The iPlant Data Commons

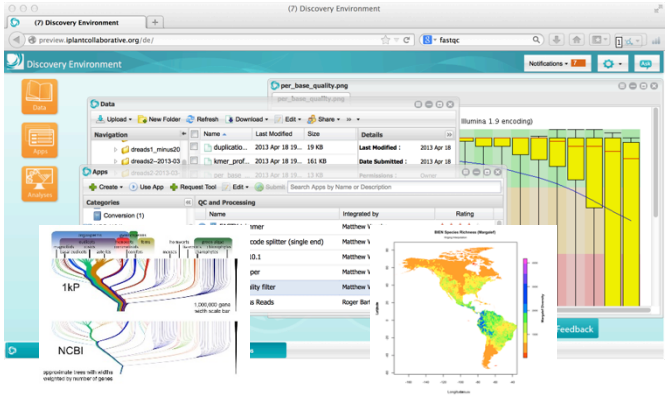
Using iRODS to Facilitate Data Dissemination, Discovery, and Reproducibility

Jeremy DeBarry, jdebarry@iplantcollaborative.org
Tony Edgin, tedgin@iplantcollaborative.org
Nirav Merchant, nirav@iplantcollaborative.org

The iPlant Collaborative

The iPlant Collaborative

We are a Cyberinfrastructure



Platforms, tools, datasets

Storage and compute

Training and support

The iPlant Collaborative

And a virtual organization

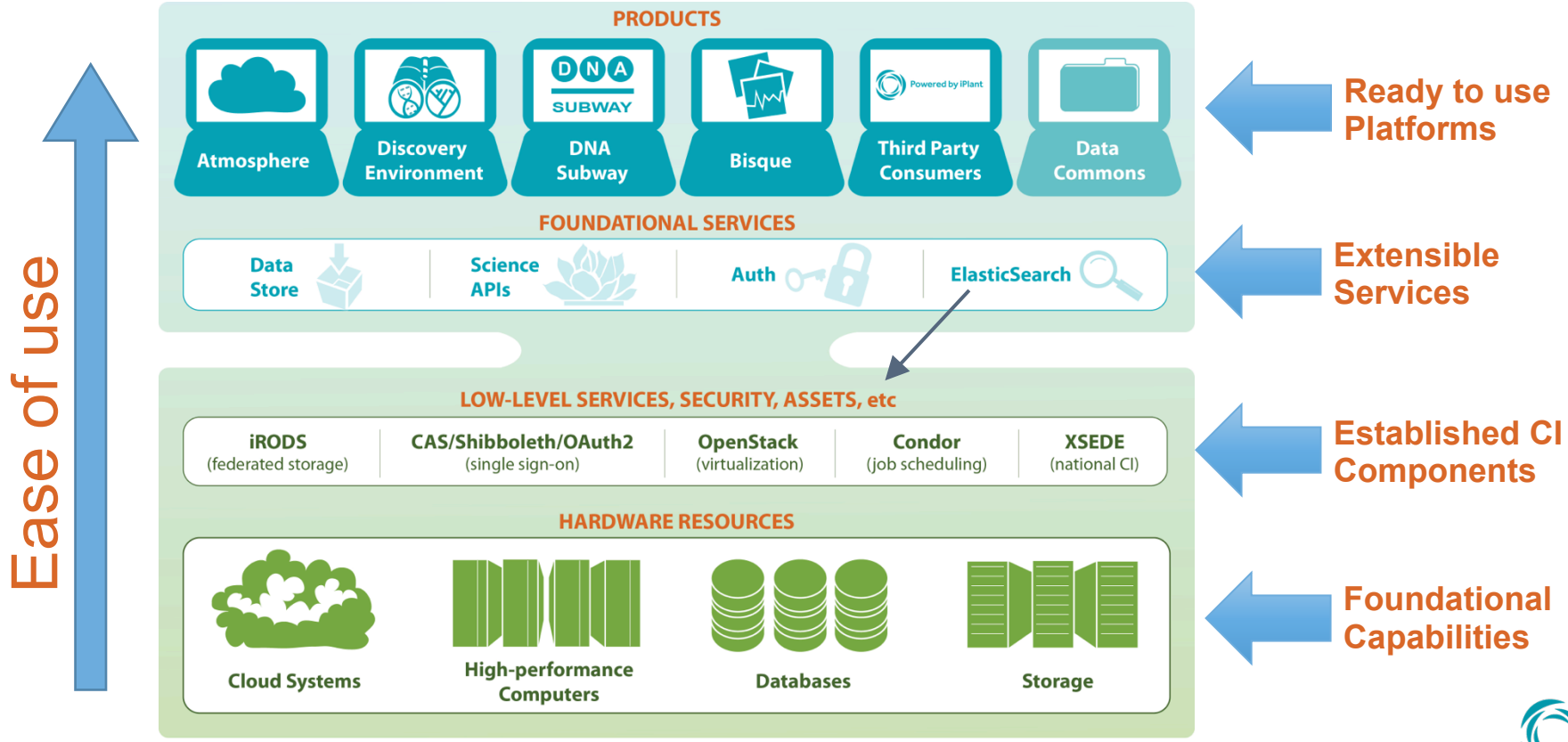
Developer Expertise
Computational Capacity
Science Domain Expertise
Training
Administrative and Organization

