The NIEHS Data Commons

Deep Patel, Mike Conway
Office of Data Science
National Institute of Environmental Health Sciences

The NIEHS Office of Data Science

Who are we?

"The mission of the Office of Data Science is to accelerate scientific discovery, foster collaborative research, and ultimately improve public health through the application of scientific data and knowledge management in the environmental health sciences."

Commons objectives

Develop a standards-based commons

- Beginning with internal researchers, managing data originating from core laboratories, including next-gen sequencing data.
  - Define organizational policies to handle data life-cycle
  - Track provenance and relationship of data sets to source data and analysis

Commons objectives

Manage metadata for discoverability and long-term usability

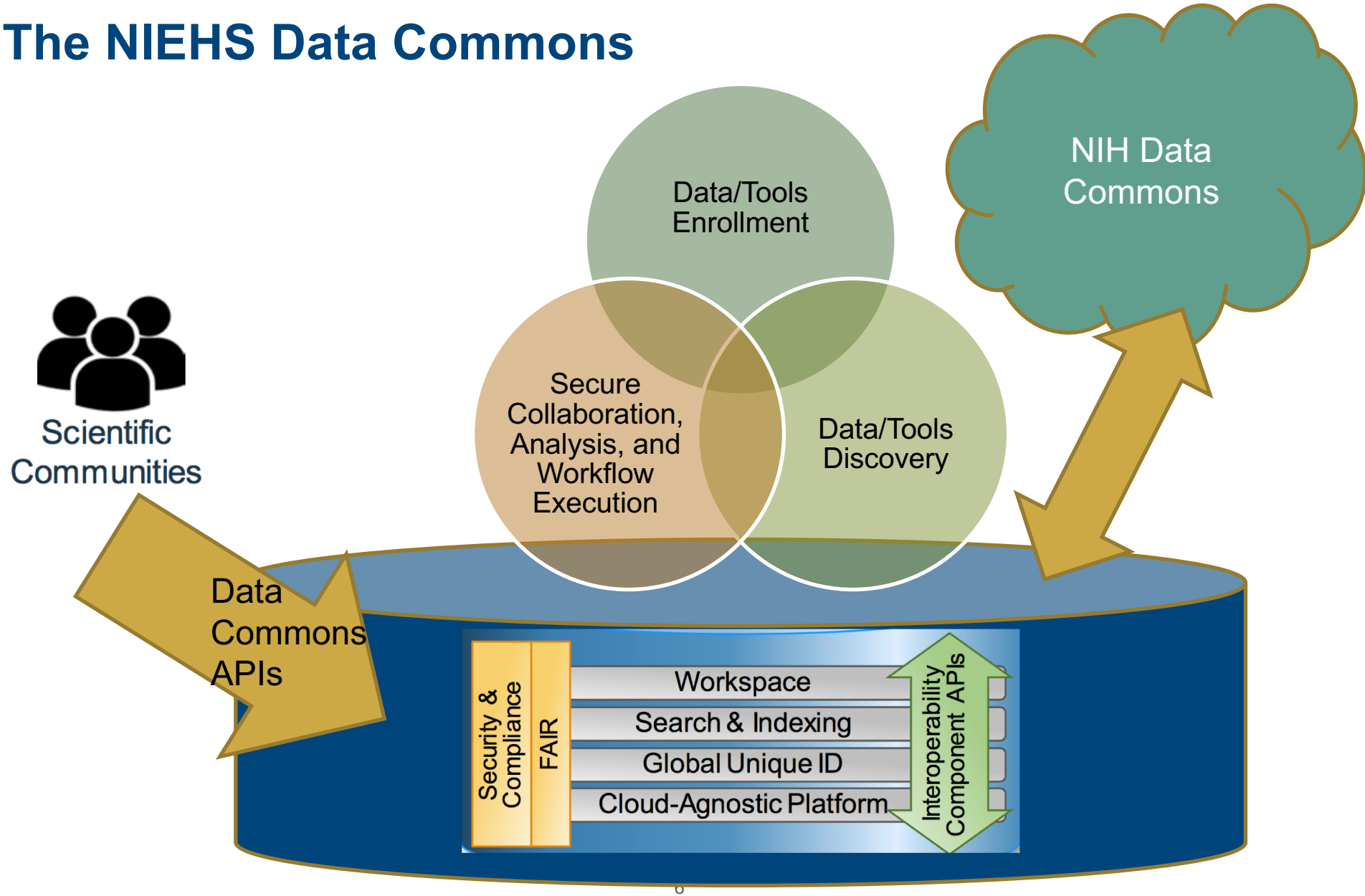
- FAIR Data
- Develop standard metadata, including controlled vocabularies and ontologies
- Automatic metadata from instruments, pipelines, and computer-actionable policies
- Support multiple indexes and search technologies for data discovery and re-use
- Allow publication to reference collections, such as NCBI GEO

