

# Next Generation Repositories & Biomedicine

Kristi Holmes

[kristi.holmes@northwestern.edu](mailto:kristi.holmes@northwestern.edu)

@kristiholmes

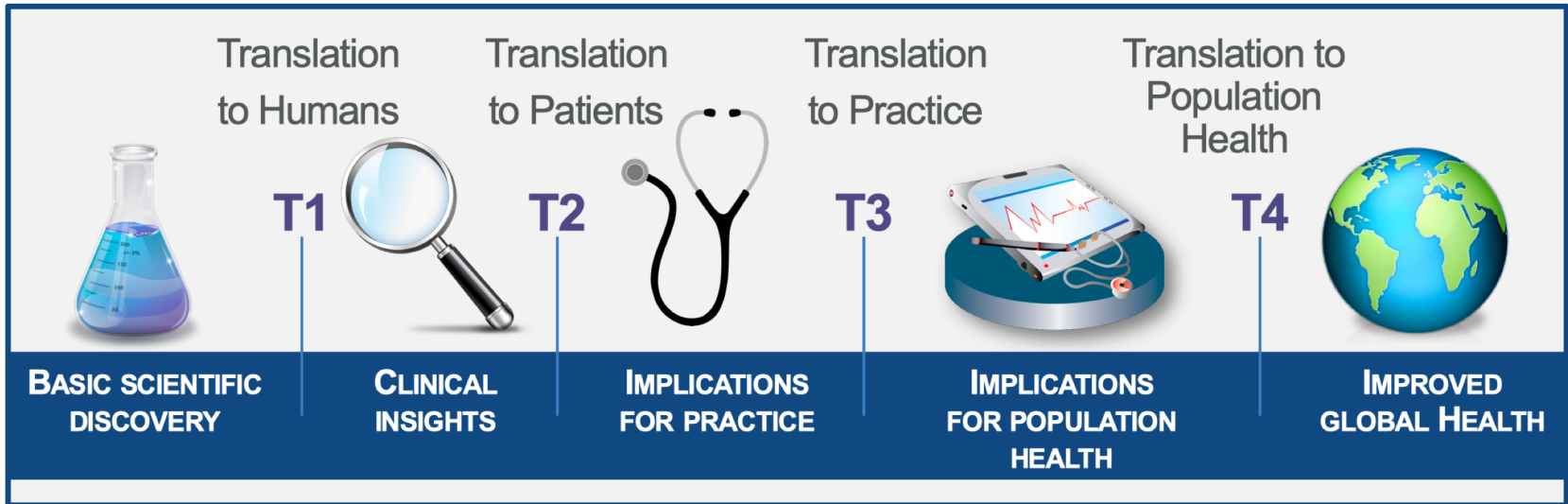
Northwestern University Feinberg School of Medicine

Galter Health Sciences Library & Learning Center

Northwestern University Clinical and Translational Sciences Institute

CTSA Program National Center for Data to Health (CD2H)

# A Focus on Translational Science



## The CTSA Program from the National Institutes of Health

- A national network of **>50 medical research institutions in the US** that work together to improve the translational research process to get more treatments to more patients more quickly.
- CTSA Program support enables research teams including scientists, patient advocacy organizations and community members to tackle system-wide scientific and operational problems in clinical and translational research that no one team can overcome.

