INVENIO RDM

Powering open science and collaboration with Invenio

Northwestern University Invenio Team
03 March 2020

OHDSI: open, collaborative science

VALUES

INNOVATION	Observational research is a field which will benefit greatly from disruptive thinking. We actively seek and encourage fresh methodological approaches in our work.
REPRODUCIBILITY	Accurate, reproducible, and well-calibrated evidence is necessary for health improvement.
COMMUNITY	Everyone is welcome to actively participate in OHDSI, whether you are a patient, a health professional, a researcher, or someone who simply believes in our cause.
COLLABORATION	We work collectively to prioritize and address the real world needs of our community's participants.
OPENNESS	We strive to make all our community's proceeds open and publicly accessible, including the methods, tools and the evidence that we generate.
BENEFICENCE	We seek to protect the rights of individuals and organizations within our community at all times.

***MISSION:** To improve health by empowering a community to collaboratively generate the evidence that promotes better health decisions and better care.*

https://www.ohdsi.org/who-we-are/mission-vision-values/

List of OHDSI Working Groups

OHDSI has a variety of ongoing projects lead by Working Group teams. We would be delighted to have your participation. Please contact the team lead to join.

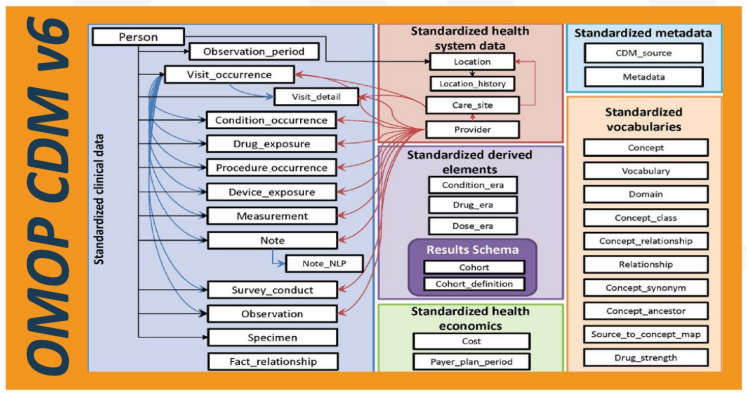
Working group leaders looking for assistance on uploading meeting recordings can find help here:

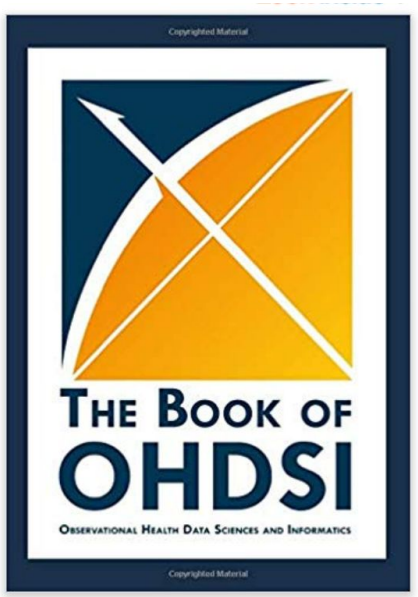
How to upload meeting recordings on the OHDSI wiki

A list of upcoming working group meetings is available here: Working Group Meeting Schedule

Active Workgroups:

- OHDSI Community
- Algorithmic Phenotyping
- Achilles WG
- Architecture WG
- Atlas & WebAPI WG
- CDM and Vocabulary Development WG
- Causal Inference
- Cerner to OMOP
- Chart Review Question Interface Project
- China WG
- Clinical Trials WG
- Devices WG
- Dissemination WG
- FHIR WG
- Genomics WG
- GIS WG
- Gold-Standard Phenotype Library
- Hadoop WG
- Latin OHDSI
- NLP WG
- Maternal & Child Health
- Metadata and Annotations WG
- OHDSI Steering Working Group
- Oncology WG
- PGHD WG
- Patient-Level Prediction WG
- Pharmacovigilance evidence investiga
- Population-Level Estimation WG
- Psychiatry WG
- Quality Measures WG
- THEMIS
- Transfusion WG
- The Book of OHDSI WG
- Women of OHDSI
- Data Quality





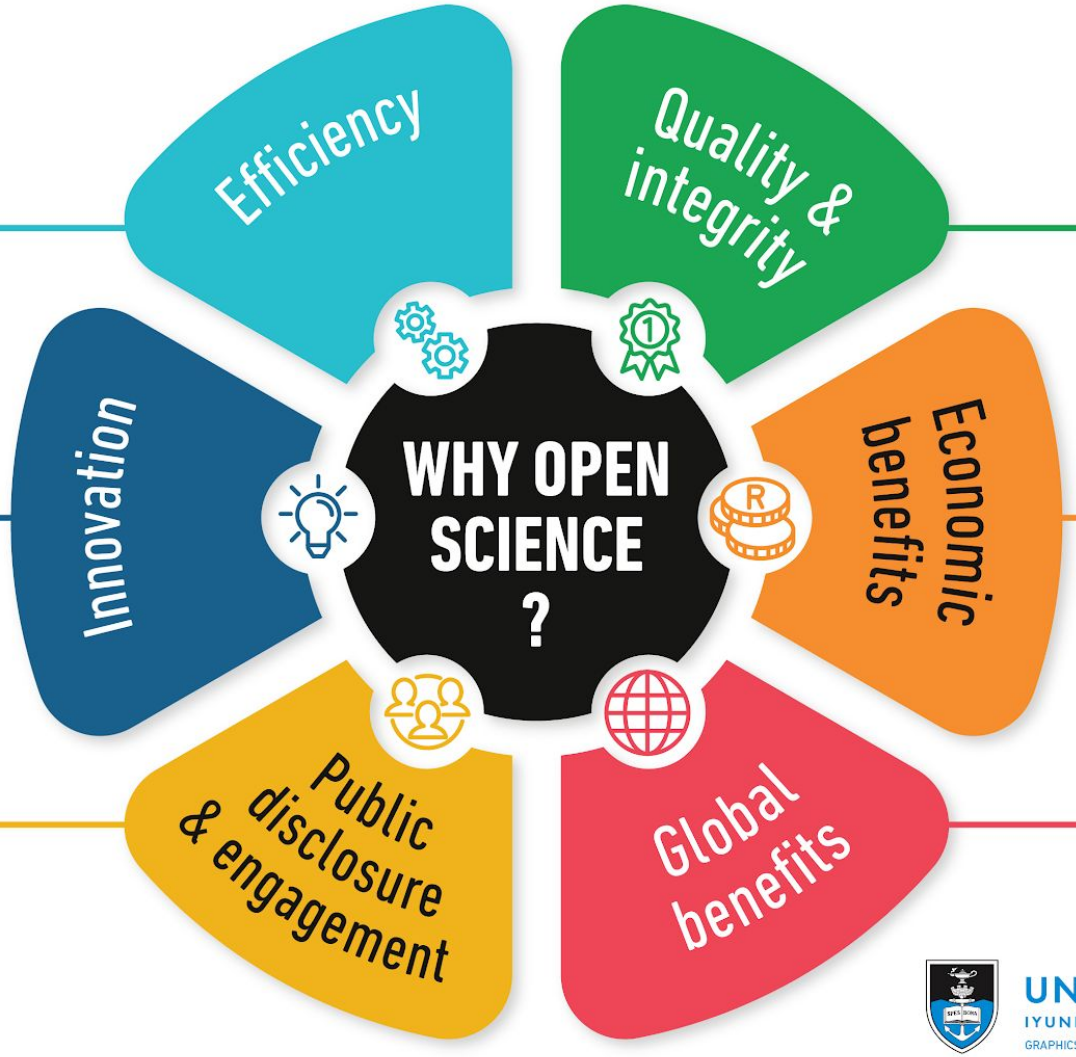
OHDSI is a multi-stakeholder, interdisciplinary collaborative to bring out the value of health data through large-scale analytics. All our solutions are open-source.

