

Secure Data Management in BIRN



Karl Helmer

Massachusetts General Hospital

September 6, 2011

For the BIRN Consortium

Talk Outline:

- BIRN Overview
- Data Services Overview
- Example Use Cases:
 - Function BIRN (FBIRN) Data Federation
 - Pathology Viewer Project
- Working with BIRN

The Purpose of BIRN

“to advance biomedical research through data sharing and online collaboration”



Challenges ... Opportunities

What is the main challenges of collaborative science?

- Not just size of datasets ...
- Not just communication bandwidth ...

The challenge is heterogeneity in both:

- User technological needs
- Sociology of communities

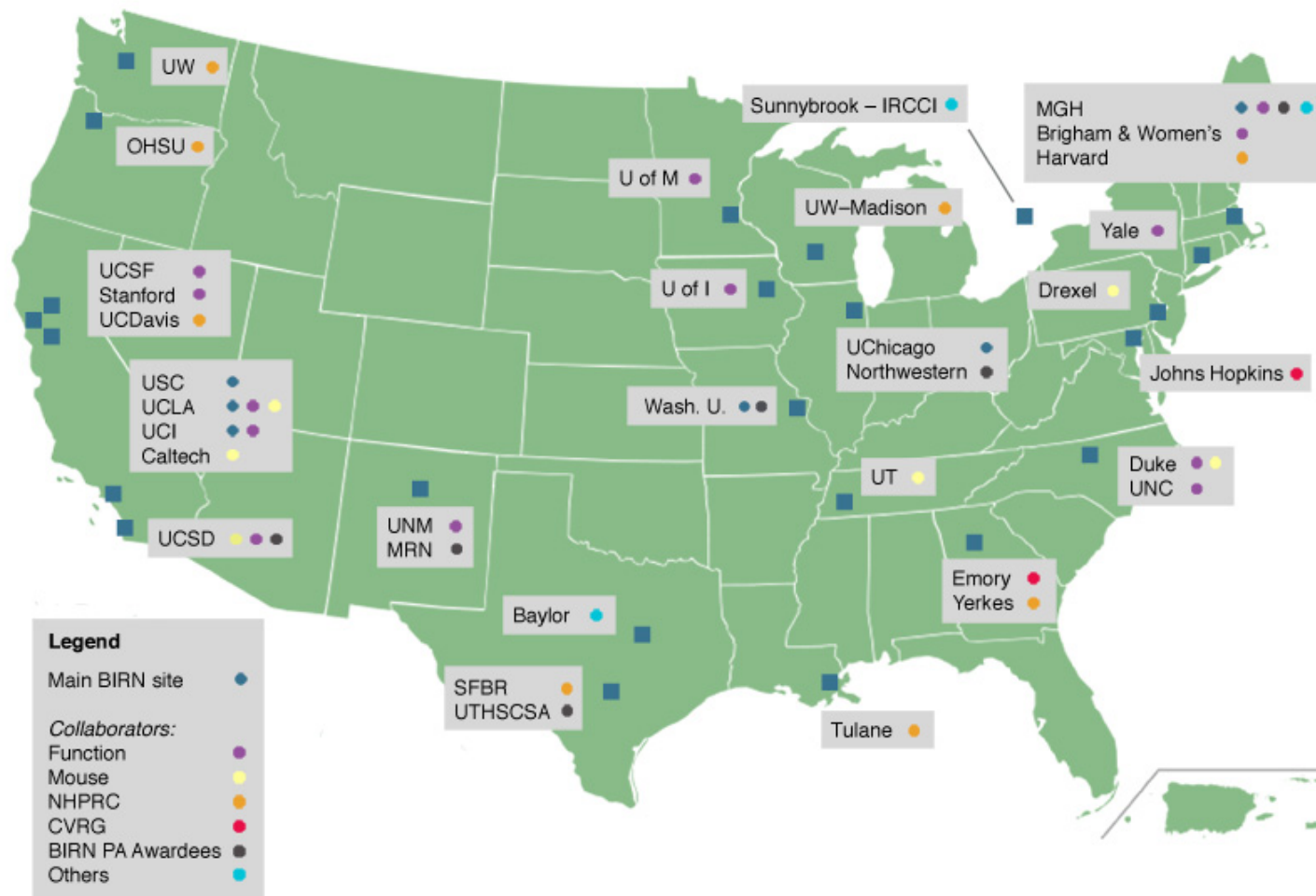
BIRN Capabilities

BIRN is driven by the things our users want

Develop *Capabilities* - *i.e.*, our ability to:

- Move or Publish data / Deploy tools
- Reconcile data across disparate sources
- Provide security (credentials, encryption)
- Develop and enable scientific reasoning systems
- Create a collaboration workspace for a new community

Sites and Collaborators



Technical Approach

- Bottom up, not top down
- Focus on user requirements - ‘users’ become ‘collaborators’
- Create solutions - factor out common requirements for reuse/modularity
 - Capability model includes both software and process
 - Avoid ‘Big Design up Front’

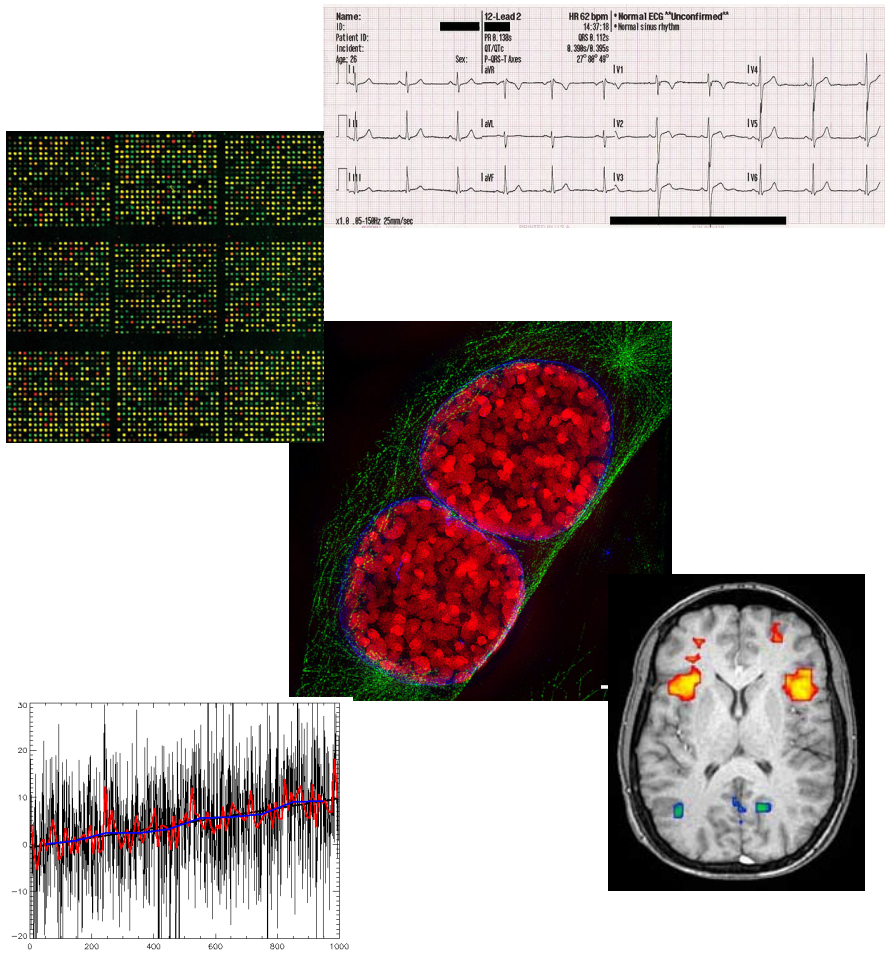
Sociological Approach

- Assist in building a community within a research domain
- Focus on “small step” initial project to demonstrate value
- Involve users in the design and implementation phases
- Acknowledgement that users may have system/expertise in place
- Seek sustainability of project

Dissemination Models

- Different problems require different kinds of solutions.
 - BIRN *operates services* for all users; e.g., user registration service
 - BIRN *provides kits* for project teams to deploy services for their members; e.g., data sharing
 - BIRN *provides downloadable tools* for individuals to use on their own; e.g., pathology databases, integration systems, reasoning systems, workflows.
- Understanding deployment needs is part of defining the problem to be solved.