# The National Center For Data to Health (CD2H)

*Informatics & data science coordinating center for the CTSA Program*

## ***Accelerating Informatics Innovation to Advance Translational Research***



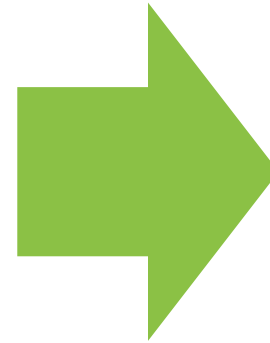
Make Data Easier to Share and Re-use



Make Tools More Accessible and Interoperable

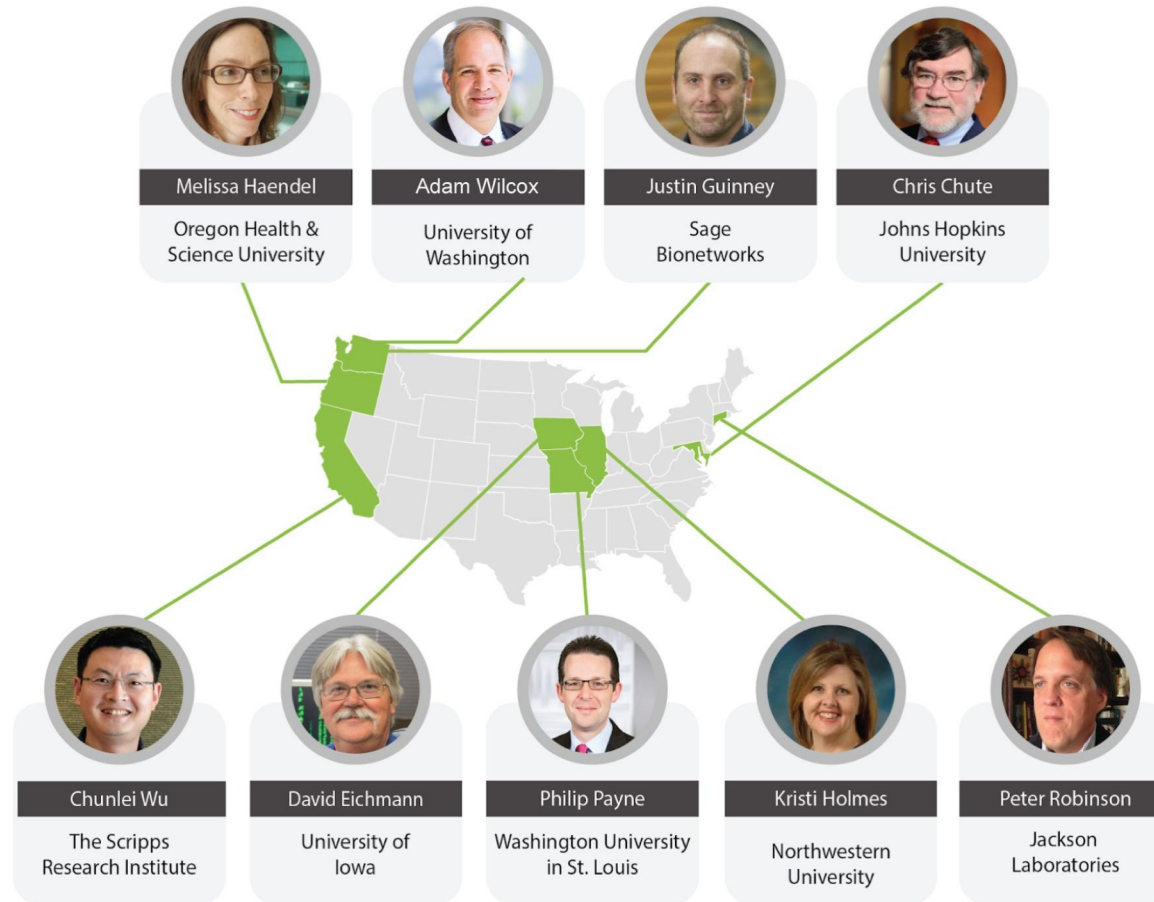


Leverage Expertise and Foster a More Collaborative CTSA Culture



Better translation of research and improved patient care

# Who we are and who we serve



Workforce  
Development



Collaboration/  
Engagement



Integration Across  
the Lifespan



Methods/  
Processes



Informatics



...& the larger informatics  
community

# What's important: patient care.



# What's important: healthy communities and empowering access to research & partnership.



[f Like us](#) [Follow us](#) [✉ Email us](#) [✍ Signup for updates](#) [💡 Questions?](#) [🌐 Google Translate](#)

[About](#) [Research](#) [Opportunities](#) [Resources](#) [Events](#) [News](#)



## About ChicagoCHEC

Our **mission** is to advance **cancer health equity** through meaningful scientific discovery, education, training, and community engagement.

[Mission](#) [Background](#) [Impact](#) [Team](#) [Institutions](#) [Community Partners](#) [Cores](#)

## Requires

- *tri-institutional partnership and a focus on cancer health equity.*
- *collaborations with the community on cancer health equity issues.*

To deliver access to research, improve patient care, it takes both technology & culture.



# We need a better way of thinking about integrated repository infrastructure – and Next Generation Repositories + Invenio v3 make this possible

*NGR make the resource, rather than the repository, the focus of services and infrastructure*  
<http://ngr.coar-repositories.org/>

## VISION:

*A foundation for a **distributed, globally networked infrastructure** on top of which layers of **value added services** can be deployed, making it more **research-centric, open to and supportive of innovation**, while also **collectively managed by the scholarly community**.*



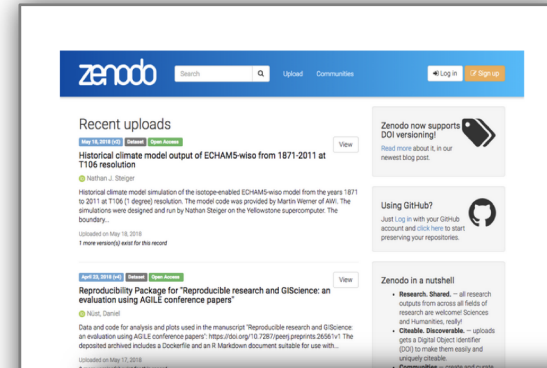
## GUIDING PRINCIPLES:

- **Distribution of control:** More sustainable and at less risk to buy-out or failure.
- **Inclusiveness and diversity:** Reflect and be responsive to different needs.
- **Public good:** Available to all via global standards
- **Intelligent openness and accessibility:** Resources are openly available and in accessible formats to increase their value and maximize re-use.
- **Sustainability:** Research organizations are major participants in the global network.
- **Interoperability:** Adopt common behaviors, functionalities, and standards for interoperability.