OHDSI has established an international network of researchers and observational health databases with a central coordinating center housed at Columbia University.

RDM

https://www.ohdsi.org/

Benefits of *opening* science...



Greater access to scientific inputs and outputs can increase scientific productivity through reducing duplication, allowing **more research from the same data** and multiplying opportunities for domestic and global participation in the research process.

Open science can **reduce delays in the re-use of scientific research** including articles and data, and promote a swifter path from research to innovation to produce new products and services.

Science, often publicly funded, should be publicly accessible **to promote a greater awareness** among citizens and to build public trust and support for public policies and investments in research. Open science also promotes citizen science in experiments and data collection.

Open access to scientific outputs allow for **greater evaluation and scrutiny** by the scientific community which means more accurate replication and validation of research results. Openness to data contributes to maintain science's self-correction principle.

Science plays a key role in today's knowledge economies and increased access to research results, including data, can positive impact not only scientific systems but also innovation.

Open science promotes collaborative efforts and faster knowledge transfer for a better understanding of global challenges and wicked problems.

UNIVERSITY OF CAPE TOWN
IYUNIVESITHI YASEKAPA • UNIVERSITEIT VAN KAAPSTAD
GRAPHICS BY GAELEN PINNOCK

RDM

https://upload.wikimedia.org/wikipedia/commons/5/5a/UCT_RDM_Why-Open-Science.png

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

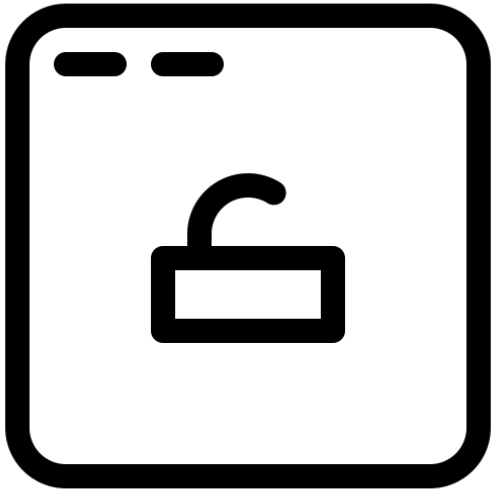
Build a turn-key research data management (RDM) repository platform based on **Invenio Framework** and **Zenodo**.

Community

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

The platform

A few highlights...