Commons objectives

Support integration and use of data in computation and analysis

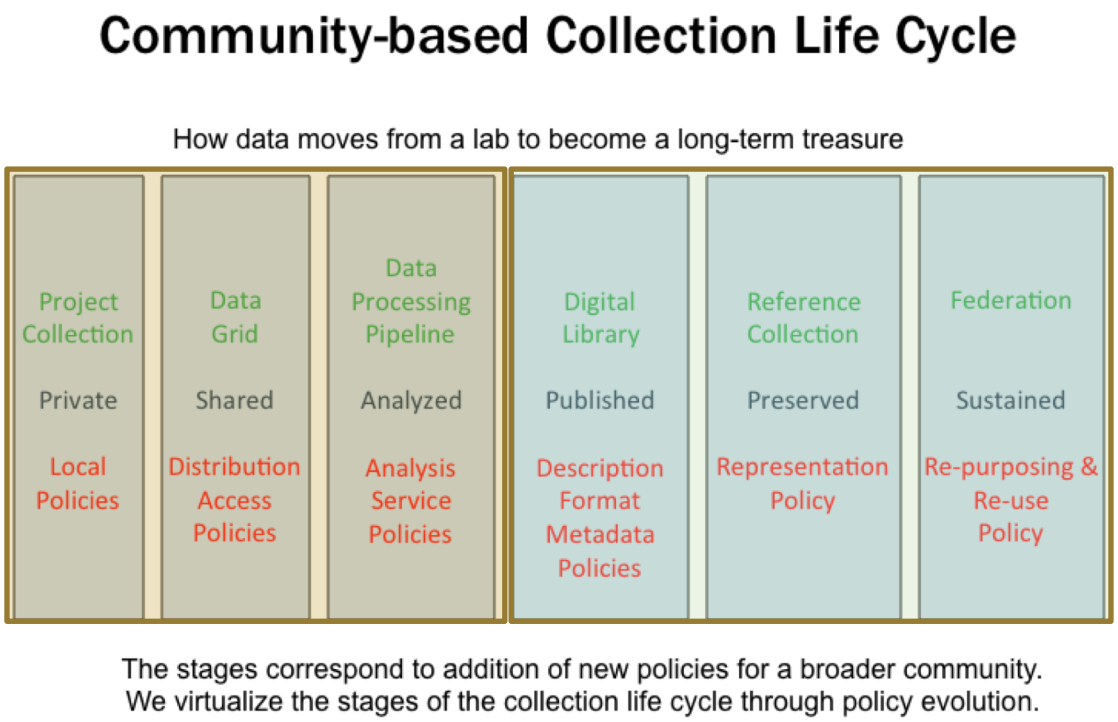
- Ease discovery and access through common tools and platforms
- Securely share data with collaborators
  - Allow audit and enforcement of access and data usage agreements
- Track provenance and authenticity
- Ensure reproducibility

The NIEHS Data Commons



Data Commons serving a full data life-cycle

Current 'commons' efforts (e.g. NIH Commons) focus on the mature part of the research data lifecycle and say less about where the data comes from!