# We're leveraging Invenio 3.0 as a strong foundation. Here's why.

- ✓ **Safe:** Invenio has been created with security and long-term preservation in mind.
- ✓ **Scalable:** Invenio is fast. Designed to manage 100+ million records and petabytes of files. All your research data can now be archived independently of the size.
- ✓ **RESTful:** Only a modern framework can create modern digital repositories. Invenio was born for the web, is JSON-native and provides RESTful APIs out of the box that will allow to build apps on top of it.
- ✓ **Open:** Invenio is 100% open source licensed under MIT license. Open standards for open science.
- ✓ **A robust community:** Large team of developers & active open source community. TIND (CERN spinoff) uses a SAAS-model for service. Invenio is used by [many organizations](#), and the underlying technology (Python, Flask) is widely supported, new RDM initiative at CERN



## Interoperability

### COAR Next Generation Repositories (NGR)

- SWORD3
- ResourceSync
- Signposting
- COUNTER
- SUSHI

See <http://ngr.coar-repositories.org/>

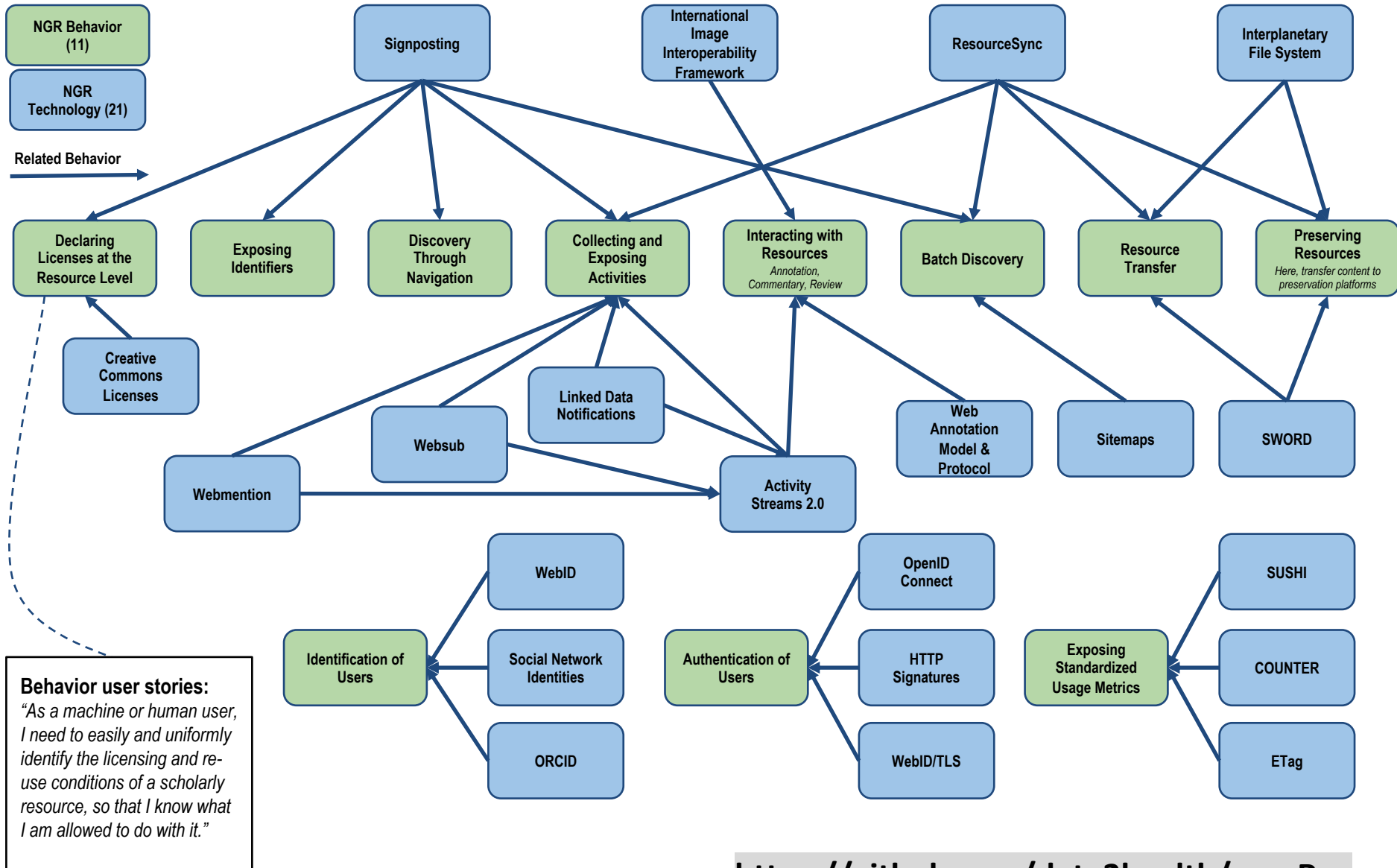
(with Japan National Institute of Informatics)