Collaboration = Sharing Data



- BIRN DM seeks to support data-intensive activities including:
 - Imaging, Microscopy, Genomics, Time Series, Analytics and more...
- BIRN utilities scale:
 - Terabyte data sets
 - 100 MB - 2 GB Files
 - Millions of Files

DM Use Cases

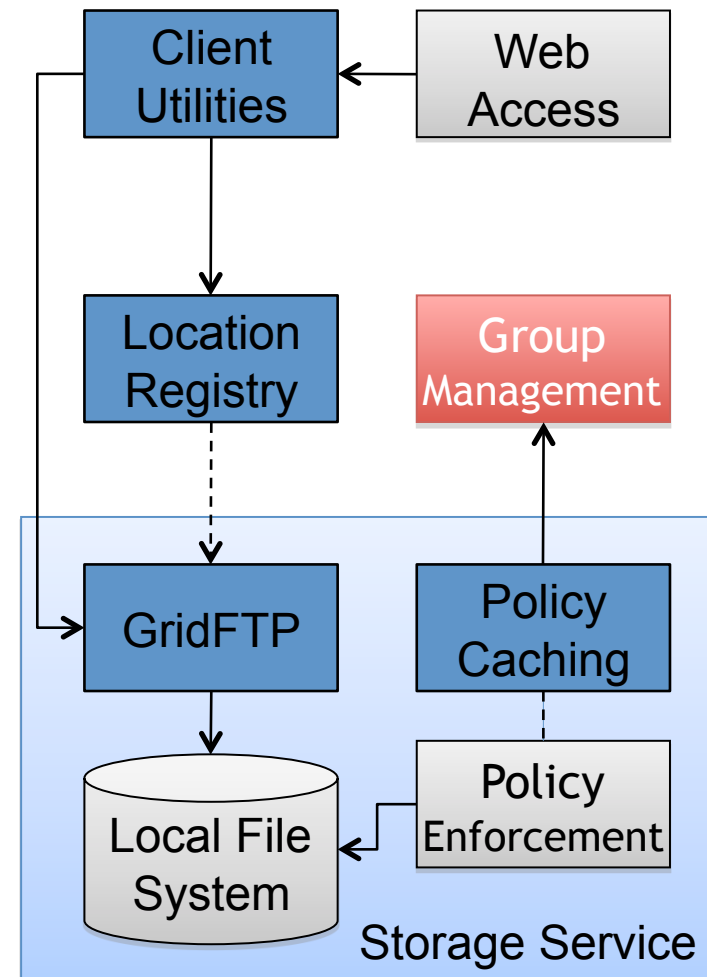
- *Data Capture: Instrument to Storage*
 - 1 to 1 transfer; typically over LAN
- *Data Transfer: Storage to Storage*
 - 1 to 1 transfer; typically over WAN
- *Data Publication and Federation*
 - Many to many transfer; WAN
- *Data Consumption: Client Access*
 - Many to/from 1 transfer; WAN and LAN

DM Key Requirements

- *Security*
 - User and Group management; protection of sensitive data, federated access to data
- *Scalability*
 - Support growth in projected data usage: data volumes, file sizes, users, sites, ops/sec
- *Infrastructure*
 - Work with commodity networking, storage, and the presence of strict firewalls.

