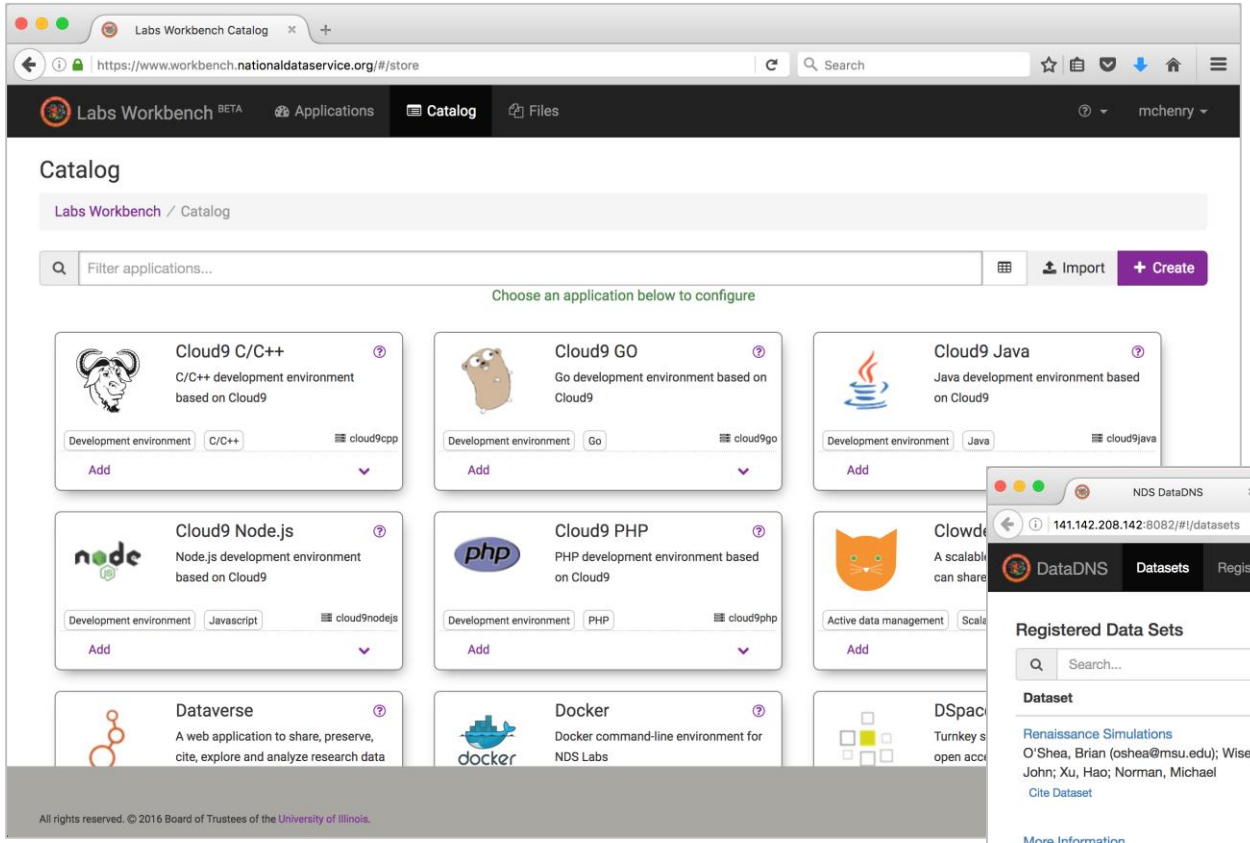
Enabling Open Science Without Impeding Open Science

Kenton McHenry
Technical Coordinator
National Data Service Consortium

Deputy Director
Scientific Software & Applications Division
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign

National Data Service Consortium

- Towards a world where it is easier to **publish**, **link**, **search**, and **reuse** data of all kinds
  - ***Advancing discovery*** by enabling open sharing of data
  - Increase collaboration within/across fields
  - Large-scale Data Service Interoperability
    - Distributed Storage & Computation
    - Spectrum of Services & Software
  - Incubator of Data Technologies, Projects, and Pilots