*Collect, record, preserve, and disseminate a wide range of research objects across the translational workforce (e.g., datasets, protocols, consent materials, education or engagement materials, technical reports, supplemental materials, survey instruments)*



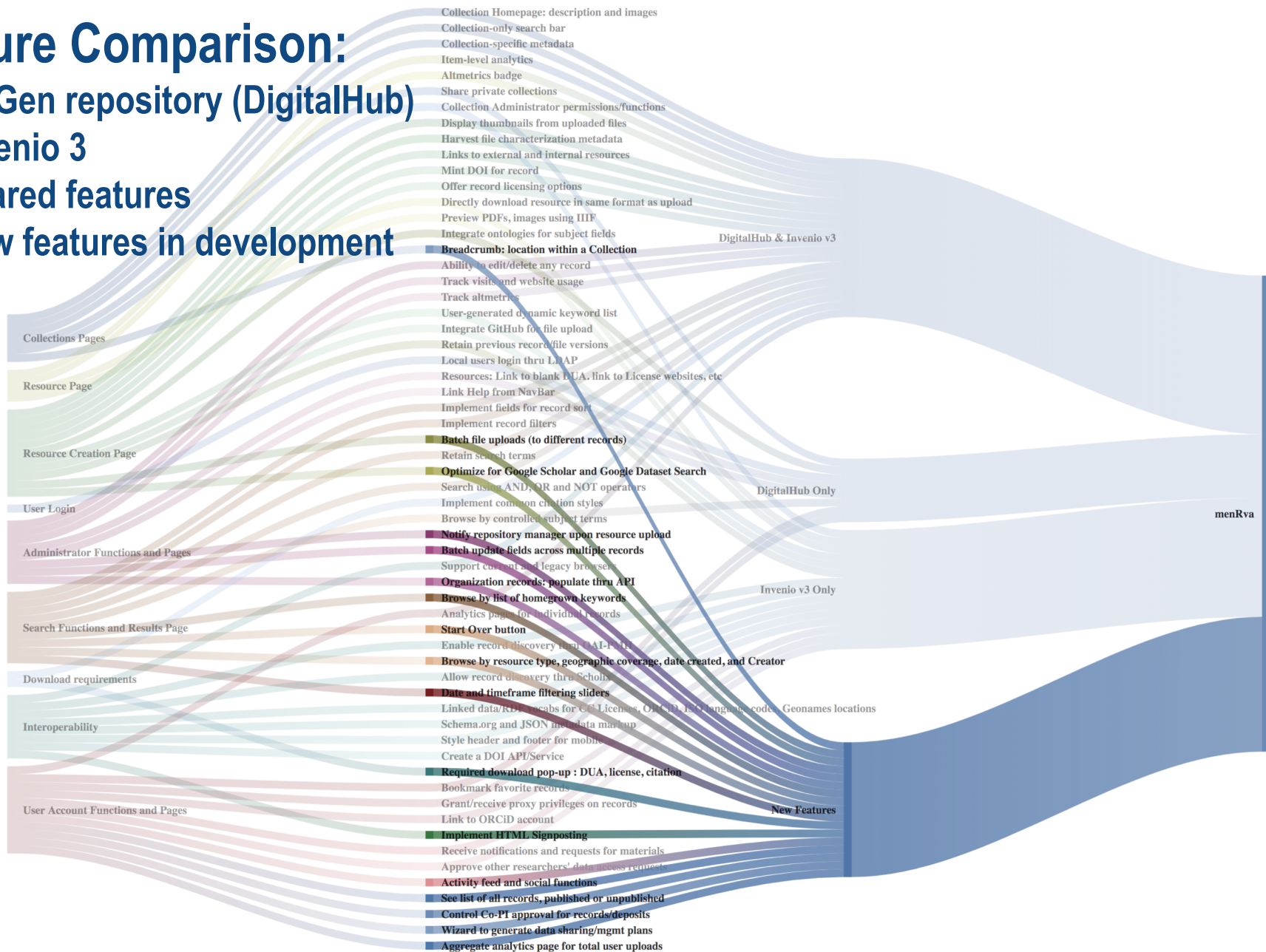
NATIONAL CENTER  
FOR DATA TO HEALTH

# Mapping COAR NGR Behaviors / Technologies helps guide our own journey: where do we want to go & how to get there

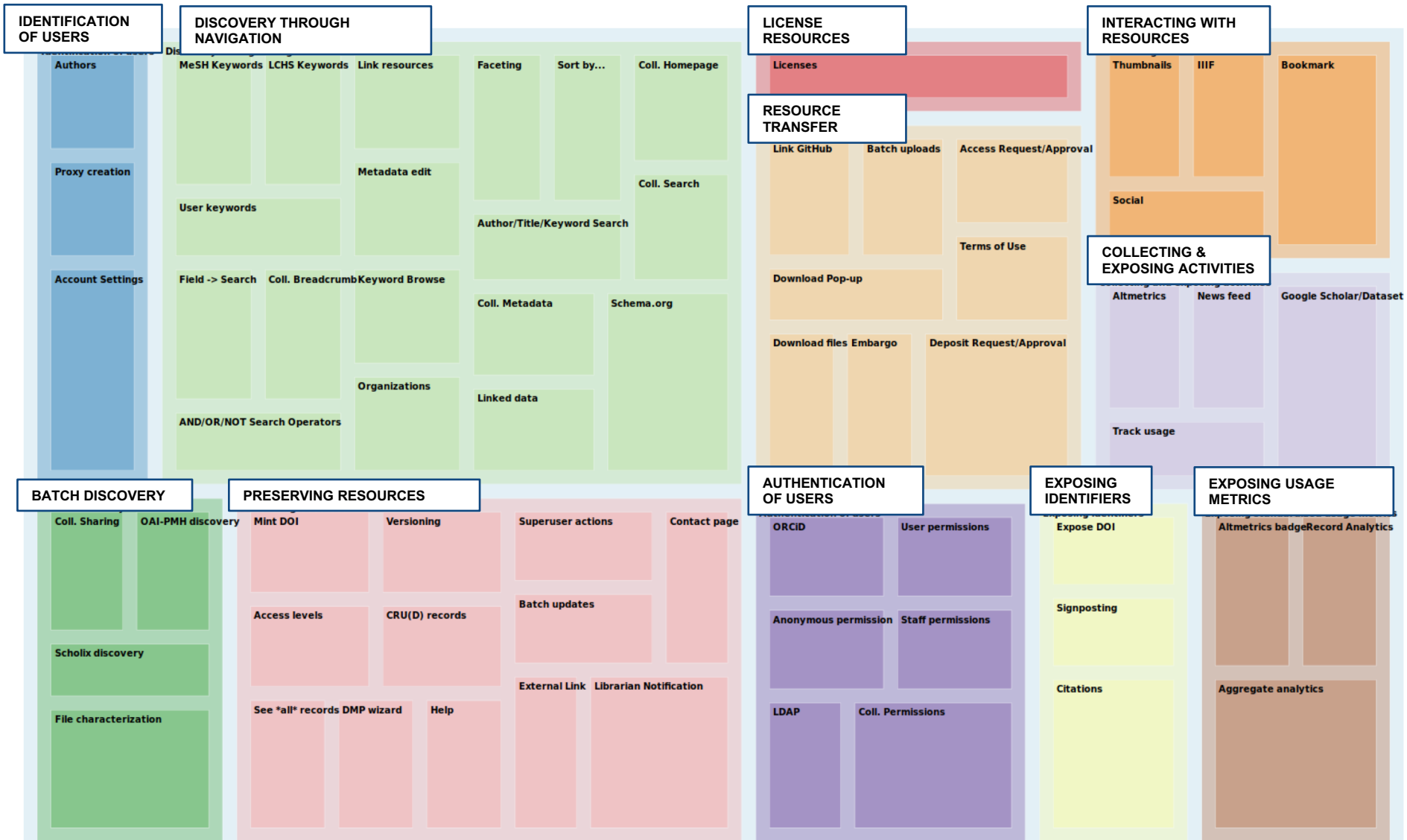


# Feature Comparison:

- 1<sup>st</sup> Gen repository (DigitalHub)
- Invenio 3
- Shared features
- New features in development



# Our own next-generation requirements mapped against the COAR's Behaviors and Technical Recommendations for NGRs



# Joining Invenio RDM (1-year project)

## Invenio RDM: a turn-key open source research data management platform

 [Lars Holm Nielsen](#)  Apr 29, 2019 

**CERN has partnered with 10 multidisciplinary institutions and companies to build a turn-key open source research data management platform called Invenio RDM, and grow a diverse community to sustain the platform.**

The Invenio RDM project is funded by the [CERN Knowledge Transfer Fund](#), as well as all the participating partners, including:

- [Brookhaven National Laboratory](#) (US)
- [Caltech Library](#) (US)
- [Data Futures](#) (UK)
- [Helmholtz Zentrum Dresden-Rossendorf](#) (DE)
- [Northwestern University](#) (US)
- [OpenAIRE](#) (GR)
- [TIND](#) (NO)
- [Tubitak](#) (TK)
- [University of Hamburg](#) (DE)
- [University of Münster](#) (DE)

The project has an ambitious one year schedule in which it will deliver:

- Invenio RDM - A research data management platform based on [Zenodo](#) and [Invenio v3 Framework](#).
- A community of public and private institutions to sustain Invenio RDM.
- Minimum two existing repositories migrated to Invenio RDM, with Zenodo being one of them.