Why Data Commons (by Phil Bourne)

- The Commons is a pilot experiment in the efficient storage, manipulation, analysis, and sharing of research output, **from all parts** of the research lifecycle.
- Should The Commons be successful we would achieve a level of **comprehensive access and interoperability** across the research enterprise far beyond what is possible today.
- Some key attributes:
  - *Should support Sharing & Accessibility*
  - *International to be maximally successful*
  - *Should allow data science become more cost-effective, hence more sustainable*
  - *Replicability, opportunity and ability to reproduce, or at least replicate, experiments*
  - *Discoverability of research output through indexing or other methods*

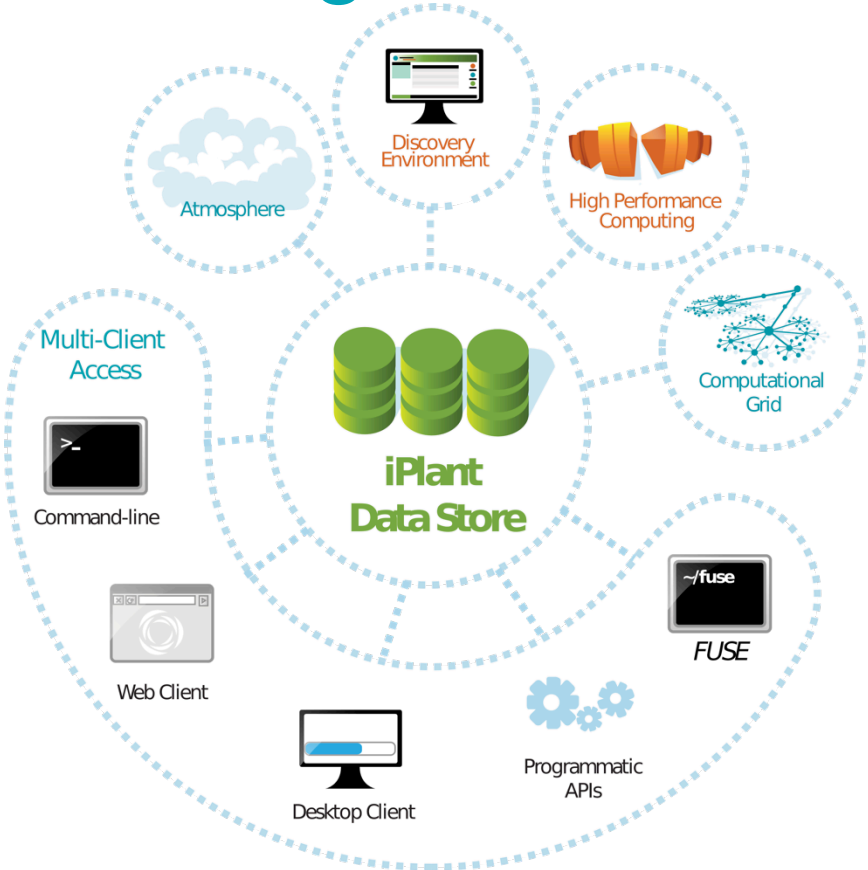
Phil Bourne is the Associate Director of Data Science (ADDS) at NIH
https://pebourne.wordpress.com/2014/10/07/the-commons/

iPlant Collaborative Products



Data Store: Sharing Data

- peer to peer
  - ACL based
  - iRODS tickets
  - public links, anonymous read-only access through URLs
- Community Data, large dataset sharing

