Pragmatic approaches for enabling data driven collaborations for plant sciences & beyond

Andrew Lenards, Edwin Skidmore
iPlant Collaborative

iPlant Collaborative™

Background: What is iPlant?

Funded by US National Science Foundation

Building a comprehensive informatics cyberinfrastructure for plant biology;

Lately, also support animal research.

iPlant Collaborative™

Background: What is iPlant?

Ecosystem of services and applications: web portals, APIs, HPC, cloud, and of course, data storage.

One of iPlant's primary goals:

Minimized the emphasis on technology.
Return the focus to **biology** and **scientific discovery**

iPlant Collaborative™

iPlant Data Store

Different Users,
Different Access Needs:
One Data Store

iPlant Collaborative™

Design Decisions

iPlant Collaborative™

Data Management

- Supporting the full lifecycle of data
- From inception, analysis, collaboration and publication for multiple data types
- Emphasis on scalability, reliability, federation
- Present a consistent view w/ all clients (i-commands, iDrop, iDrop Lite, WebDAV, & iPlant tools)

iPlant Collaborative™

Data Management

- Integrate with external systems (provenance)
- Ensure metadata is first class citizen of the infrastructure across all systems
- Provide multiple modes of access to data
- Promote and support the use standards compliant metadata (but offer flexibility)