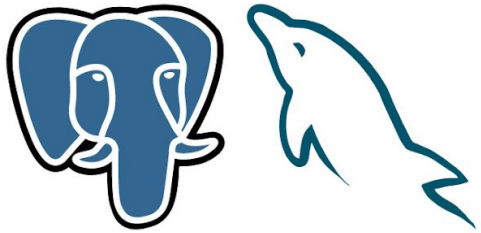


https://thenounproject.com/

InvenioRDM stack

Elasticsearch

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

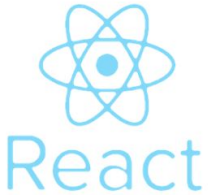


PostgreSQL or MySQL

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

InvenioRDM roadmap

February

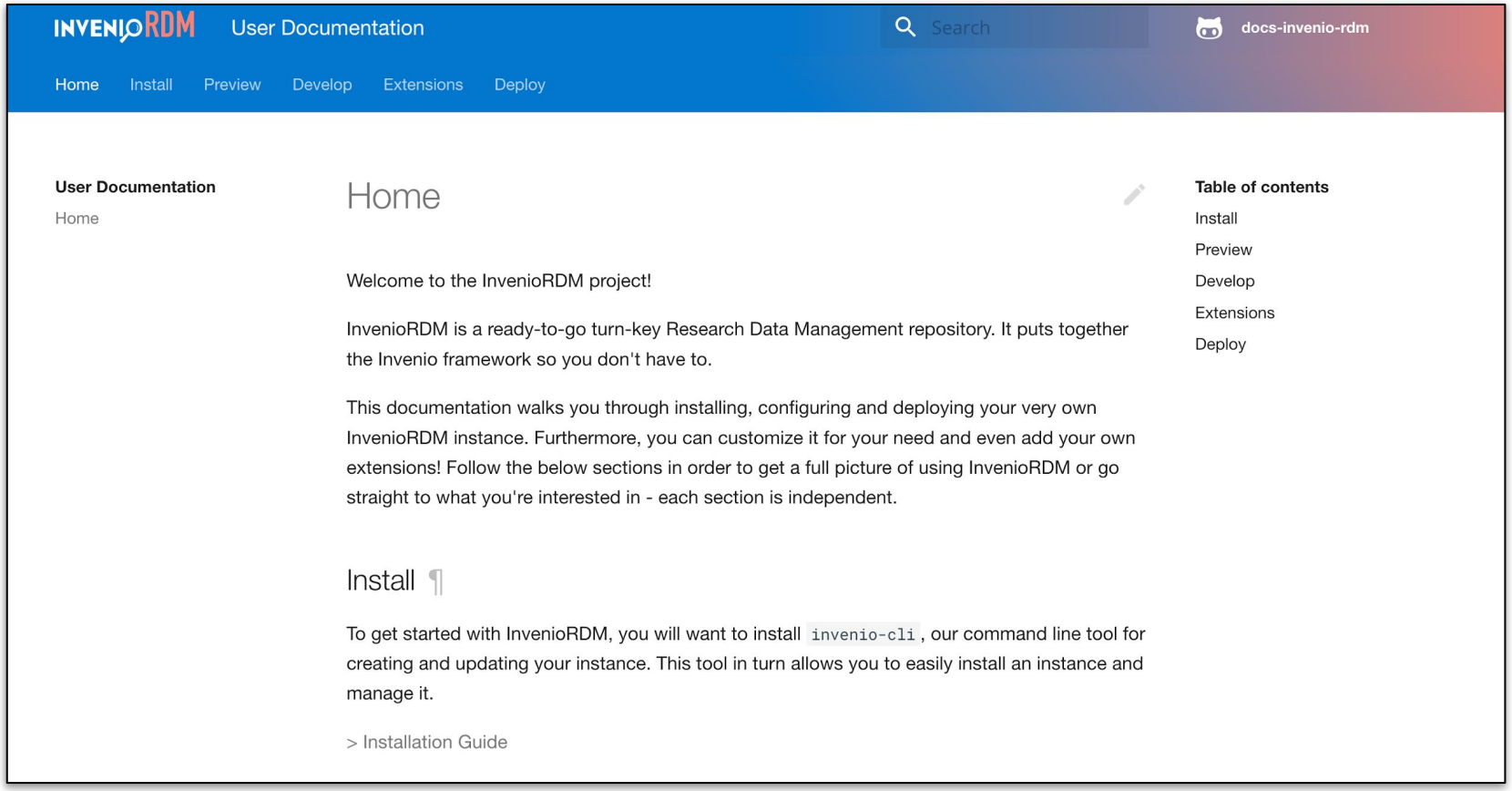
- Milestones: Draft governance and sustainability plan, mock-up feedback from collaborators
- Release
  - Branding customization - institutional theming can be applied
  - First iteration of the Data model
  - Improved CLI with improved workflow
  - New documentation site for developers (http://inveniordm.docs.cern.ch)
  - Closer project tracking with enhanced structure and outreach

March

- Milestones: First release for core plugins, review of February release
- Release
  - Search permissions
  - Deposit page
  - Improved record page
  - Data model update

- To see further ahead: https://invenio-software.org/products/rdm/roadmap/

Standing up InvenioRDM



1- Install invenio-cli
pip install invenio-cli

2- Initialize your project
invenio-cli init --flavour=RDM

3- Run it
cd <project name>
invenio-cli containerize

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

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

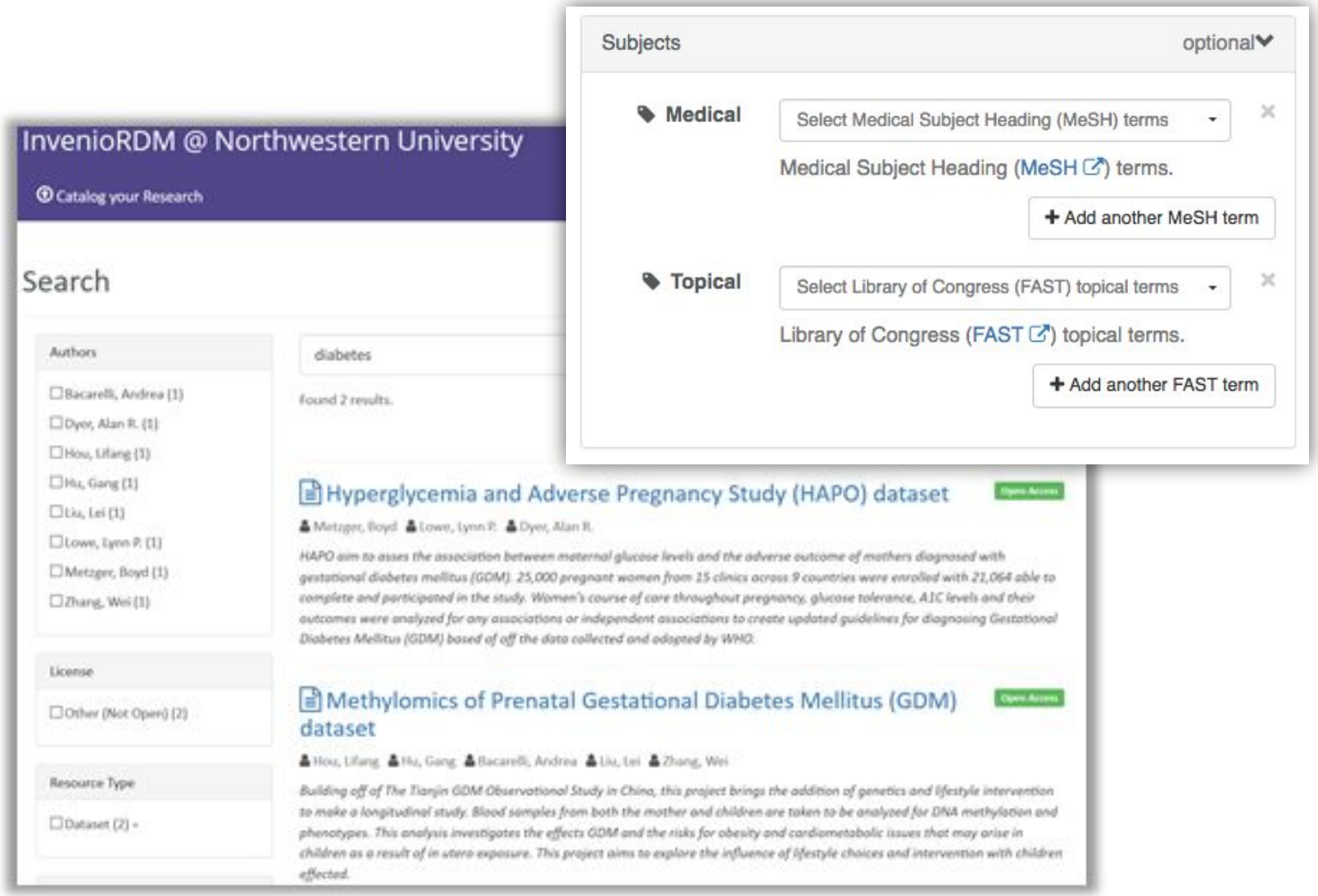
Large installation:

- 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

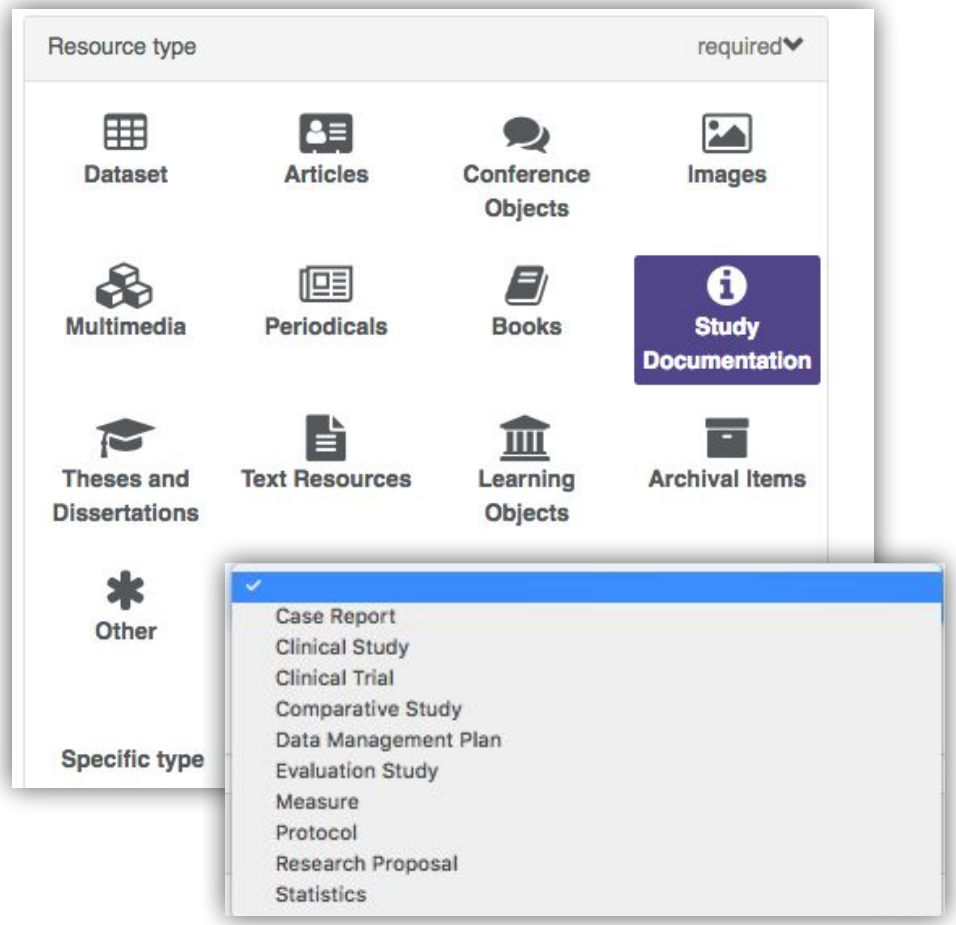
Robust search enhanced by:

- 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



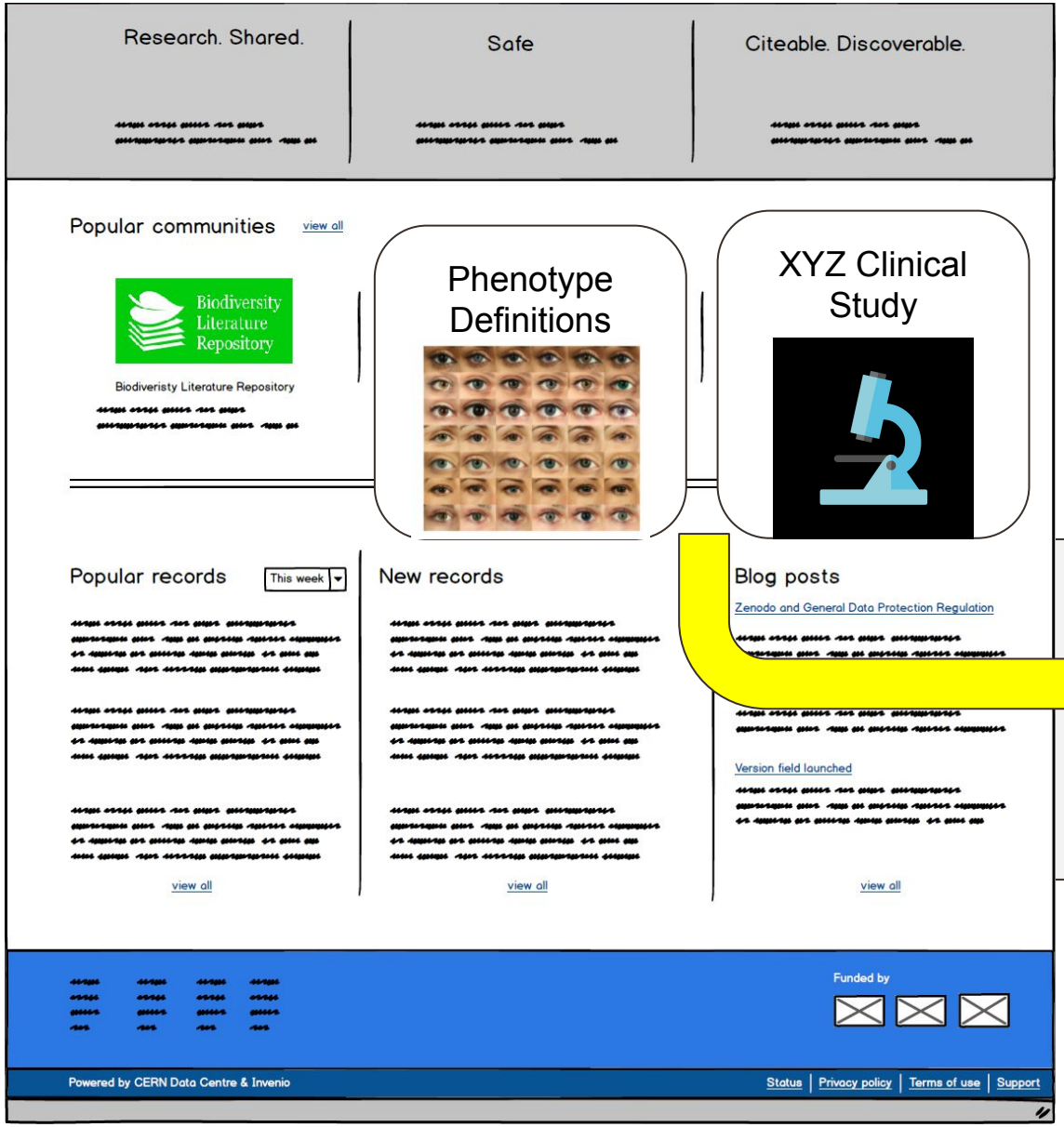
RDM

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

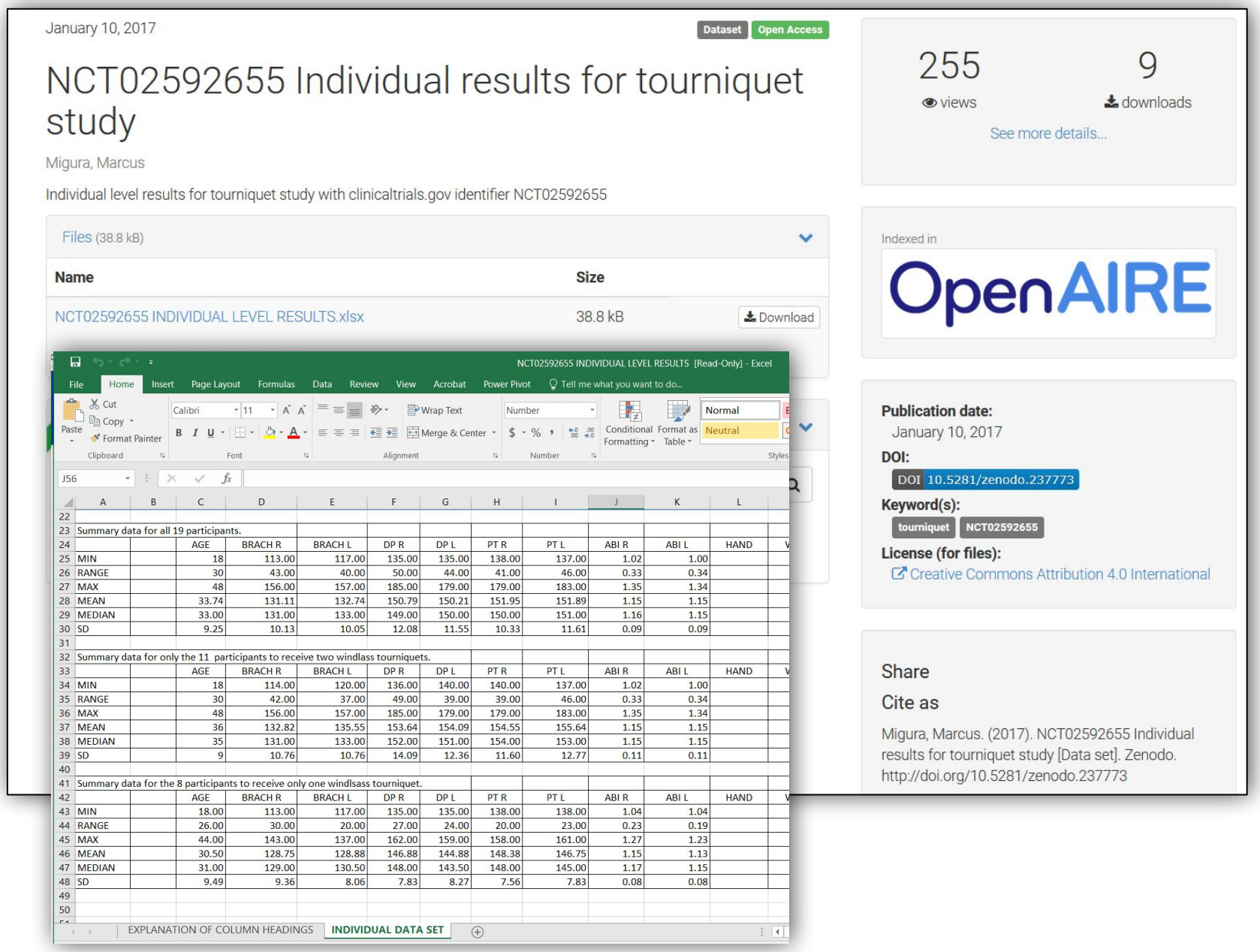
Collections & Clinical Studies

Store multiple datasets with large numbers of detailed results from each analysis and re-use of data generated by a single study.

Results presented in InvenioRDM are:

- easy to find
- browsable
- publicly available
- citeable

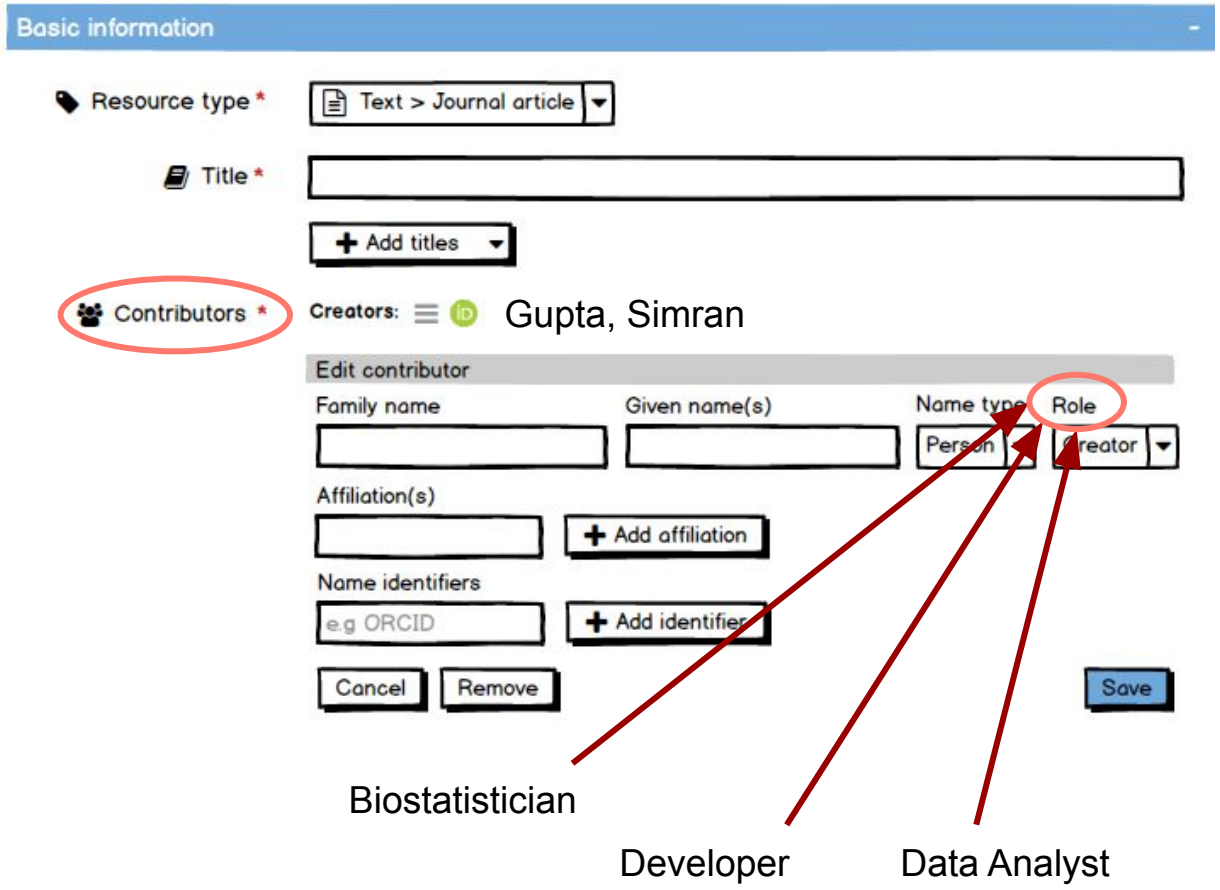
Hone in on the results you seek using InvenioRDM's robust metadata of subject and resource type terms.



Summary data for all 19 participants.												
		AGE	BRACH R	BRACH L	DP R	DP L	PT R	PT L	ABI R	ABI L	HAND	
MIN		18	113.00	117.00	135.00	135.00	138.00	137.00	1.02	1.00		
RANGE		30	43.00	40.00	50.00	44.00	41.00	46.00	0.33	0.34		
MAX		48	156.00	157.00	185.00	179.00	179.00	183.00	1.35	1.34		
MEAN		33.74	131.11	132.74	150.79	150.21	151.95	151.89	1.15	1.15		
MEDIAN		33.00	131.00	133.00	149.00	150.00	150.00	151.00	1.16	1.15		
SD		9.25	10.13	10.05	12.08	11.55	10.33	11.61	0.09	0.09		
Summary data for only the 11 participants to receive two windlass tourniquets.												
		AGE	BRACH R	BRACH L	DP R	DP L	PT R	PT L	ABI R	ABI L	HAND	
MIN		18	114.00	120.00	136.00	140.00	140.00	137.00	1.02	1.00		
RANGE		30	42.00	37.00	49.00	39.00	39.00	46.00	0.33	0.34		
MAX		48	156.00	157.00	185.00	179.00	179.00	183.00	1.35	1.34		
MEAN		36	132.82	135.55	153.64	154.09	154.55	155.64	1.15	1.15		
MEDIAN		35	131.00	133.00	152.00	151.00	154.00	153.00	1.15	1.15		
SD		9	10.76	10.76	14.09	12.36	11.60	12.77	0.11	0.11		
Summary data for the 8 participants to receive only one windlsass tourniquet.												
		AGE	BRACH R	BRACH L	DP R	DP L	PT R	PT L	ABI R	ABI L	HAND	
MIN		18.00	113.00	117.00	135.00	135.00	138.00	138.00	1.04	1.04		
RANGE		26.00	30.00	20.00	27.00	24.00	20.00	23.00	0.23	0.19		
MAX		44.00	143.00	137.00	162.00	159.00	158.00	161.00	1.27	1.23		
MEAN		30.50	128.75	128.88	146.88	144.88	148.38	146.75	1.15	1.13		
MEDIAN		31.00	129.00	130.50	148.00	143.50	148.00	145.00	1.17	1.15		
SD		9.49	9.36	8.06	7.83	8.27	7.56	7.83	0.08	0.08		

Properly attribute all contributors to research

https://data2health.github.io/CTS-Personas/

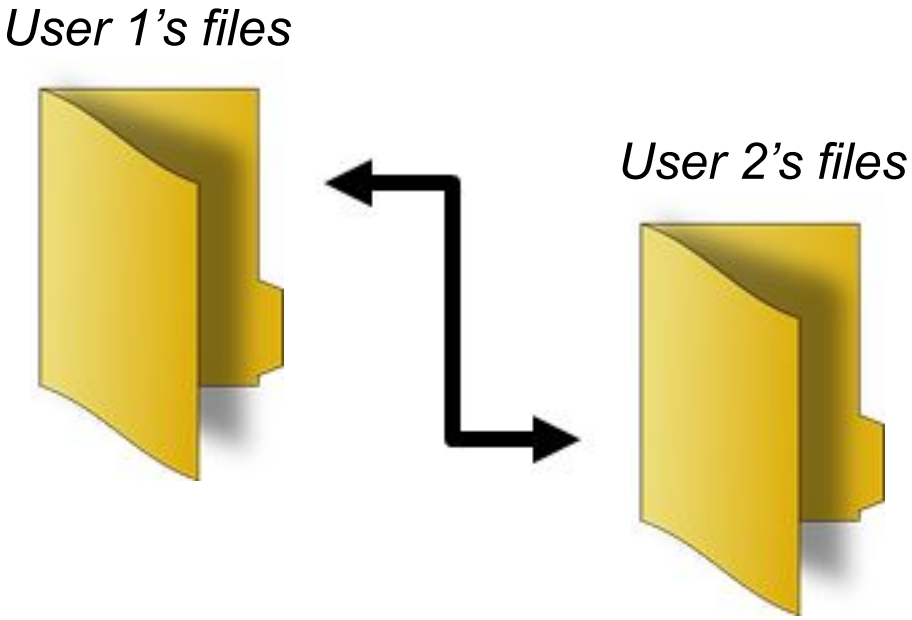


InvenioRDM incorporates **contributor roles** for all records. Deposit your SQL code, statistical analysis plan, database code, and other study documentation; receive credit, and group all documents in a Collection

