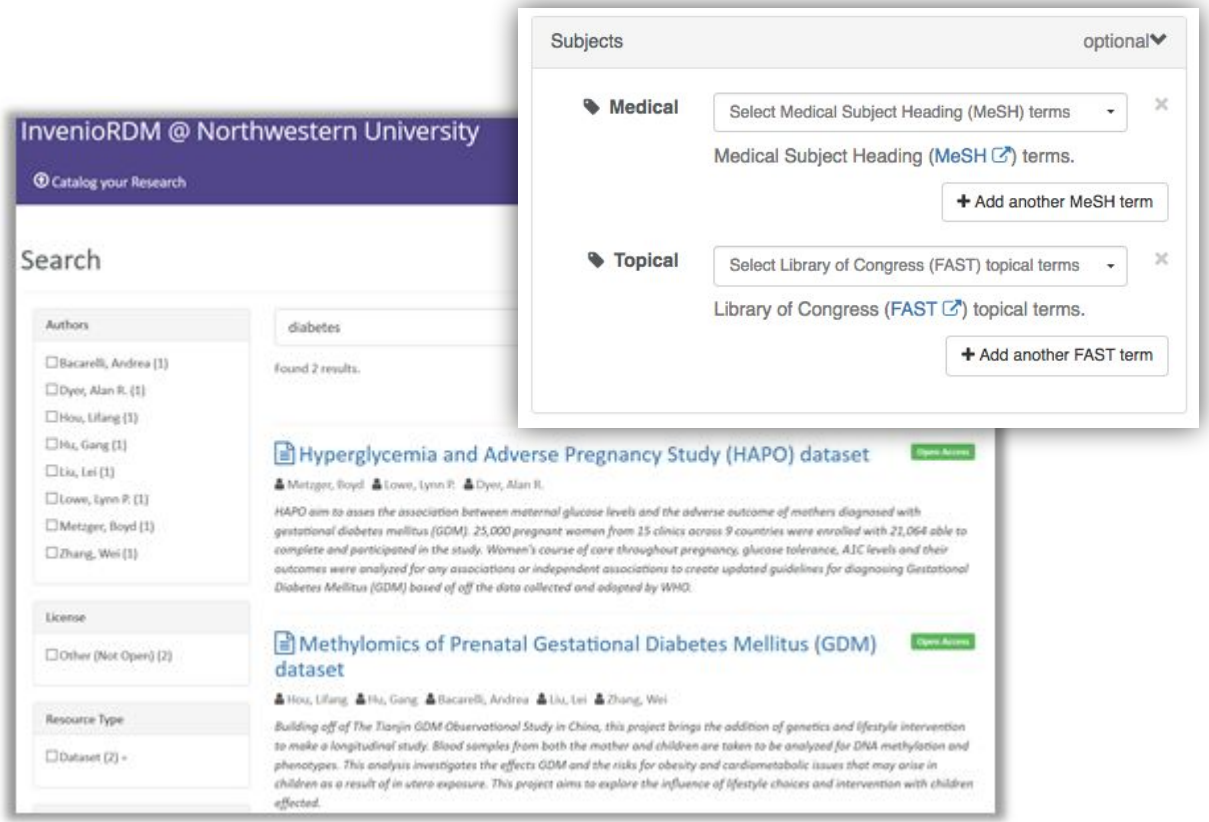
Large installation:

- Load balancer: 2 node (with DNS load balancing)
- Web/app servers: 3+ nodes
- Background workers: 3+ nodes
- Database: 2 nodes (master/slave)
- Elasticsearch: 5 nodes (3 data, 2 clients)
- Redis: 3 nodes (HA setup)
- RabbitMQ: 2 nodes (HA setup)