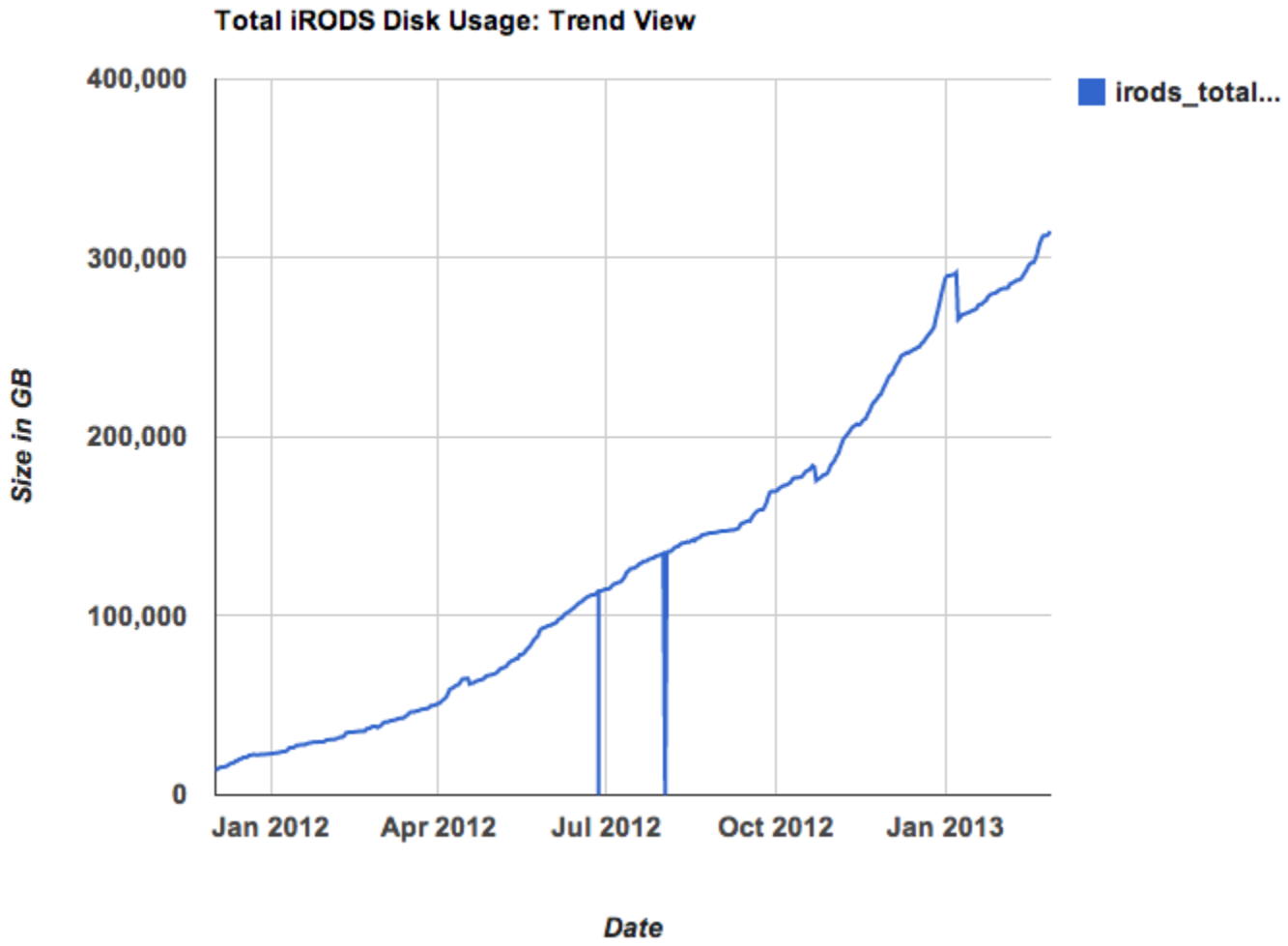
iPlant Collaborative™

Deployment / System Metrics

- users: 9900 users
  - a growth of about 350 new users per month
- data object count: 68M objects
- total data size: 313TB
  - 1TB per day growth
- deployment information
  - one iCAT
  - one database server
    - will be setting up a master-slave with SSDs and partitioning in March
  - several resource servers
  - one resource server for mirroring at TACC