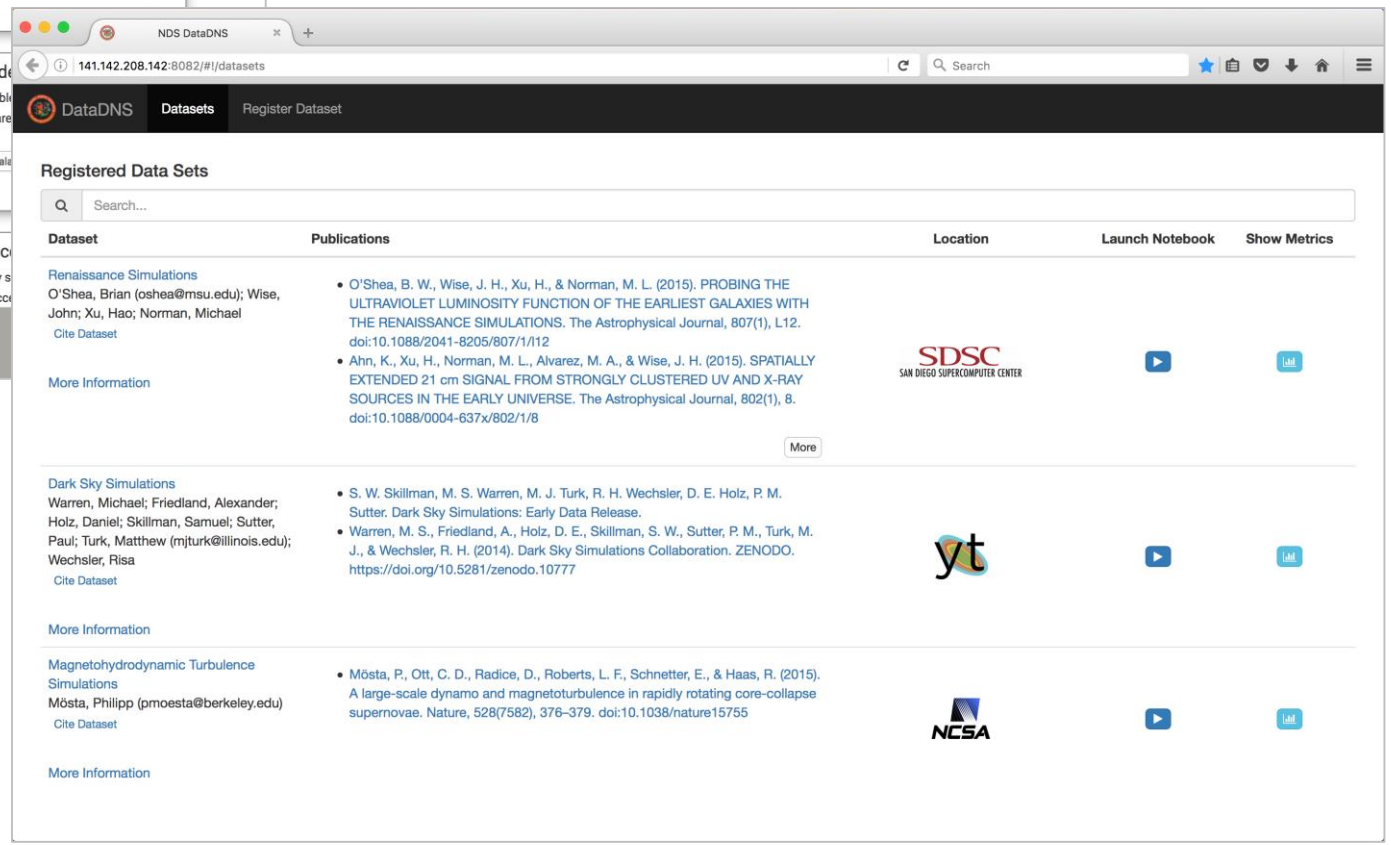


NDS Labs Workbench

- Incubate data technologies/projects
- Experiment with tools, perfect stack
- Run data science training environments
- Promote your data tools!

NDS Data DNS

- Share your data without moving it
- Contribute to reproducible science
- Invite new analysis
- Services for finding, indexing data



Examples of Requirements that Risk Impeding Open Science

- Sharing
  - Data Restrictions
  - Academic Value
  - Commercial Value

- Usability
  - Curation
  - Reusability
  - Storage

Data Restrictions

- Data that can not be shared broadly due to restrictions
  - e.g. human subject data

- ***Enabling Open Science***
  - Capture data in repositories with access restrictions
  - Secure facilities (e.g. HIPAA, FISMA)

- ***Impediment to Open Science***
  - Limited access to the data

→ ***Non-consumptive Data Analysis***

Academic Value

- Additional discoveries to be made with the data

- ***Enabling Open Science***
  - Embargo periods on the data

- ***Impediment to Open Science***
  - Repositories with little viewable data
  - Precarious for new repository technologies

Commercial Value

- Some aspect of the data or derivation from the data may have commercial value

- ***Enabling Open Science***
  - Capture in repositories with access restrictions

- ***Impediment to Open Science***
  - Repositories with little viewable data long term
  - Precarious for repository

Curation Overhead

- Organizing data, assigning metadata, describing data layout so others can use the data
  - Slow, tedious, not yet rewarded academically
- ***Enabling Open Science***
  - Creation of user friendly data management tools with collaborative/automatic curation support
- ***Impediment to Open Science***
  - MANY different tools
  - Redundancies
  - Useful features scattered across tools and communities
  - Competition in reaching critical mass of users

Reusability

- Necessary tools to access and use the data
  - Workflows, Indexing and search, analysis tools for unstructured data, analysis tools in general, transfer tools for large datasets, transformation tools, format specifications and data loaders, etc.
- ***Enabling Open Science***
  - Build widely accessible tools & services to support these needs
- ***Impediment to Open Science***
  - MANY different tools
  - Separating instances vs services
  - Long term viability

Storage

- Available and useful storage

- ***Enabling Open Science***
  - Tiered storage resources to meet budget constraints
  - Adjacent to HPC or cloud resources

- ***Impediment to Open Science***
  - Long term viability

Takeaways

- Uncertainty in the components of an open science enterprise, in one form or another, appears to be a significant impedance to open science
  - Security of data/intellectual property
  - Long term viability

- Some of this uncertainty may be avoidable!