Search and retrieve datasets using standards-based documentation

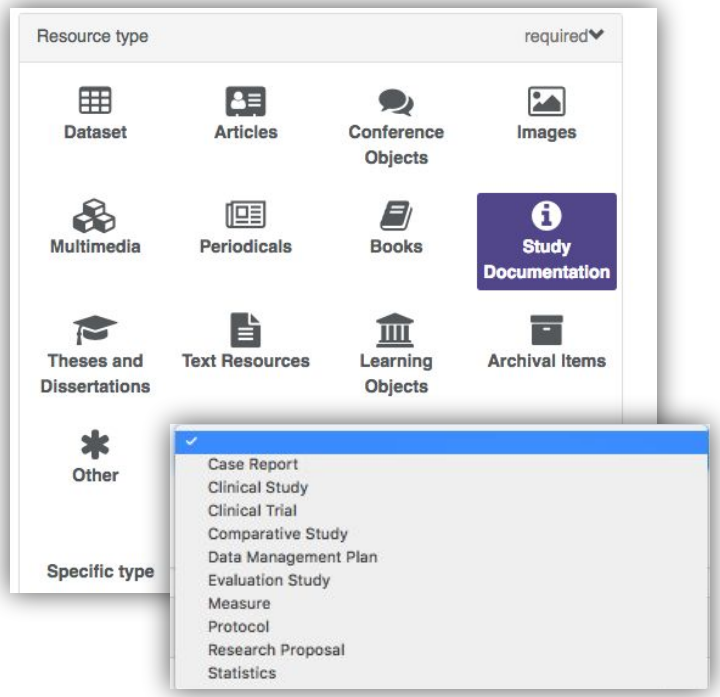
Robust search enhanced by:

- Standardized forms of name (LDAP + ORCiD coming soon)
- Standard subject terms (MeSH, Library of Congress Subject terms)
- Standardized citation formats
- Clear levels of access
- Standard application of licenses



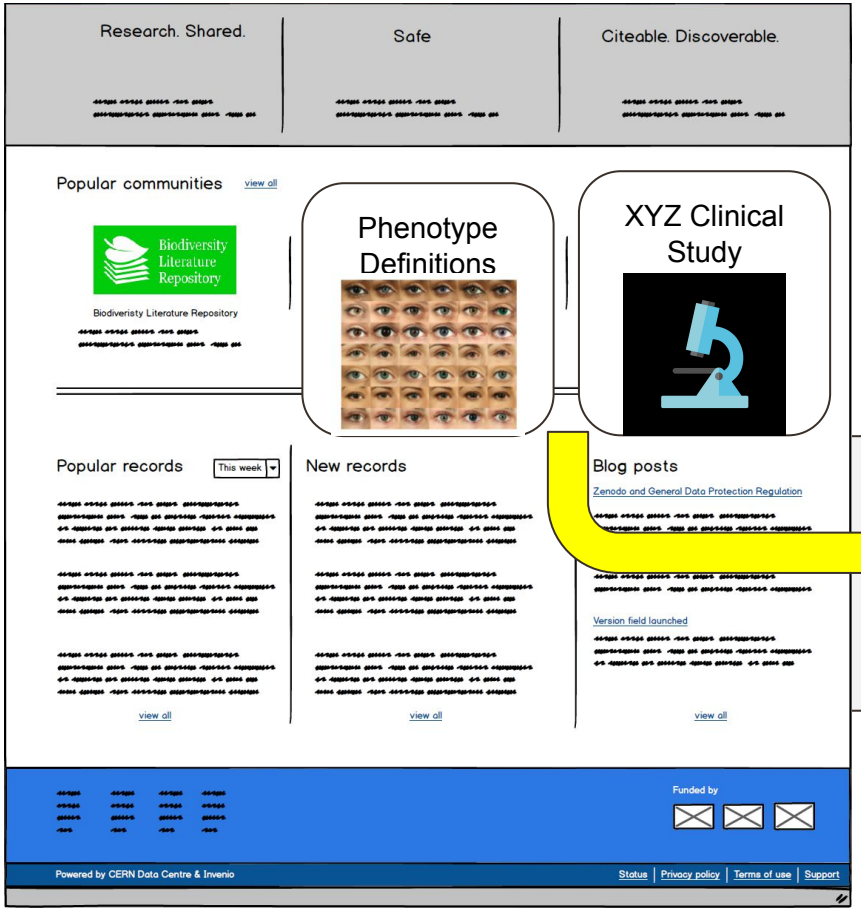
RDM

Data management for reproducibility and Open Access: study-focused resource types