iPlant Collaborative™

Total Disk Usage: 313 TB



iPlant Collaborative™

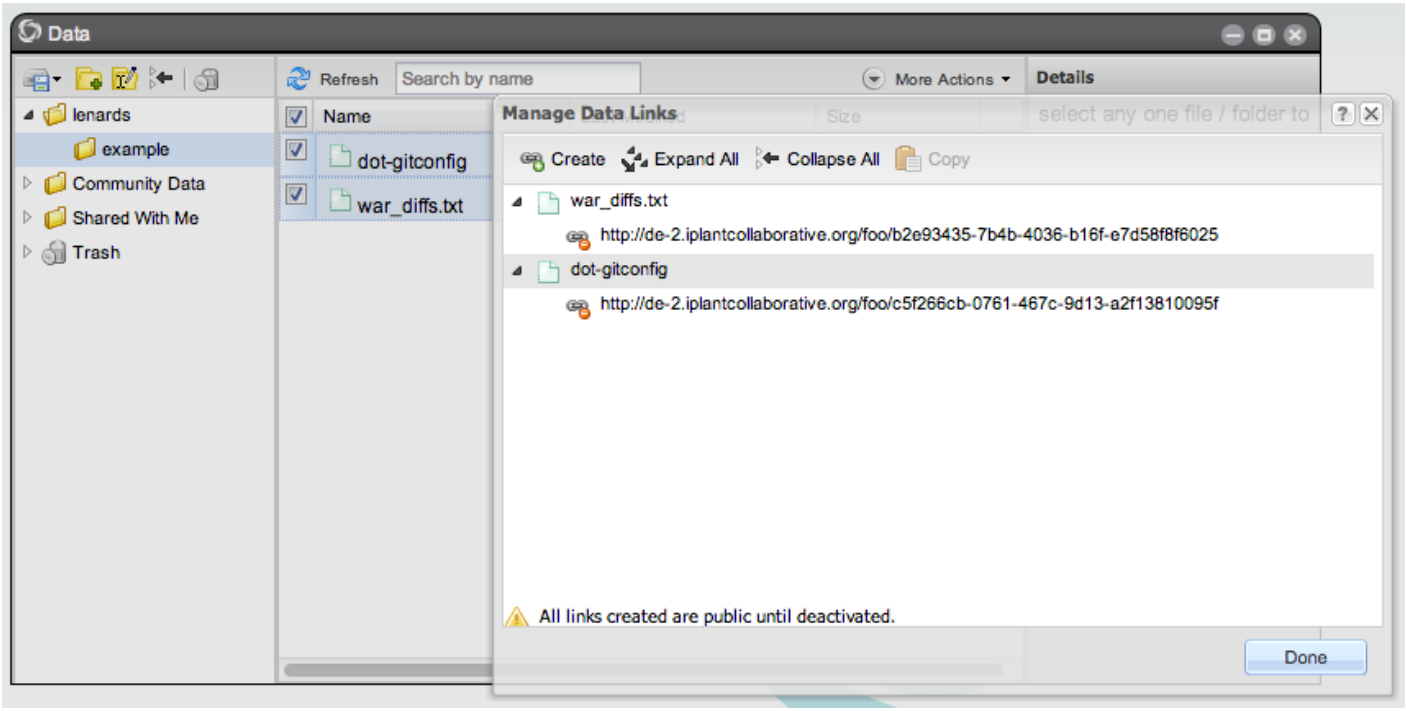
Features: Collaboration

- Data Links (the concept, using tickets, and the web landing page)
- Sharing
- Using a convention for community shared folders