Moore, Reagan W., et al. "White Paper: National Data Infrastructure for Earth System Science."

NIEHS Concerns:

- Metadata quality
- Delivery to PI
- Appropriate sharing within project
- Retention, compliance
- Ingest pipelines

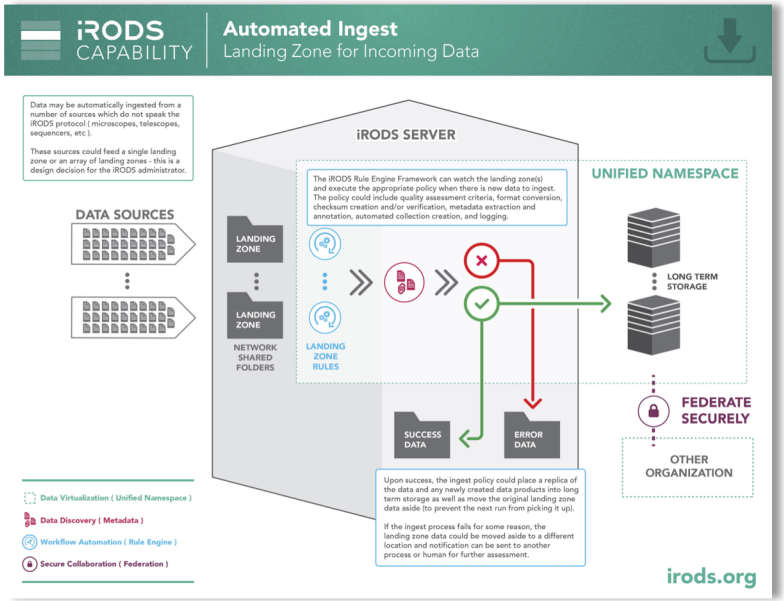
NIH Concerns:

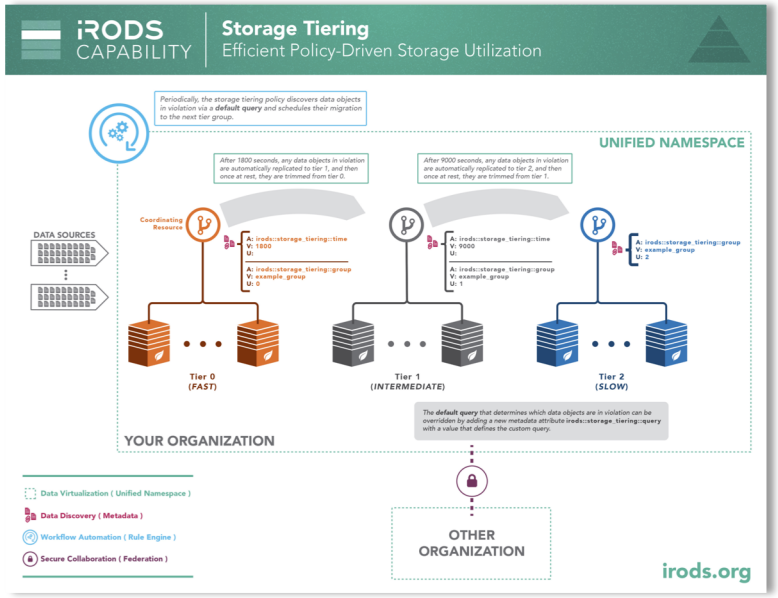
- FAIR
- Publishing
- Data sharing/licensing
- Discoverability
- Analytics and derived data

Commons 'Patterns'