InvenioRDM helps you store, manage and, if needed, share your study's outputs:

- Study-based resource types to manage a large range of assets
- **Reproducibility** is enhanced: store research proposals, datasets, code
- Be **compliant** with data sharing mandates
- **Cite** and **attribute** the work of all contributors to research
- **Reuse** deposited data or measures from other studies



Communities & Collections

Community: Define your research group or other collaborative unit

Collection: Create multiple Collections under the umbrella of the Community. Within Collections, deposit and describe your:

Phenotype Definitions
- Definitions
- Characterizations
- Evaluations
- Metadata
- Dissemination Strategy

Clinical Studies
- Research Proposals
- Protocols
- Data Management Plans
- Methods Descriptions
- Measures
- Case Reports
- Datasets and Analyses

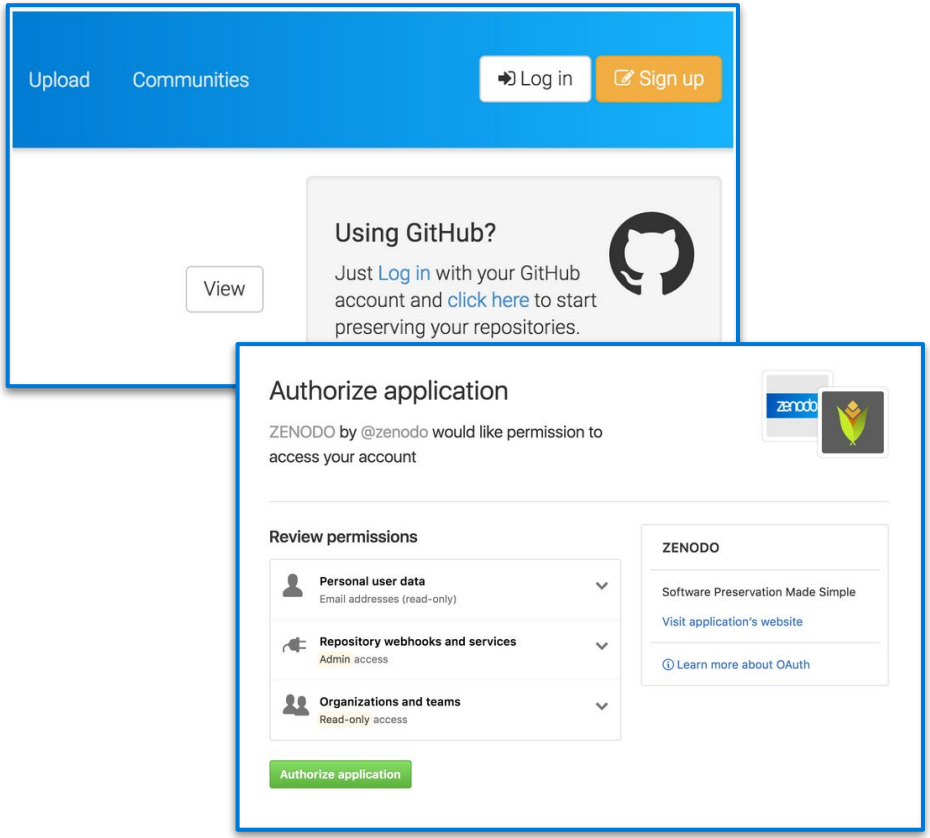
Collections bring together related groupings of documentation to communicate process, enable sharing of results, and support publication, compliance, and reproducibility