  - e.g. /iplant/home/shared
- Rules for "Powered by iPlant" partners (bisque, coge etc)

iPlant Collaborative™

Features: Data Links

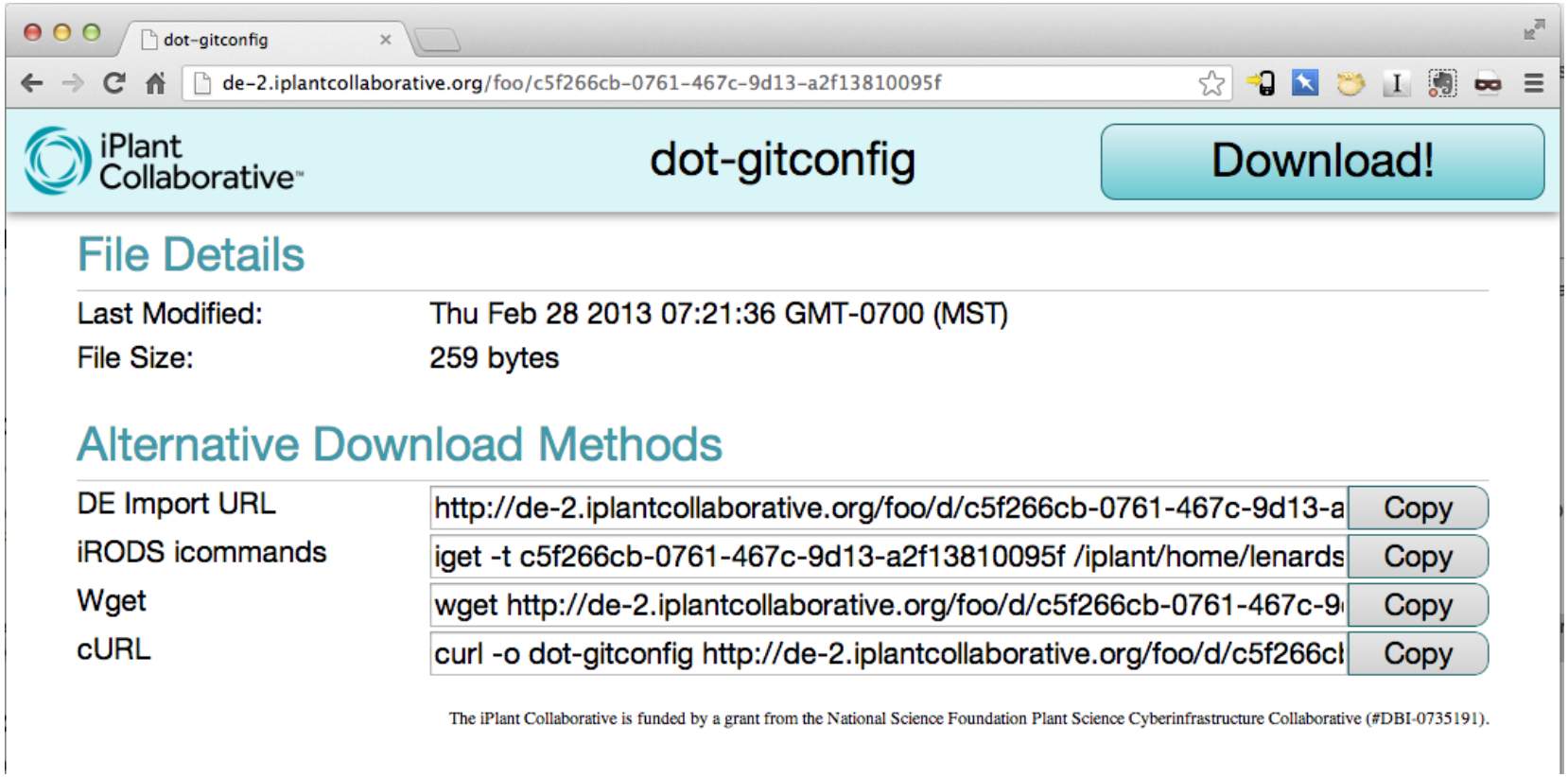
- Concept: provide a URL reference for fetching data
- How: using the 'tickets' feature from iRODS 3.1