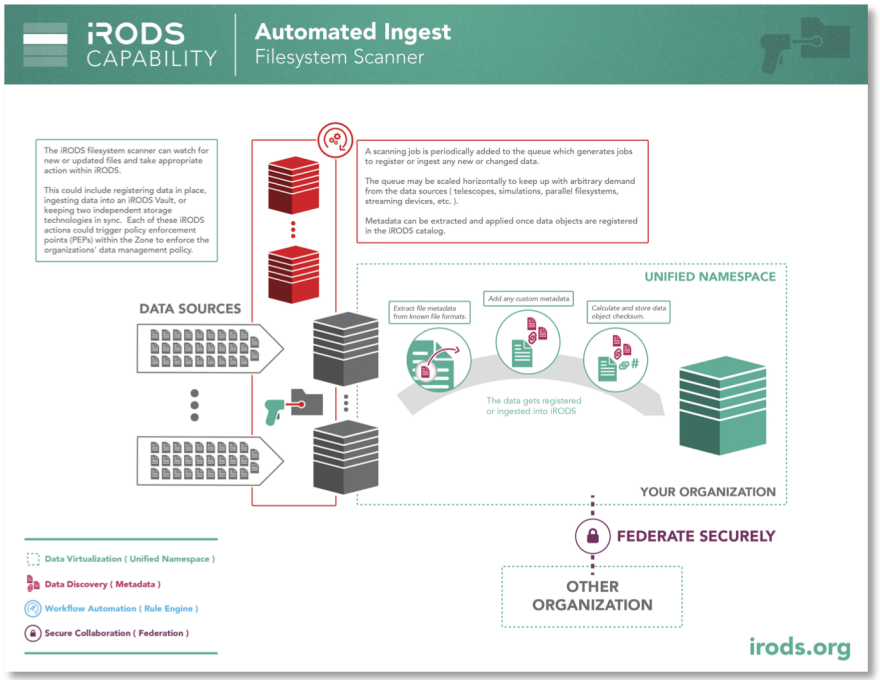
- Let's look at the NIEHS Commons and see where patterns come into play.
  - How do we as a community develop frameworks around iRODS capabilities and the philosophy of policy-based data management that ease development?
  - How do we develop a pattern language and architectural discipline and talk with each other about systems that support FAIR and Big Data?
  - The Consortium is already developing a pattern catalog, and this is a **Good.Thing**.