RDM

Software publication and citeability

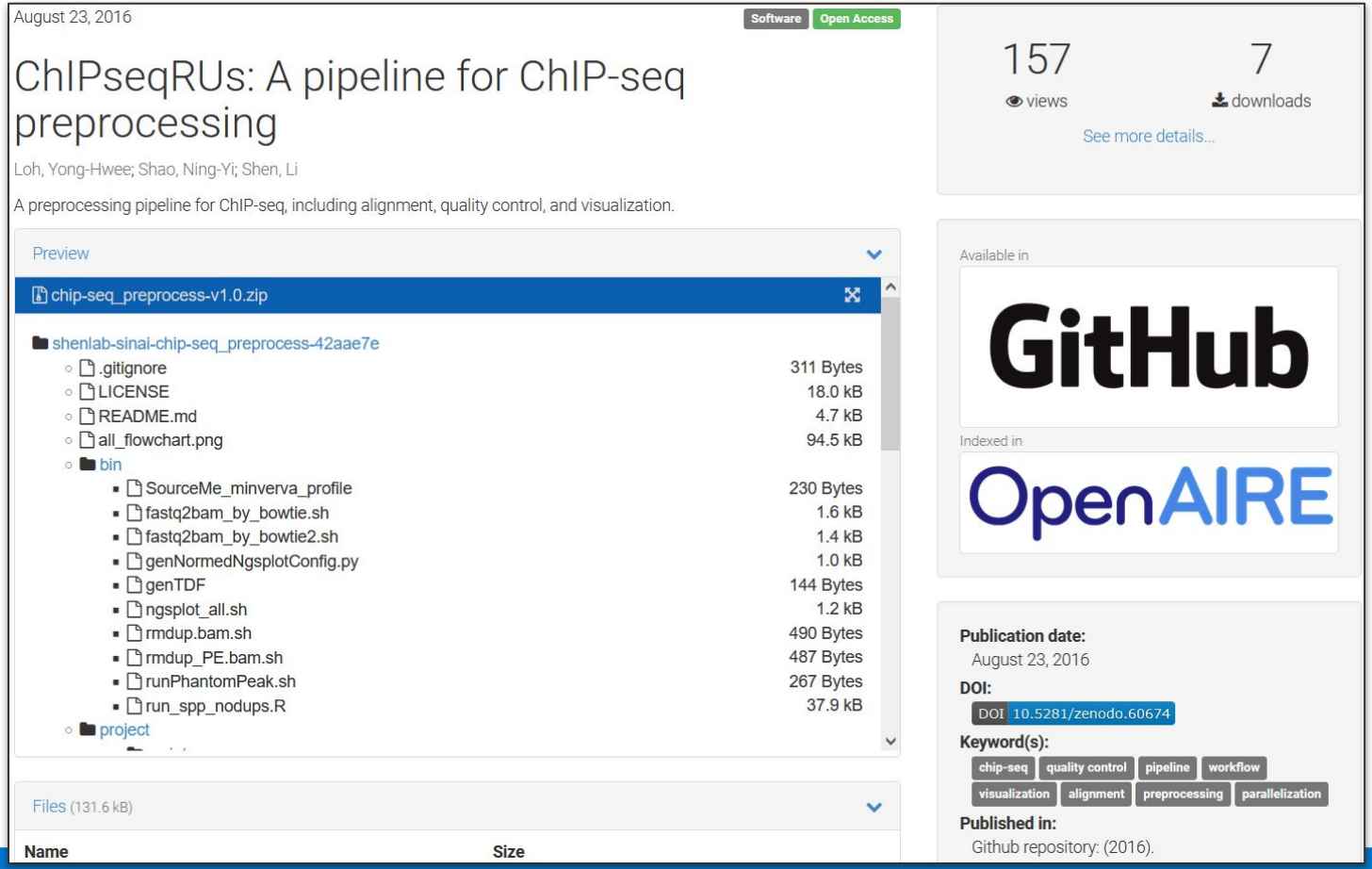
InvenioRDM will support software publication or deposit in a number of ways, including:

- Generation of citations for all resource types, including software
- Planned support for custom fields which may be necessary to integrate software metadata
- 3rd party integrations allowing a direct link to GitHub, as can currently be seen in Zenodo
  - Login to GitHub from Zenodo
  - Authorize GitHub to give Zenodo needed permissions
  - Pick the repository you want to archive
  - Create a new release



https://guides.github.com/activities/citable-code/

Software publication example



August 23, 2016

Software | Open Access

ChIPseqRUs: A pipeline for ChIP-seq preprocessing

Loh, Yong-Hwee; Shao, Ning-Yi; Shen, Li

A preprocessing pipeline for ChIP-seq, including alignment, quality control, and visualization.