iPlant Collaborative™

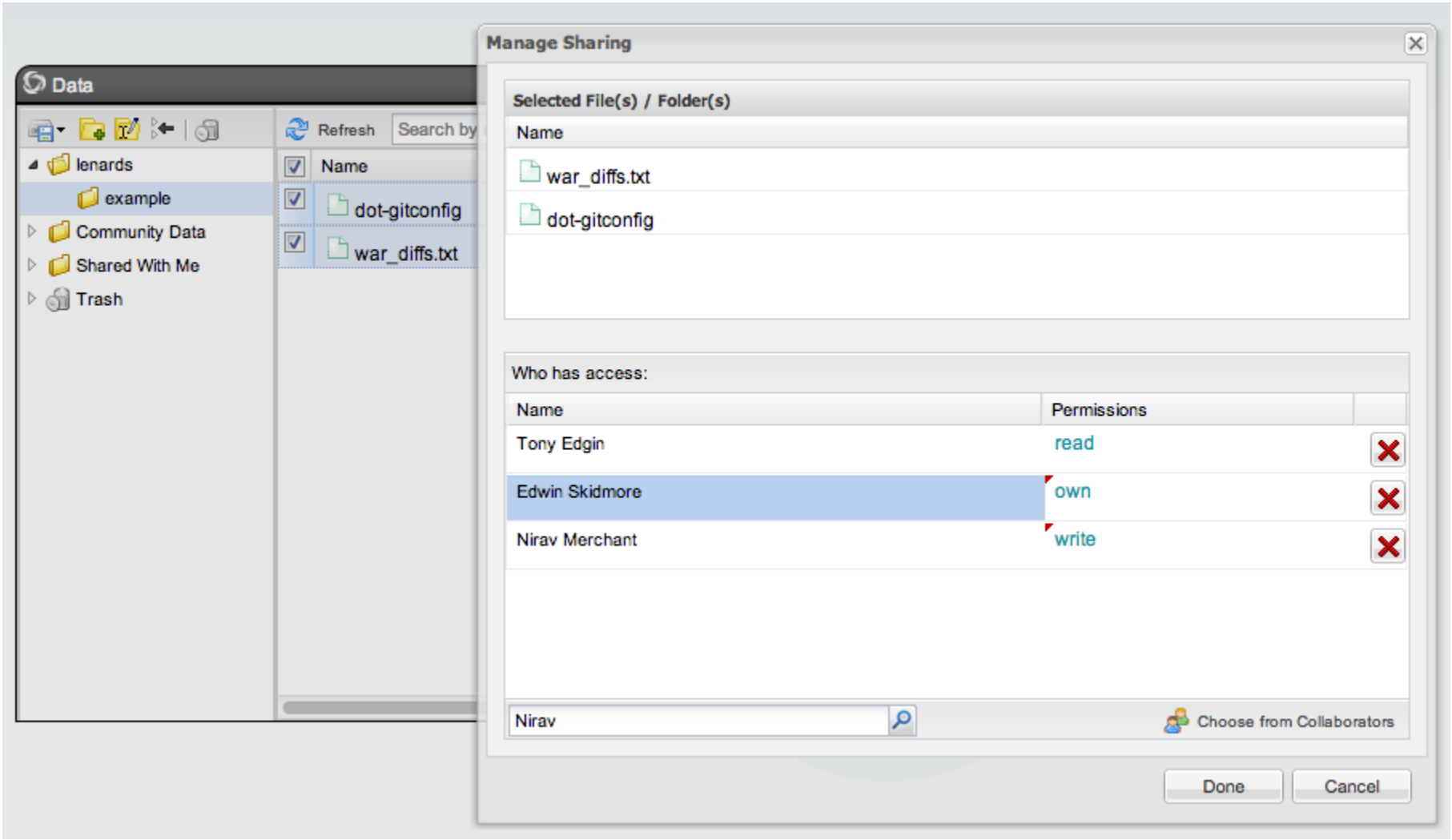
Features: Data Links

- In Action: and the web landing page



iPlant Collaborative™

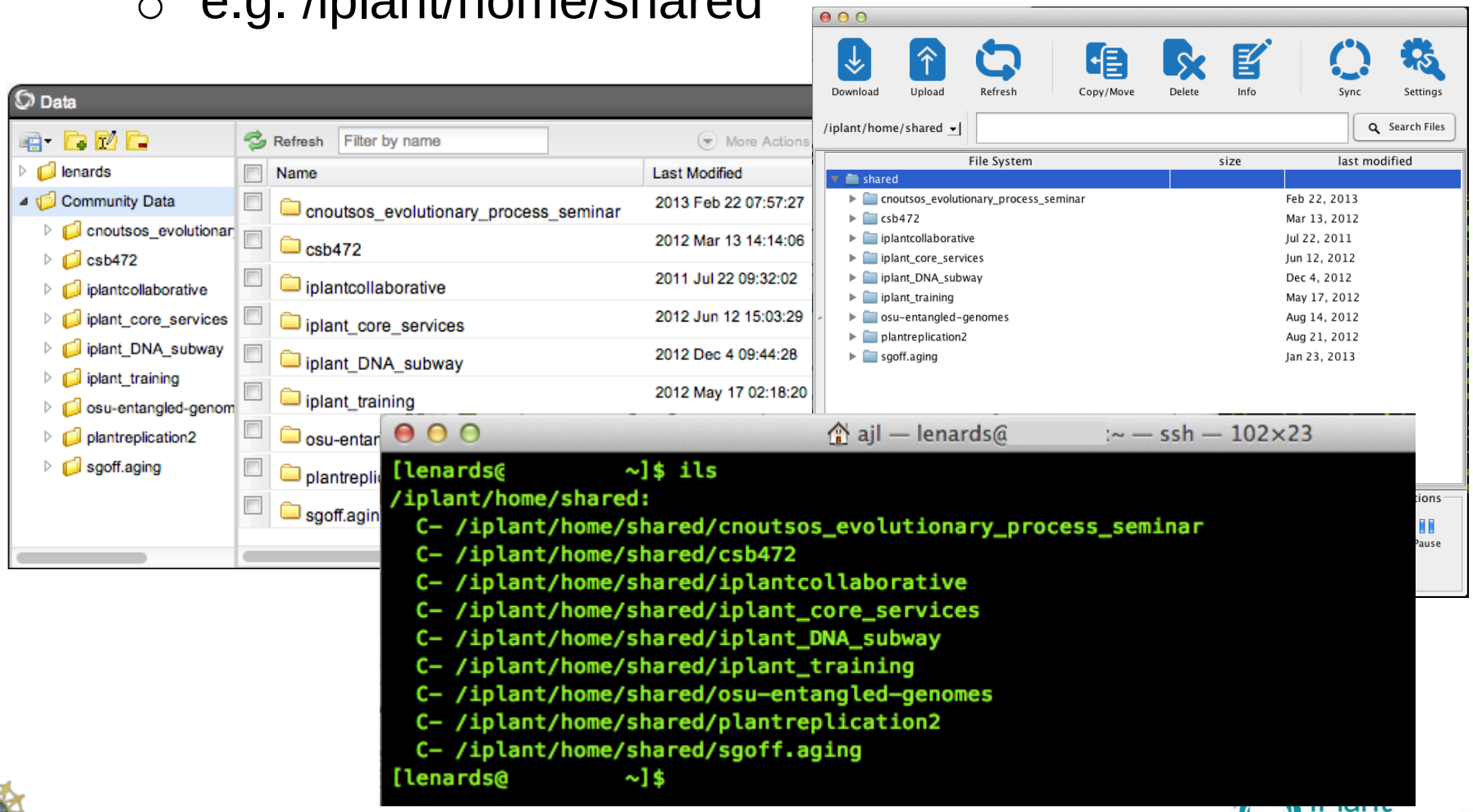
Features: Sharing



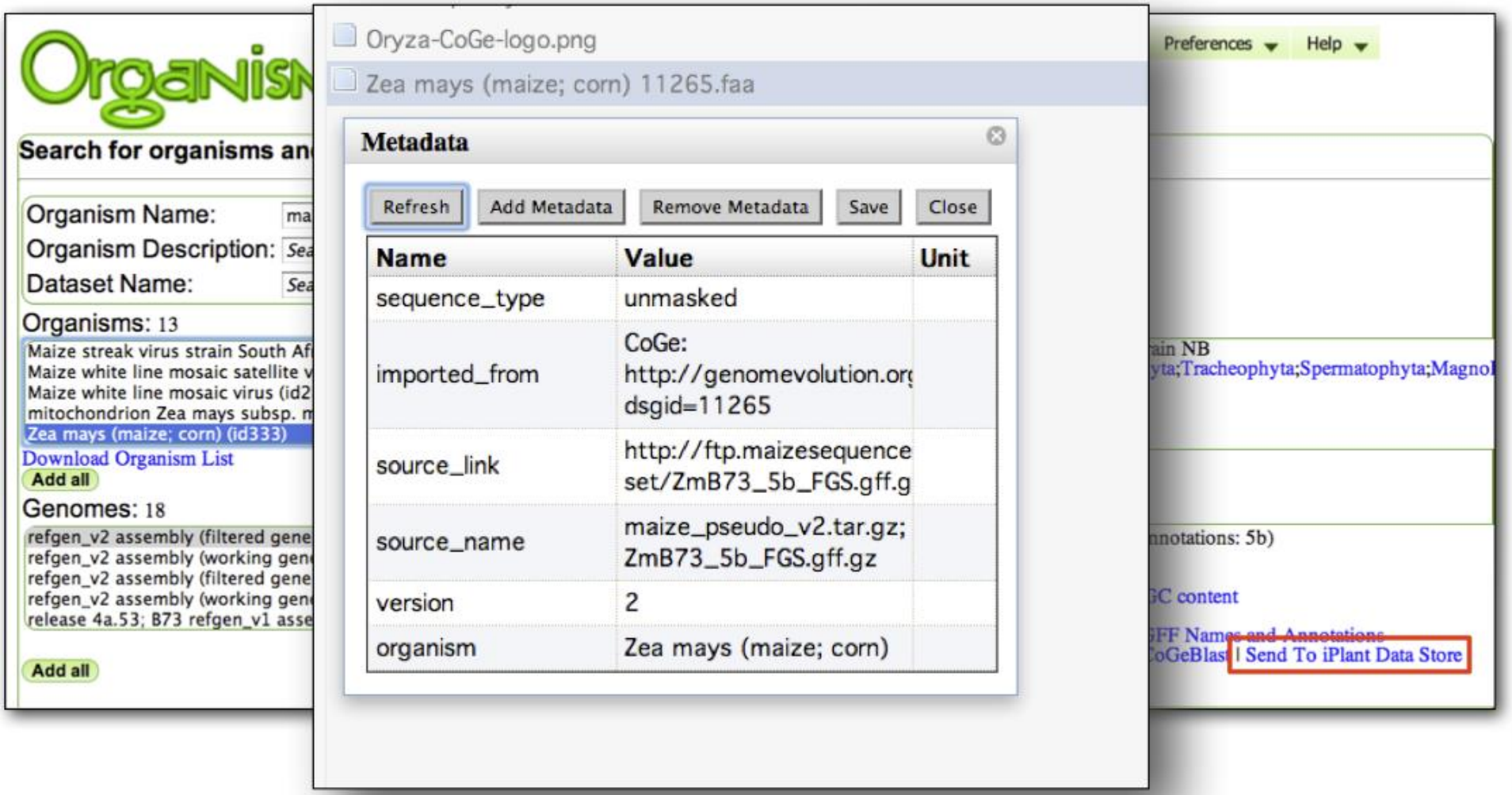
iPlant Collaborative™

Features: Community Data

- Using a convention for community shared folders
  - e.g. /iplant/home/shared



Embedded Metadata



iPlant Collaborative™

Challenges

- 9900+ users
  - diversity of use cases
  - eclectic technology background
- ACLs: 632M ACLs for an iCAT that has 68M objects
  - iRODS 3.2 improved ACLs performance
  - Allowed us to reduce the ACLS from ~280M

iPlant Collaborative™

Challenges

- General purpose file repository challenges
  - Dealing with firewall issues (e.g. idle connection timeouts!)
  - Network provider issues & mirroring large data sets
  - Java issues (both browser & desktop)
  - Searching and indexing
  - Quota management
    - current model doesn't map well to our use cases

iPlant Collaborative™

Hacks and Optimizations