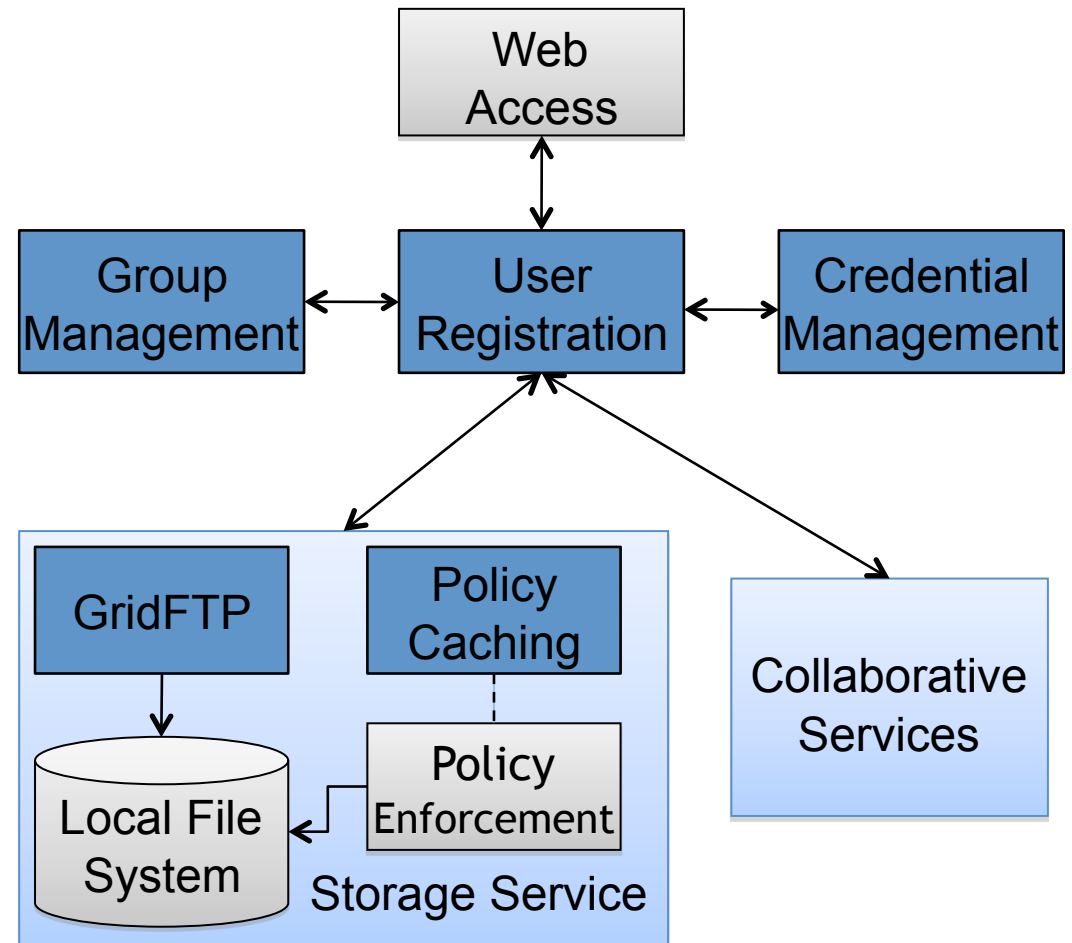
Data Service Architecture

- Client Utilities
 - UNIX utilities
 - Java & C APIs
- Shared Services
 - Manage File Locations
 - Manage Users
 - Manage User Groups
- Local Services
 - Integrate with conventional file systems