Preview

chip-seq_preprocess-v1.0.zip

- shenlab-sinai-chip-seq_preprocess-42aae7e
  - .gitignore 311 Bytes
  - LICENSE 18.0 kB
  - README.md 4.7 kB
  - all_flowchart.png 94.5 kB
  - bin
    - SourceMe_minverva_profile 230 Bytes
    - fastq2bam_by_bowtie.sh 1.6 kB
    - fastq2bam_by_bowtie2.sh 1.4 kB
    - genNormedNgsplotConfig.py 1.0 kB
    - genTDF 144 Bytes
    - ngsplot_all.sh 1.2 kB
    - rmdup.bam.sh 490 Bytes
    - rmdup_PE.bam.sh 487 Bytes
    - runPhantomPeak.sh 267 Bytes
    - run_spp_nodups.R 37.9 kB
  - project

Files (131.6 kB)

Name	Size

157 views

7 downloads

See more details...

Available in GitHub

Indexed in OpenAIRE

Publication date:
August 23, 2016

DOI:
DOI 10.5281/zenodo.60674

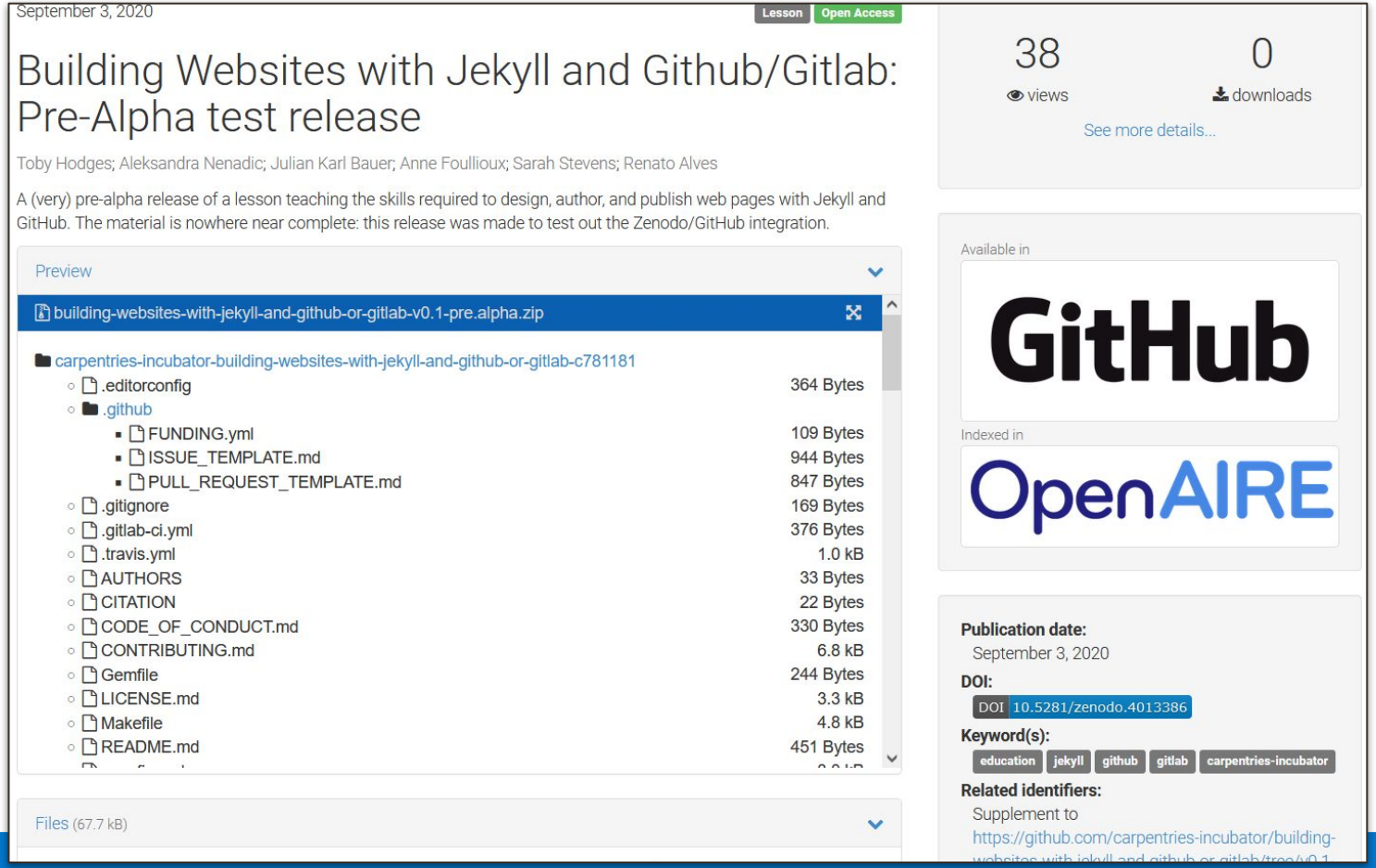
Keyword(s):
chip-seq | quality control | pipeline | workflow | visualization | alignment | preprocessing | parallelization