- using AVUs for a "view" of shared data
- using GenQuery in Jargon for read-only operations
- SQL optimizations -> 600x faster (HT to Nirav)
  - partitioning
  - PostgreSQL ANY(Array) for some subselects

iPlant Collaborative™

Continued growth in user community...

iPlant Collaborative™

Animal Genomics

The problem

- Multiple emergent large-scale projects in agricultural animal genomics, genetics, and stress physiology using next-gen sequencing
- National Animal Genome Research Program (NRSP-8) : *Minimal* agency-level resources for scalable computing, storage, and collaboration
- Coordinator for NRSP-8 contacted iPlant based on word-of-mouth

Our approach

- Extend iPlant support and resources to this program
  - iPlant Data Store
  - Foundation API
  - iPlant Discovery Environment
- Educate community members to develop and implement scalable versions of their own pipelines and algorithms

Results

Rapid adoption of iPlant Data Store

- 1000 Bull Genomes
- Water Buffalo SNP-chip
- Swine and Chicken Heat Stress Genetics
- Bovine, Sheep, & Horse Genome Projects
- ~40 TB data (and growing)

Advanced training: Three-day onsite "Introduction to Developing for iPlant" at TACC in July 2012 for animal genomics bioinformaticians

Scalable science:

- 192 CPU BWA alignment pipeline
- 768 CPU GATK-based genotyping pipeline
- 32 CPU RNAseq mapping pipeline
- More on the way…

We have radically transformed the process of animal genomics for these communities…




"The ability to transport 2 TB of data overnight using the iRODS system was particularly helpful because previously, we had been mailing hard drives which is not an optimal solution to sharing big data"

"We've successfully used iPlant to map buffalo sequencing data from [multiple] breeds to the bovine genome and the [water] buffalo genome for SNP and INDEL detection. This took only a few days, where it would have taken more than a month previously. That allowed us to help the buffalo community quickly create a SNPchip on a short timeframe and allowed us to more quickly provide variants for use in defining genetic diversity in water buffalo."