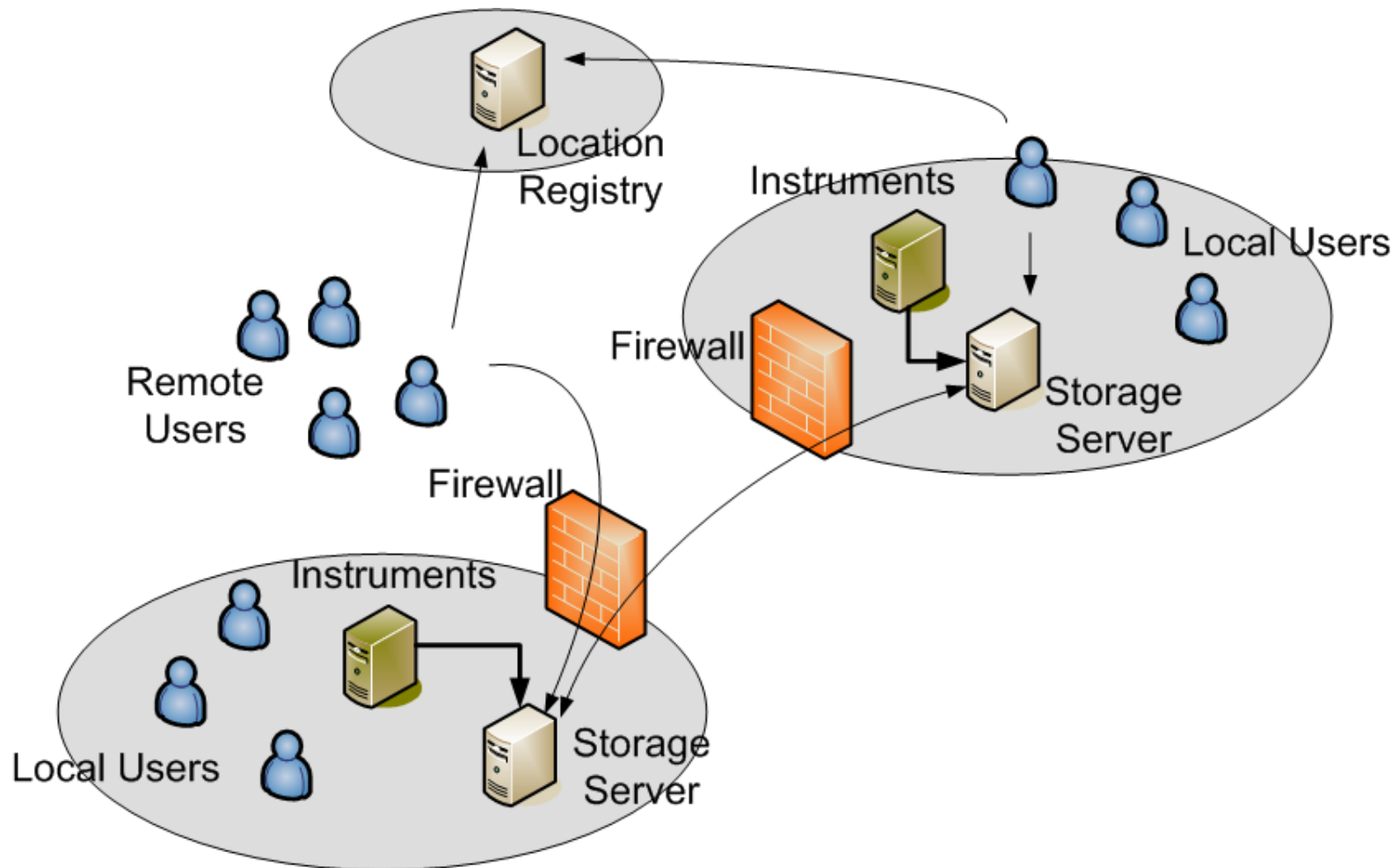


Security Service Architecture

- Client Utilities
 - UNIX utilities
 - Java & C APIs
- Shared Services
 - Group management service
 - Registration service
 - Credential service
- Local Services
 - Integration with local applications



Typical Data Grid Deployment



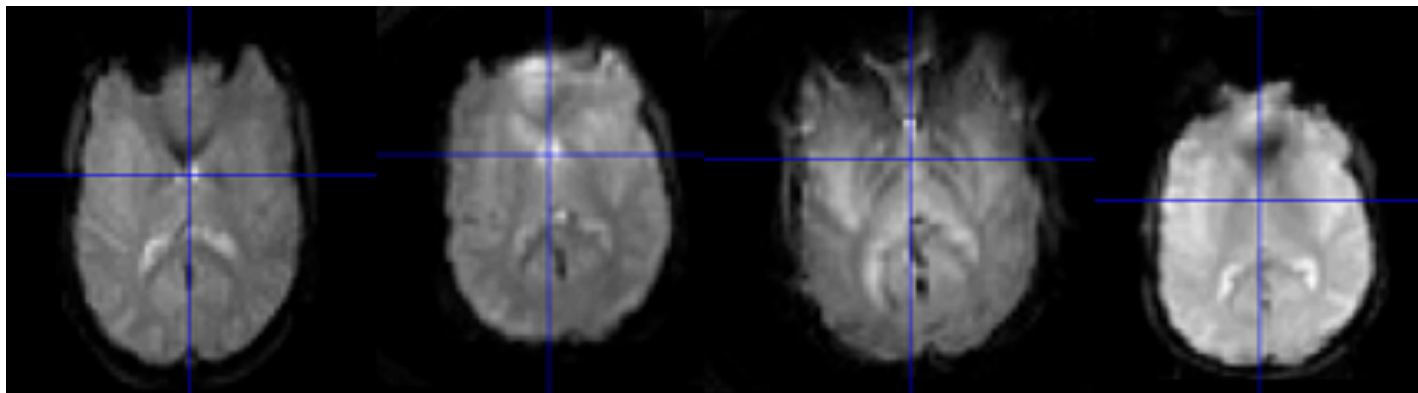
Use Case 1: Function BIRN (FBIRN)