Extracting Patterns…



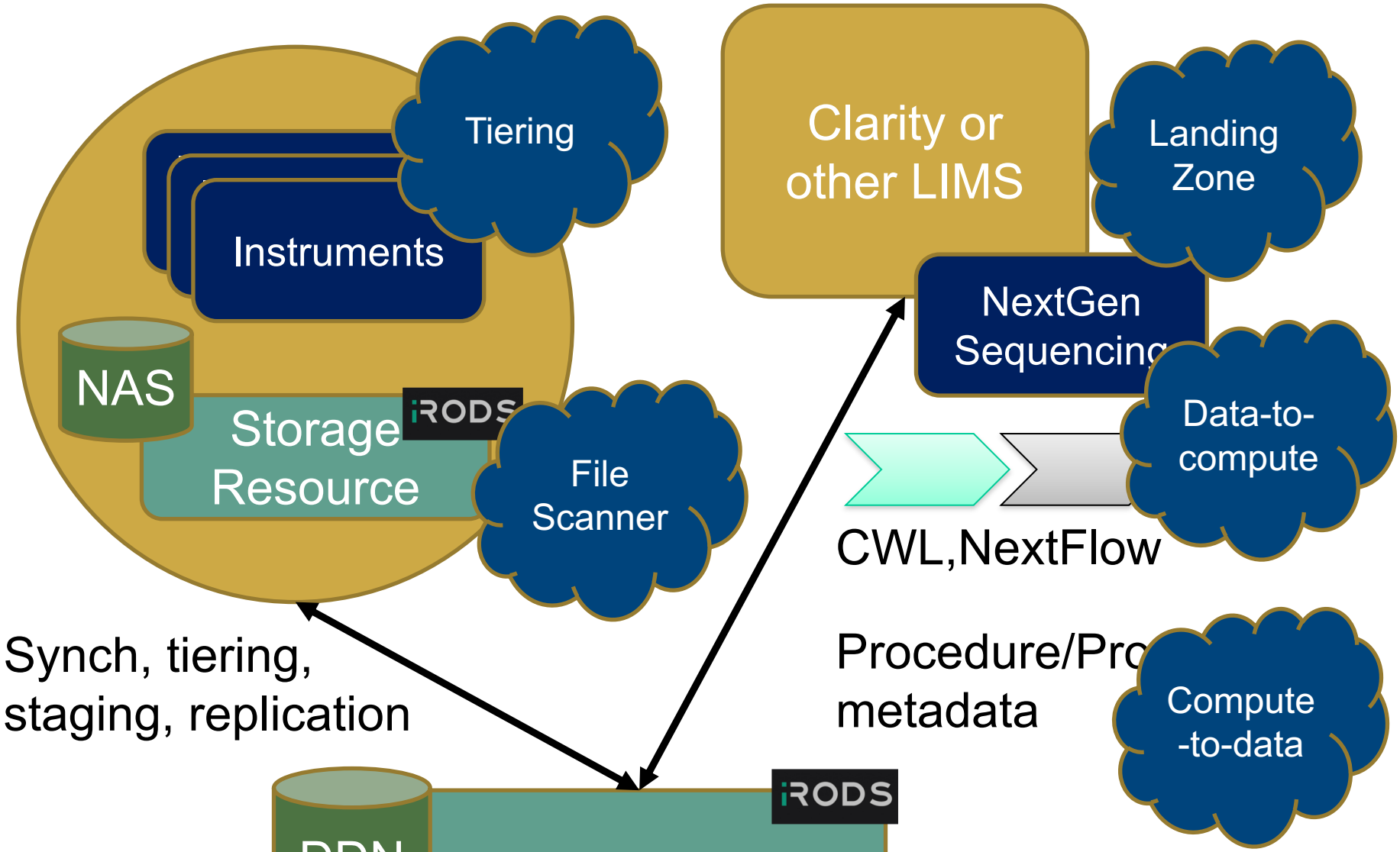


- Shout out to the Consortium folks, this may be the ‘next thing’.
- How would a good catalog of patterns translate into frameworks and capabilities in iRODS?