"Among the most helpful aspects of using iPlant has been the ability to more efficiently conduct collaborative research... Our collaborators have been able to use tools at iPlant to conduct RNAseq analyses and variant calling. [iPlant resources] have helped individuals with very limited programming experience do bioinformatics quickly so that they can spend more time working on understanding the biology related to their areas of study."

iPlant Collaborative™

Future Work

- data-driven analyses
  - (rules engine driven, aka “smart data”)
- search and indexing
  - (making crawling a first class citizen)
- deeper integration with cloud
  - (integration with Atmosphere and OpenStack)
- more NetCDF, HDF5 use for Bio data
- SAM/BAM/VCF other popular indexed NGS file native support

iPlant Collaborative™

Future Work (continued)

- Metadata & ontology guidelines for our community
  - very early draft being developed by staff & community
- Usability Efforts
- Software Defined Networks
- Sensor data streaming from Data Turbine for project SEGA
  - http://www.dataturbine.org/
  - http://nsf.gov/awardsearch/showAward?AWD_ID=1126840

iPlant Collaborative™

Questions

iPlant Collaborative™

Thank you

Mike Conway, Antoine de Torcy, Wayne Schroeder, Sheau-yen Chen, Arcot Rajasekar, Reagan Moore

The iPlant Collaborative is funded by a grant from the United States National Science Foundation (#DBI-0735191).

URL: www.iplantcollaborative.org

iPlant Collaborative™

The iPlant Collaborative

Executive Team:

Steve Goff

Dan Stanzione

Postdocs:

Barbara Banbury
Christos Noutsos
Solon Pissis
Brad Ruhfel

Students:

Peter Bailey
Jeremy Beaulieu
Devi Bhattacharya
Storme Briscoe
Ya-Di Chen
David Choi
Barbara Dobrin
John Donoghue
Yekatarina Khartianova
Chris La Rose
Amgad Madkour
Aniruddha Marathe
Andre Mercer
Kurt Michaels
Zack Pierce
Andrew Predoehl
Sathee Ravindranath
Kyle Simek
Gregory Striemer
Jason Vandeventer
Nicholas Woodward
Kuan Yang

Metadata → Data → Tools → Workflows → Viz

Faculty Advisors & Collaborators:

Ali Akoglu
Kobus Barnard
Timothy Clausner
Brian Enquist
Damian Gessler
Ruth Grene
John Hartman
Matthew Hudson
David Lowenthal
Eric Lyons
B.S. Manjunath
Nirav Merchant
David Neale
Brian O'Meara
Sudha Ram
David Salt
Mark Schildhauer
Doug Soltis
Pam Soltis
Edgar Spalding
Alexis Stamatakis
Doreen Ware
Steve Welch

Staff:

Greg Abram
Sonali Aditya
Ritu Arora
Roger Barthelson
Rob Bovill
Brad Boyle
Gordon Burleigh
John Cazes
Mike Conway
Victor Cordero
Rion Dooley
Aaron Dubrow
Andy Edmonds
Tony Edgin
Dmitry Fedorov
Melyssa Fratkin
Michael Gatto
Utkarsh Gaur
Cornel Ghiban
Steve Gregory
Matthew Hanlon
Natalie Henriques
Uwe Hilgert
Nicole Hopkins
Eun-Sook Jeong
Logan Johnson
Chris Jordan
Kathleen Kennedy
Mohammed Khalfan
David Knapp
Lars Koersterk
Sangeeta Kuchimanchi
Kristian Kvilekval
Sue Lauter
Tina Lee
Andrew Lenards
Monica Lent
Zhenyuan Lu
Aaron Marcuse-Kubitz
Naim Matasci
Sheldon McKay
Robert McLay
Dave Micklos
Nathan Miller
Steve Mock
Martha Narro
Shannon Oliver
Benoit Parmentier
JMatt Peterson
Dennis Roberts
Paul Sarando
Jerry Schneider
Bruce Schumaker
Edwin Skidmore
Brandon Smith
Mary Margaret Sprinkle
Sriram Srinivasan
Josh Stein
Lisa Stillwell
Jonathan Strootman
Peter Van Buren
Hans Vasquez-Gross
Matthew Vaughn
Rebeka Villarreal
Ramona Wallls
Liya Wang
Anton Westveld
Jason Williams
John Wregglesworth
Weijia Xu

References / Resources

- Data Sharing & Management Snafu:
  - http://bit.ly/YBPhr3

iPlant Collaborative™