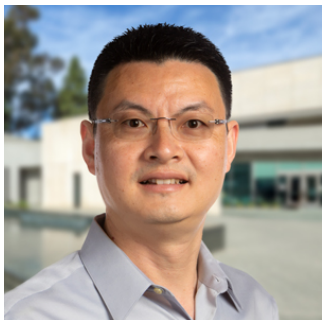
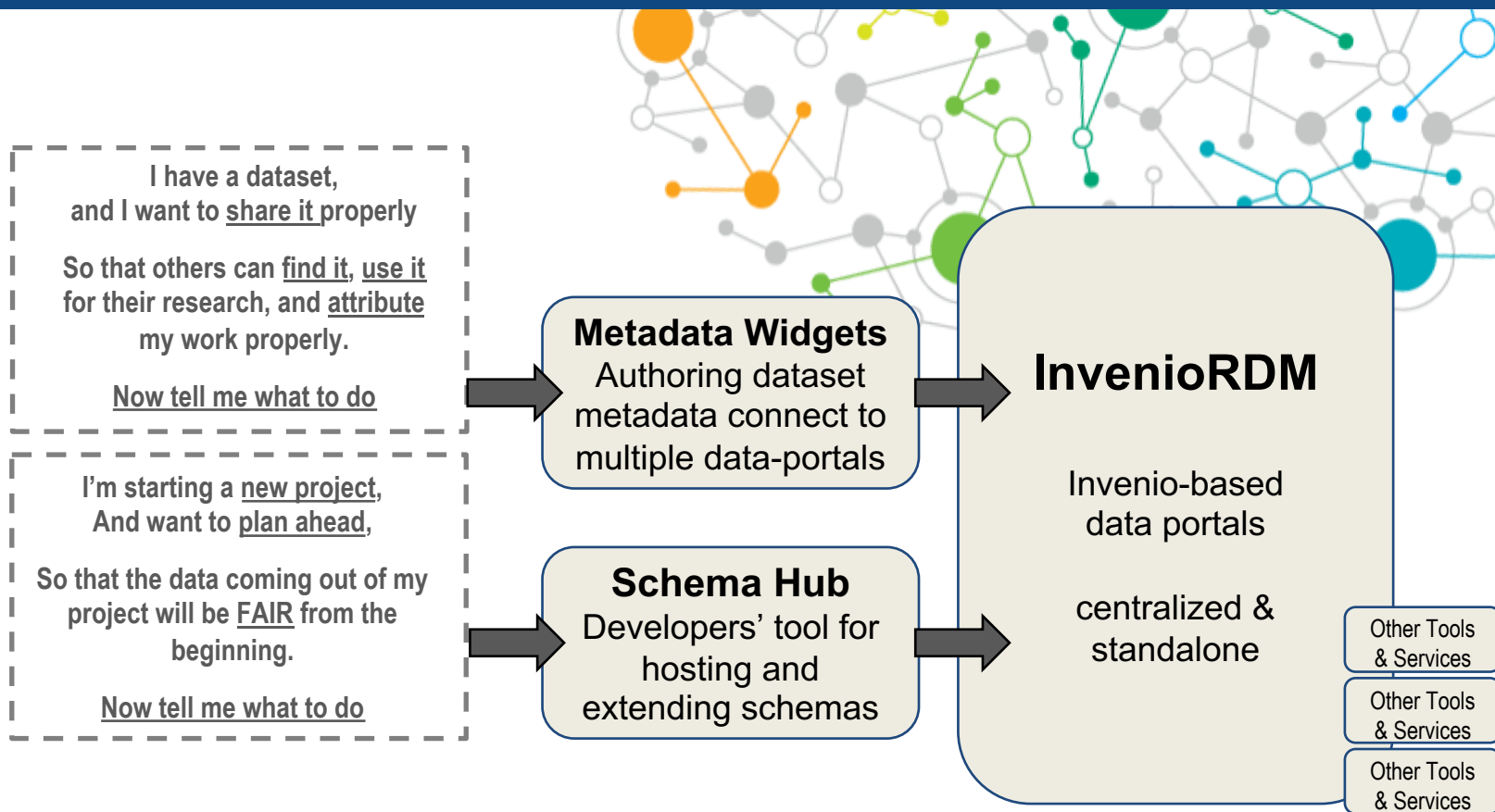
**For more information about the Invenio RDM project, please contact:**

Lars Holm Nielsen  
Invenio Product Manager  
CERN IT Department  
[info@inveniosoftware.org](mailto:info@inveniosoftware.org)