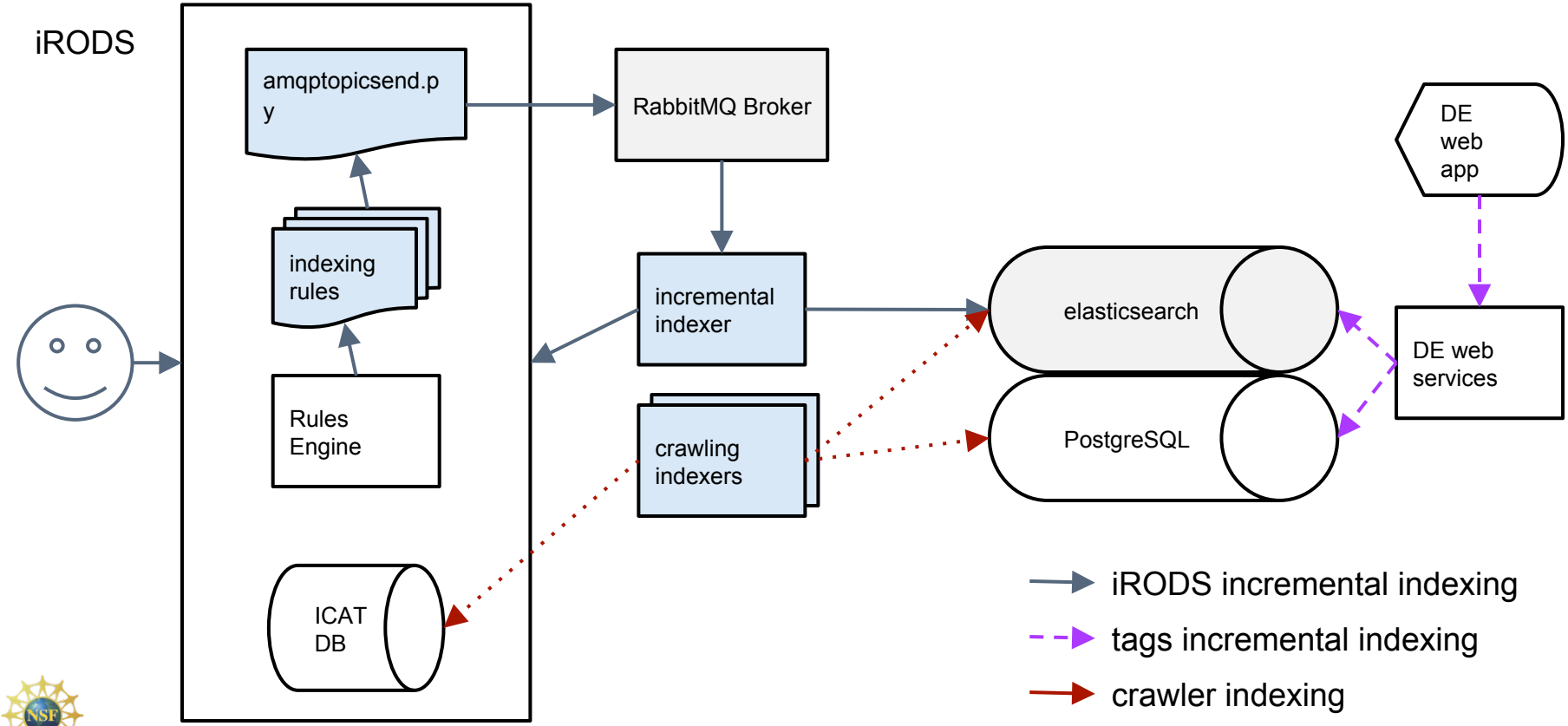


Sharing Usage Statistics

- User Statistics
  - 27000 user accounts
  - 4900 users with data
  - 2600 users (53% of users with data) made at least 1 share
  - 2100 shares per user
  - 42 million files (58% shared)
  - 59 million (1.1 million/month) shares
- Community Data Statistics
  - 5 million files
  - 55 million (1.0 million/month) shares
- ~1PB of data

A share is an instance of a user sharing a single file with a single user or group.

Indexing and Search



Data Analysis in the DE

- Data analysis tools are installed on Condor cluster or HPC. (user can request new tools)
- An *app* is an analysis template for a tool usable from the DE. (user can create and publish new apps)
- A *workflow* combines apps into an analysis pipeline (user can create and publish new workflows).
- Batches of files may be processed by apps and workflows.
- Ability to annotate results with metadata, comments and relationships to support reproducibility and manage large data sets

Current Tool Deployment

- Results reproduction is difficult outside of our infrastructure.
- Updates can make results reproduction impossible.
- Deployment requires support staff.
- Software conflicts may make certain tool combinations impossible.

Containerized Tool Deployment*
(or Docker to the rescue!)

- Containers are portable, so results can be reproduced outside of CI.
- Different versions of tool are placed in different containers. Old tool results are still reproducible after an upgrade.
- Users can bring custom tools in own containers.
- No conflicts occur between tools, since each in its own container.

** Coming July 2015*

The iPlant Data Commons

- It is a part of the Data Store for high-value, public datasets with links to external repositories.
- Its data are available throughout the CI.
- Its data are searchable and discoverable.
- Its data will be available and useful to the community, not buried in Data Store.

Data Commons Project Management

- Associate collaborators with a project.
- Add data and metadata to a project
- Organize data using standardized and project-specific metadata.
- Suggest analyses based on data type.
- Track history of operations. (provenance)

Data Commons Staging Area

- Distill project artifacts into package for publication to Data Commons and external repository.
- Combine and edit project metadata.
- Select appropriate licenses and persistent ids for the data and for the chosen repository.

Acknowledgements

- This project is funded by NSF.
- Thanks to all of the staff at the iPlant Collaborative.
- The Data Commons would not exist without the efforts of Kapeel Chougule, Jeremy DeBarry, Maria Esteva, Matthew Hanlon, Nirav Merchant, Stephen Mock, Sriram Srinivasan, Ramona Walls, and Liya Wang.

If you have additional questions, please don't hesitate to contact me at tedgin@iplantcollaborative.org.