Published in:
Github repository: (2016).

RDM

https://doi.org/10.5281/zenodo.60674

Software publication example



September 3, 2020

Lesson Open Access

Building Websites with Jekyll and Github/Gitlab: Pre-Alpha test release

Toby Hodges; Aleksandra Nenadic; Julian Karl Bauer; Anne Foullioux; Sarah Stevens; Renato Alves

A (very) pre-alpha release of a lesson teaching the skills required to design, author, and publish web pages with Jekyll and GitHub. The material is nowhere near complete: this release was made to test out the Zenodo/GitHub integration.

Preview

building-websites-with-jekyll-and-github-or-gitlab-v0.1-pre.alpha.zip

carpentries-incubator-building-websites-with-jekyll-and-github-or-gitlab-c781181

.editorconfig	364 Bytes
.github	
FUNDING.yml	109 Bytes
ISSUE_TEMPLATE.md	944 Bytes
PULL_REQUEST_TEMPLATE.md	847 Bytes
.gitignore	169 Bytes
.gitlab-ci.yml	376 Bytes
.travis.yml	1.0 kB
AUTHORS	33 Bytes
CITATION	22 Bytes
CODE_OF_CONDUCT.md	330 Bytes
CONTRIBUTING.md	6.8 kB
Gemfile	244 Bytes
LICENSE.md	3.3 kB
Makefile	4.8 kB
README.md	451 Bytes

Files (67.7 kB)

38 views

0 downloads

See more details...

Available in

GitHub

Indexed in

OpenAIRE

Publication date:
September 3, 2020

DOI:
DOI 10.5281/zenodo.4013386

Keyword(s):
education jekyll github gitlab carpentries-incubator

Related identifiers:
Supplement to
https://github.com/carpentries-incubator/building-websites-with-jekyll-and-github-or-gitlab/tree/v0.1

RDM

https://doi.org/10.5281/zenodo.4013386

With thanks…

Teams

- The Invenio team @ CERN & RDM collaborators (here)
- Galter Health Sciences Library & Learning Center
- Northwestern University Clinical and Translational Sciences Institute
- The NU Institute for Innovations in Developmental Sciences
- Confederation of OA Repositories (COAR)

Sara Gonzales

Guillaume Viger

Matt Carson

Kristi Holmes

Support

Work presented here is supported in part by:

CERN Knowledge Transfer Fund

NUCATS: UL1TR001422 (NCATS)

all of the InvenioRDM project partners

CERN

Northwestern University NUCATS Clinical and Translational Sciences Institute

RDM

Thank you!

Where to learn more about project HERMES?

Stephan Druskat, DLR, PI, *@stdruskat*

Oliver Bertuch, FZJ, PI, *@poi_ki_lo_therm*

Guido Juckeland, HZDR, PI, *@GuidoJuckeland*

Oliver Knodel, HZDR, *@olikno1*

Tobias Schlauch, DLR, *@TobiasSchlauch*

- Find us on Twitter
- Write an email to team@software-metadata.pub
- Go to software-metadata.pub