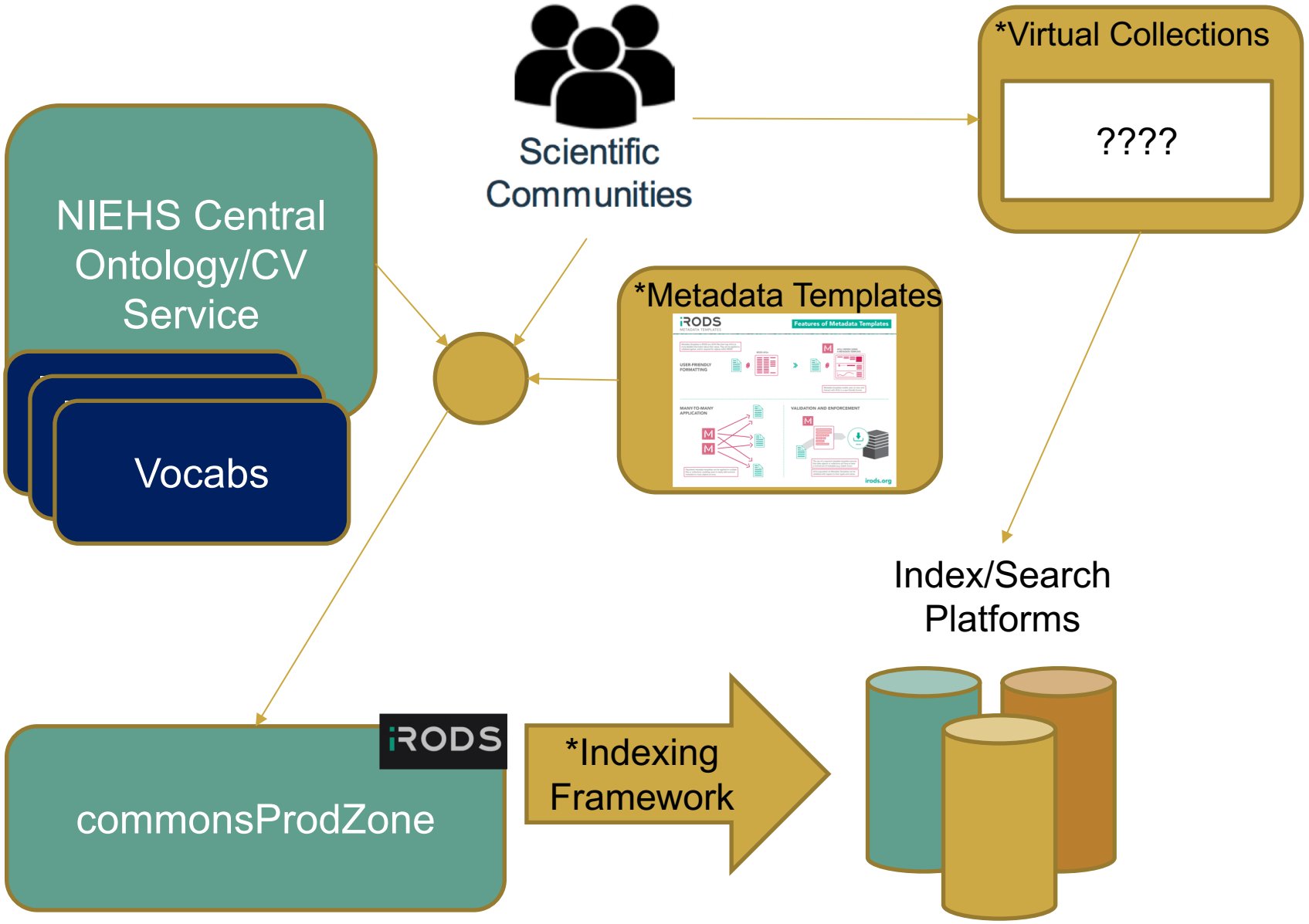


Patterns from https://irods.org/documentation/

Core Labs Ingest and Pipelines



Metadata Support



New Challenges

- Managing immutable archives (e.g. BDBag) and persistent identifiers
- Managing federated authn/authz
- Integrating the Data Commons into the workflows and daily routines of researchers in non-disruptive ways that make their work easier, not more difficult
- Keeping the focus on science, not cyberinfrastructure

Big Data is Big Preservation

- Let's not forget our roots, and how this applies now more than ever.
- OAIS and related concepts, including trusted digital preservation provide lots of useful language and a good conceptual framework to add to the 'cloud', 'FAIR', and NSF 'Cyberinfrastructure for the 21st Century' frame.
- FAIR does not matter if the data turns out to be lost!

Acknowlegements

NIEHS: Beth Bowden, John Bucher, Allen Dearry, Leesa Deterding, Michael Devito, Christopher Duncan, Matthew Edin, Thomas Van'T Erve, John Grovenstein, Guang Hu, Mary Jacobson, Jeffrey Kuhn, Beth Lauderdale, Jian-Liang Li, Alex Merrick, Geoffrey Mueller, Suzanne Osborne, Scott Redman, Andy Shapiro, Troy Simpson, Chris Stone, Cheryl Thompson, Paul Wade, Deborah Wales, Jason Williams, Rick Woychik;

Renaissance Computing Institute: (RENCI);

iRODS Consortium