Piloting a Data Publication Service for the BD2K Commons

Ian Foster, Kyle Chard, Aditya Parameswaran, Ivo Dinov, Ben Heavner, Gustavo Glusman, Mike D'Arcy, Ravi Madduri, Judy Pa, Carl Kesselman, Eric Deutsch, Nathan Price, John Van Horn, Joseph Ames, Kristi Clark, Leroy Hood, Jiawei Han, and Arthur Toga

University of Chicago, University of Illinois at Urbana-Champaign, University of Southern California, Institute for Systems Biology, University of Michigan

ark:/88120/r8h59h

Abstract

Objective: Develop a scalable cloud-hosted data publication system for the BD2K Commons.

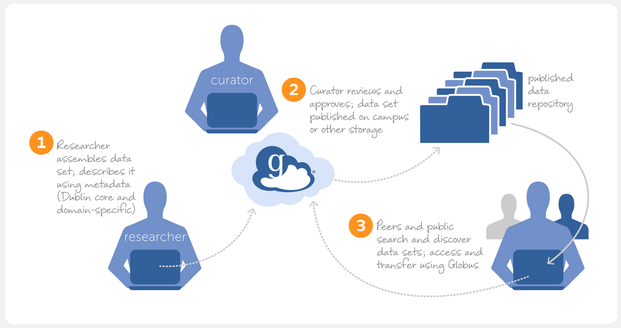
Approach: Integrate BDDS data publication capabilities, KnowEng data preparation capabilities, and National Data Service (NDS) infrastructure and capabilities.

How it works

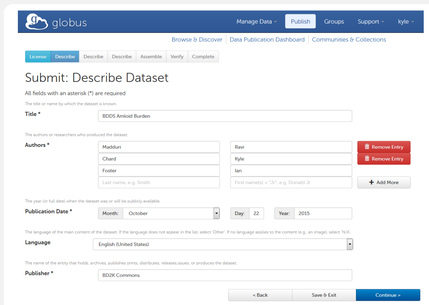
Publication capabilities are delivered through a hosted service with metadata stored and indexed in the cloud and data storage provided by a specified remote storage provider made accessible via Globus.

Published datasets are organized by "communities" and their member "collections". A variety of specific policies can be set on communities or collections to manage:

- Metadata (schema, requirements)
- Access control (user and group based)
- Submission and curation workflows
- Submission and distribution license
- Storage endpoint
- Persistent identifier provider (DOI, Handle)



Data Publication

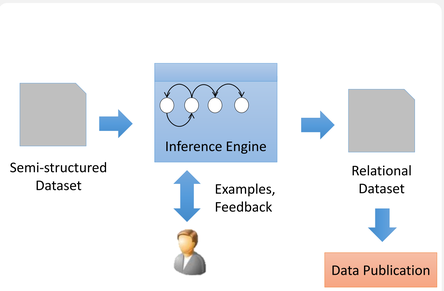


1. Dataset Description

- Describe your dataset with metadata from standard schemas (e.g., DataCite) or custom domain-specific schemas
- All metadata indexed for future discovery

2. Metadata/Structure Extraction

- Extract relational data using user-provided examples
- HMM (Hidden-Markov-Model)-based technique
- Inconsistencies or noise identified as by-product of output

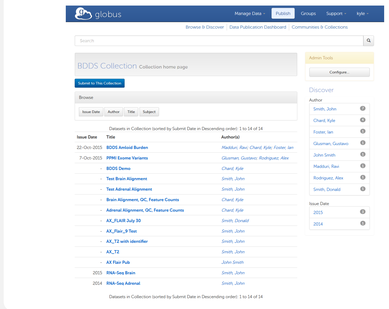


3. Dataset Assembly

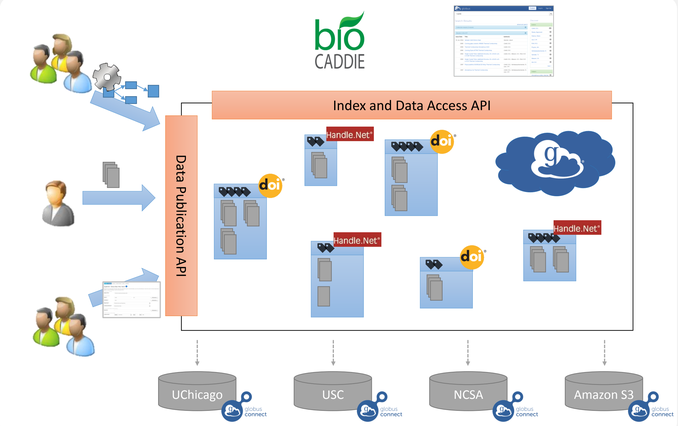
- Asynchronously assemble large datasets using Globus transfer
- Supports collaborative assembly
- Restricted data access for submission and curation

4. Dataset Discovery

- Discover datasets with free text search
- Facets generated to drill down into search results
- Search results link to datasets for easy retrieval



Publication Model



Capabilities

Publish large datasets
- Leverage Globus for file transfer, identities, and groups
- Handle large datasets with ease
- Use local or institutional storage

Customizable metadata descriptions
- Build custom metadata schema for specific research data
- Re-use and import existing metadata schemas

Flexible data sharing
- Share with individuals, groups, or the public
- Change access dynamically for one or more published datasets

Customizable user-oriented workflows
- Customize submission workflows with arbitrary steps
- Define different curation workflows

Arbitrary unique identifiers
- Associate unique identifiers with datasets (e.g., DOI)
- Improve dataset discovery and citability

Rich discovery support
- Search on standard, custom, and file metadata
- Goal: intuitive, "Google-like" search for biomedical data