RDM

Gonzales, S., O'Keefe, L., Gutzman, K., Viger, G., Wescott, A., Farrow, B., . . . Holmes, K. (n.d.). Personas for the Translational Workforce. *Journal of Clinical and Translational Science,* 1-27. doi:10.1017/cts.2020.2

Collaborators: Work with them and discover new ones



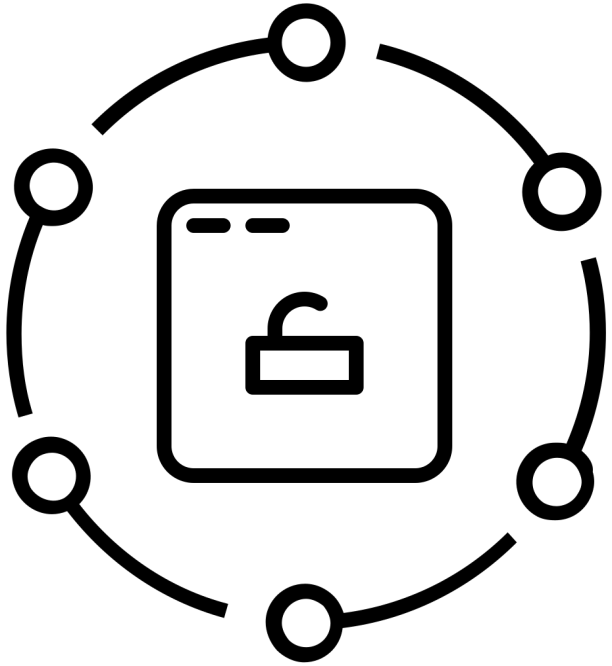
InvenioRDM will allow **private record sharing**, so researchers can:

- Share files with each other, but not anyone else in the university community or the public
- Vet materials collaboratively and privately before switching records to 'public' for open access/data sharing

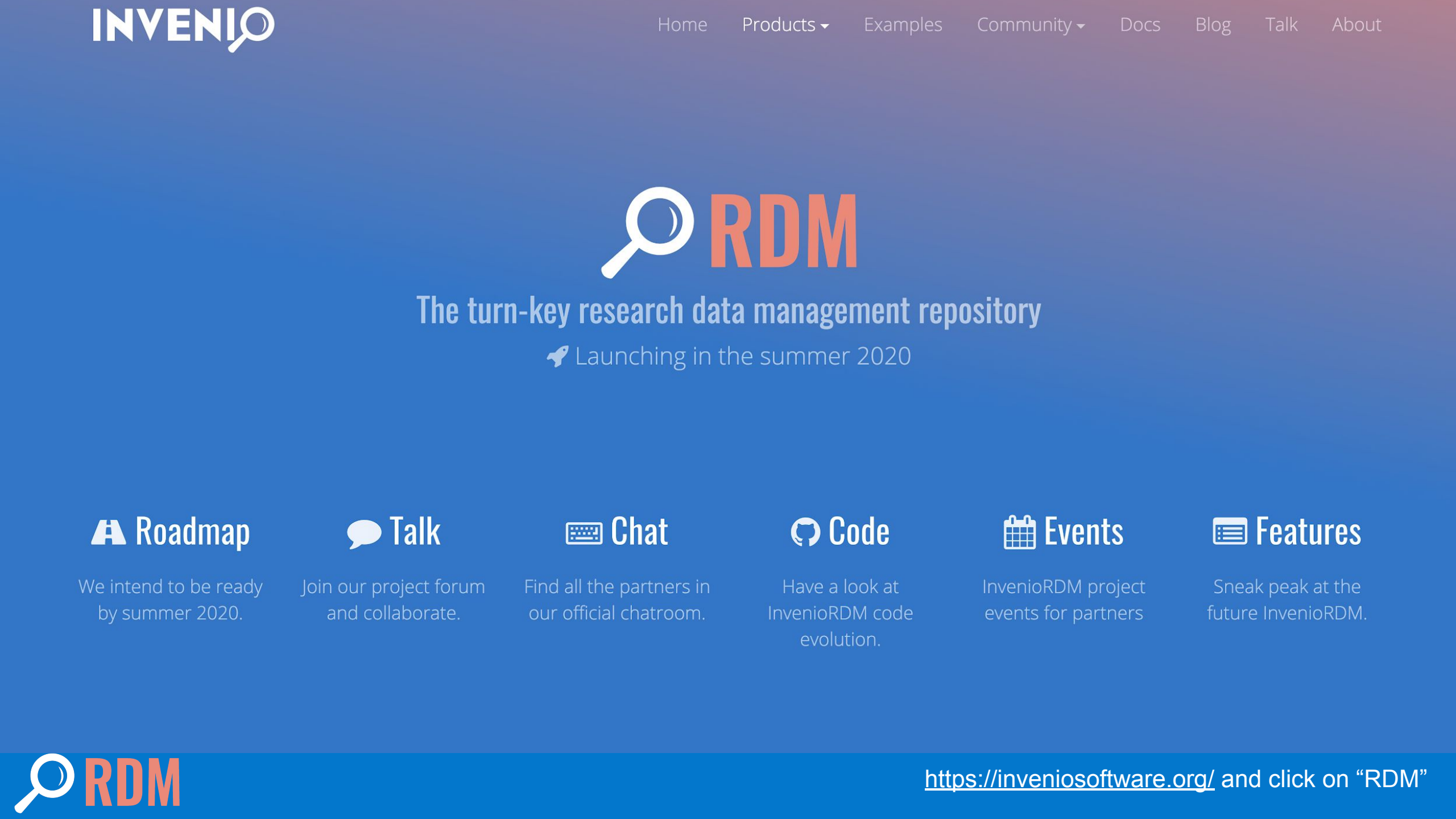
InvenioRDM will have a social component, allowing researchers to:

- Follow other researchers
- Receive updates when someone they follow deposits something
- Manage requests to access files represented by a metadata-only record

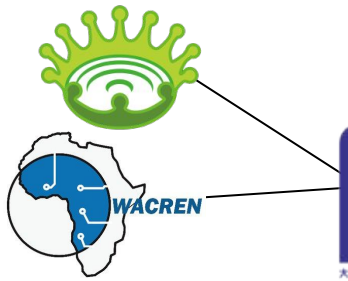
The community



https://thenounproject.com/


INVENIO
Home
Products
Examples
Community
Docs
Blog
Talk
About
RDM
The turn-key research data management repository
Launching in the summer 2020
Roadmap
We intend to be ready by summer 2020.
Talk
Join our project forum and collaborate.
Chat
Find all the partners in our official chatroom.
Code
Have a look at InvenioRDM code evolution.
Events
InvenioRDM project events for partners
Features
Sneak peak at the future InvenioRDM.
RDM
https://inveniosoftware.org/ and click on “RDM”

InvenioRDM collaborators



BROOKHAVEN NATIONAL LABORATORY

Caltech Library

CERN

NATIONAL CENTER FOR DATA TO HEALTH

data futures

HZDR HELMHOLTZ ZENTRUM DRESDEN ROSSENDORF

INFN

Joint Research Centre JRC

NII 大学共同利用機関法人 情報・システム研究機構 国立情報学研究所 National Institute of Informatics

NORTHWESTERN UNIVERSITY

OpenAIRE

TIND

TÜBİTAK ULAKBİM

TRECCANI LA CULTURA ITALIANA

TU Graz

UHH Universität Hamburg DER FORSCHUNG | DER LEHRE | DER BILDUNG

WWU MÜNSTER

How can Invenio support the OHDSI community?

Some Use Cases

We're managing a large multi-site project, harmonizing data from numerous sources and managing research projects. We want to create communities of practice to integrate theories, data, techniques, and tools.

I lead a large basic science research group. We use InvenioRDM to support reproducible science by packaging combined with big data mining, a desire to process collected data using the latest bioinformatics tools.

I am a clinical researcher. I need a way to pre-register protocols or research proposals, search on demographics of participants in similar studies, get insights into recruitment, share portions of study for compliance.

My team wants to find out about clinical trial opportunities to offer patients all options for treatment. It is important to us to openly share the latest research with patients. InvenioRDM communities give us a way to make these materials openly available and packaged in a cohesive and attractive manner. As resources are updated, we can upload the new versions and track access.

Our multi-institution health equity project uses InvenioRDM to collaborate with our community-based partners and credit these partnerships. We can share materials from community health events, project materials, training materials, annual reports, and lay summaries of research. InvenioRDM helps us to be better partners, accountable to collaborators and the community,

I'm an early career researcher just getting started on my research career. I need to "put my best foot forward" to showcase my work and demonstrate my expertise and collaborations. Invenio gives me a way to make all of my research efforts findable and the metrics are helpful for reporting. and highlighting my impact to my leadership.

Our institute wants a way to publish and disseminate content such as our handbook, lay summaries, and more. We want to credit all contributors and produce an attractive and interactive resource that can be easily updated.

FAIR: OHDSI & InvenioRDM

Findable

InvenioRDM's records are made **findable** through each being issued a Digital Object Identifier (DOI), and through their metadata being indexed and made searchable immediately.

OMOP database summaries can be published in InvenioRDM as findable descriptor records to reference the database for reproducibility and citation

Accessible

Metadata in InvenioRDM are **accessible** because they are retrievable using a standardized communications protocol which is free and universally implementable.

OMOP data can be mapped through similar open protocols through SQL interfaces, though largely for secure querying. Results of analyses in multiple OMOP databases can be cataloged in InvenioRDM, and these records retrieved through the open protocol OAI-PMH.

Interoperable

InvenioRDM leverages metadata encoding (JSON) and vocabulary (FundRef, OpenAIRE, COAR Resource Types, etc.) standards to ensure maximum **interoperability** for records describing digital assets.

OMOP similarly ensures interoperability through its CDM and standardized vocabulary, and the OHDSI community goes beyond this work by providing a platform to enable an interoperable understanding of the analysis methods for healthcare data.

Reusable

Ensuring the **reusability** of digital assets deposited in InvenioRDM is key and is achieved through assigning licenses and establishing provenance through registering users.

OHDSI's Metadata Working Group is actively working toward attaching provenance information to OMOP records.

Links

- **Official InvenioRDM site:** https://inveniosoftware.org/products/rdm/
- **Roadmap:** https://inveniosoftware.org/products/rdm/roadmap/
- **GitHub:** https://github.com/inveniosoftware/invenio-app-rdm
- **Documentation:** https://invenio-app-rdm.readthedocs.io/en/latest
- **Project Boards:** https://github.com/orgs/inveniosoftware/projects
- **RFC (Request for Comments):** https://github.com/inveniosoftware/rfcs

Northwestern's Proof of Concept: http://bit.ly/inveniordm-at-nu

- Test Login: gla3975
- Password: InvenioRDM@NU_2019

Install your own instance! https://inveniordm.docs.cern.ch/

With thanks…

Teams

- The Invenio team @ CERN & RDM collaborators (here)
- Galter Health Sciences Library & Learning Center
- Northwestern University Clinical and Translational Sciences Institute (NUCATS)
- CTSA Program Center for Data to Health (CD2H) team
- The NU Institute for Innovations in Developmental Sciences
- Confederation of OA Repositories (COAR)

Sara Gonzales

Guillaume Viger

Lisa O'Keefe

Matt Carson

Support

Work presented here is supported in part by:
CERN Knowledge Transfer Fund
CD2H: U24TR002306 (NCATS)
NUCATS: UL1TR001422 (NCATS)

all of the InvenioRDM project partners

CERN

Northwestern University NUCATS Clinical and Translational Sciences Institute

NATIONAL CENTER FOR DATA TO HEALTH

RDM