<https://invenio-software.org/blog/2019-04-29-rdm/>

**Next steps for CD2H.**

# Moving toward a Data Discovery Engine for biomedicine




Work with Chunlei Wu (Scripps)  
<https://wulab.io/>

*Invenio-RDM nodes and other trusted repositories can enable a distributed architecture. Empowering a range of existing repositories & data catalogs will result in a robust, collaborative community.*




# CD2H Project Highlight: Data Discovery Engine

*Make data more discoverable and reusable*



NATIONAL CENTER  
FOR DATA TO HEALTH

CD2H DATA DISCOVERY ENGINE



A PROJECT OF THE CD2H DATA WORKGROUP

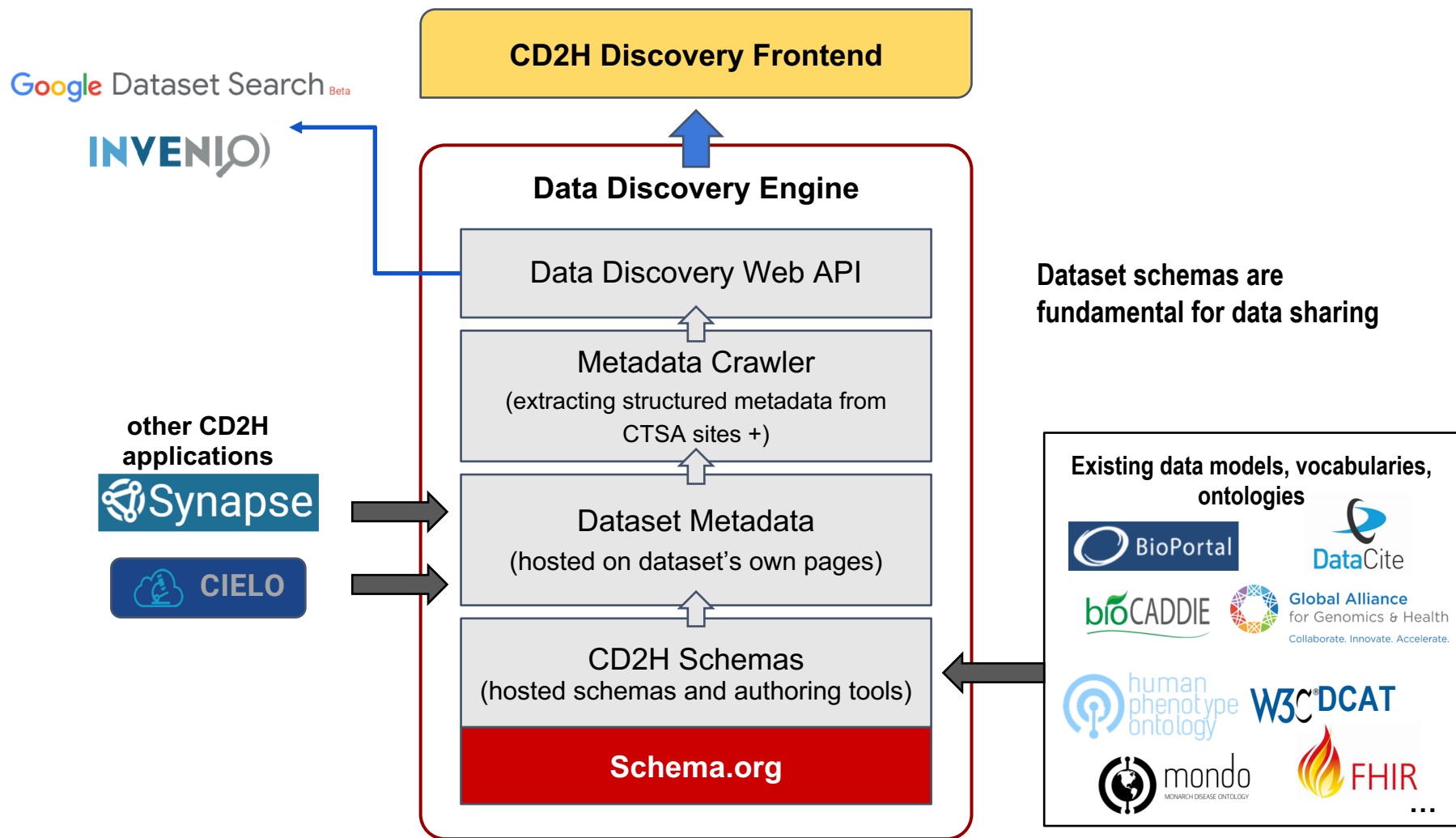
<http://discovery.biothings.io>

 **Scripps  
Research**

 **Northwestern Medicine**  
Feinberg School of Medicine



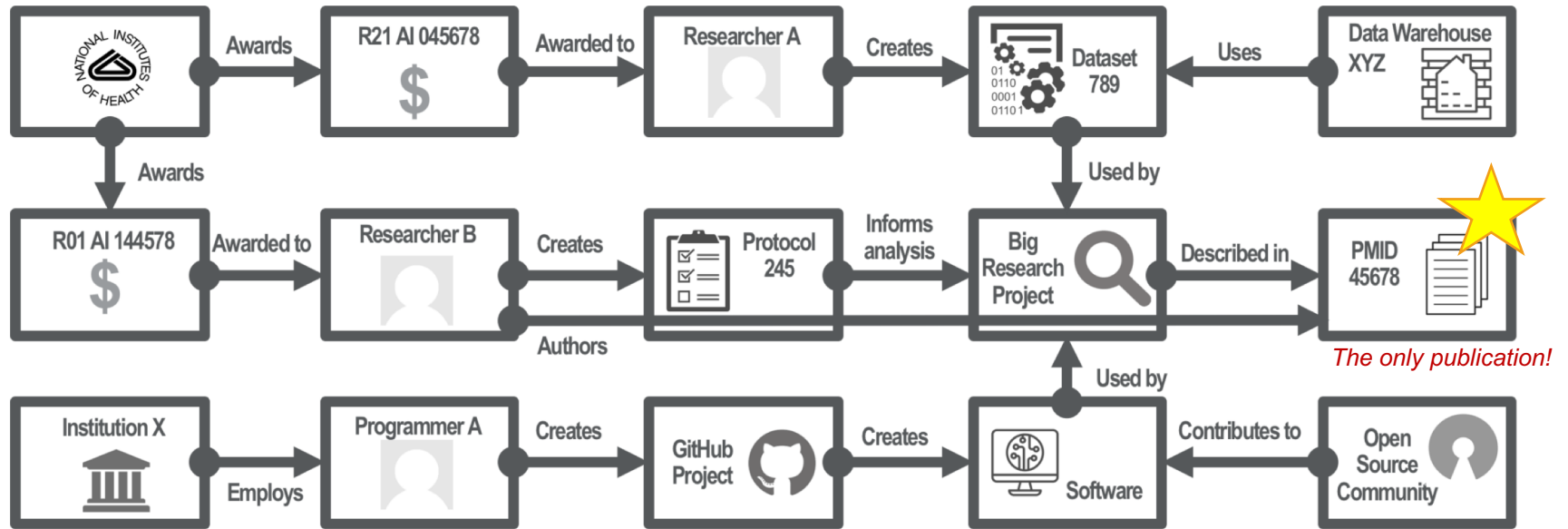
# CD2H Project Highlight: Data Discovery Engine



Prototype at: <http://discovery.biothings.io>

**One more note on technology + culture.**

# Architecting Attribution



Adapted from original by @figgyjam

A wide range of contributions beyond traditional papers are required to drive research. We're building on CRedIT and community input to make it possible to describe, give credit for, highlight the impact of non-traditional contributions to research

[https://github.com/data2health/architecting\\_attribution](https://github.com/data2health/architecting_attribution)

Interested in learning more or joining us? sign-up at the bottom of the page

# With thanks...



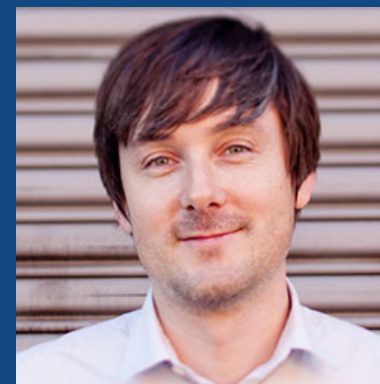
*Guillaume Viger*



*Sara Gonzales*



*Lisa O'Keefe*



*Matt Carson*

## Teams

- Galter Health Sciences Library & Learning Center
- Center for Data to Health (CD2H)
- Lars & the Invenio team
- Chunlei Wu, Scripps
- Northwestern University Clinical and Translational Sciences Institute (NUCATS)
- Collaborators: ChicagoCHEC, FIRST DailyLife, Health for All, OpenVIVO

## NIH Support

- U24TR002306 (NCATS)
- UL1TR001422 (NCATS)
- U54CA202995, U54CA202997, U54CA203000 (NCI)
- P30AR072579 (NIAMS)
- G08LM012688 (NLM)

COMMENT

Writing

Study conception

# Credit where credit is due

Investigation

Formal analysis

Liz Allen, Amy Brand, Jo Scott, Micah Altman and Marjorie Hlava are trialling digital taxonomies to help researchers to identify their contributions to collaborative projects.

Through the endorsement of individuals' journal articles could be classified using a 14-role taxonomy (see "Who did what?").

**Game changer:** Perhaps one of the biggest shifts in "culture" was the development, release, and implementation of the CRediT taxonomy, making it easier to give people credit for their specific contribution in a published work.



## CRediT

**CRediT ontology in OWL:**  
<https://github.com/data2health/credit-ontology>

CRediT is high-level taxonomy, including 14 roles, that can be used to represent the roles typically played by contributors to scientific scholarly output. The roles describe each contributor's specific contribution to the scholarly output.