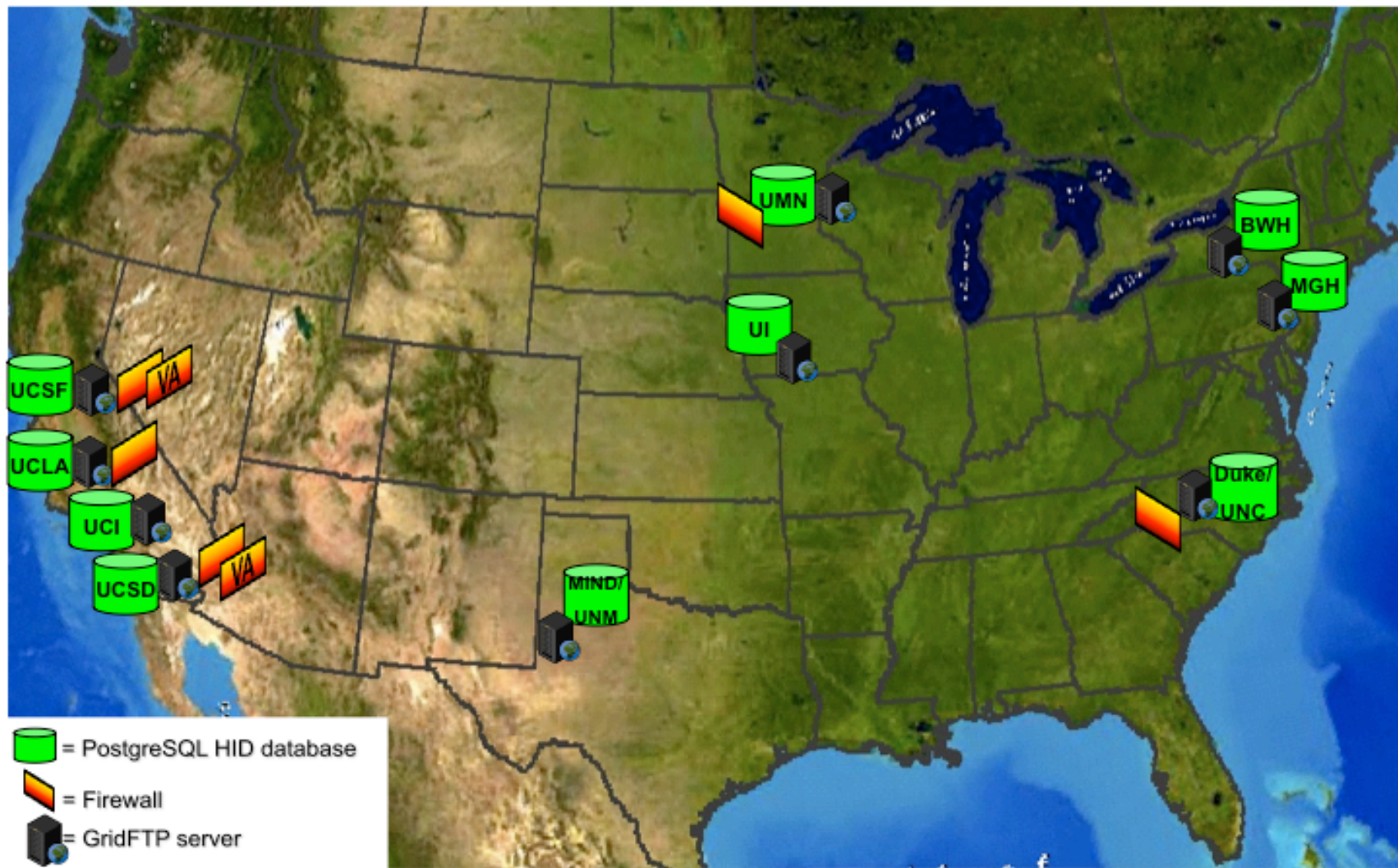
Multisite Study of Schizophrenia

Q: Multisite Data Collection?

- FBIRN is testing improved methods for collecting, sharing, and analyzing multisite fMRI data in clinical populations.
- Why? - Same person, scanned at MRI centers using different machines and “best” protocols, does not produce the same picture.



FBIRN



FBIRN DM Requirements

- High transfer success with verification for consistency
- Failures at other sites must not prevent users from accessing their own local data
- Experiment will produce over a million files with size of 100s of bytes to several MBs
- System must store files in a manner that is interoperable with conventional systems
- Users part of one or more groups, system must support group-based access control

FBIRN Data Lifecycle

- Project lifecycle:
 - data capture phase
 - small number of large bulk writes to local storage
 - multiple concurrent writes to dataset not common
 - data quality assurance phase
 - data access phase
 - Local and remote bulk data reads
 - multiple concurrent reads are common

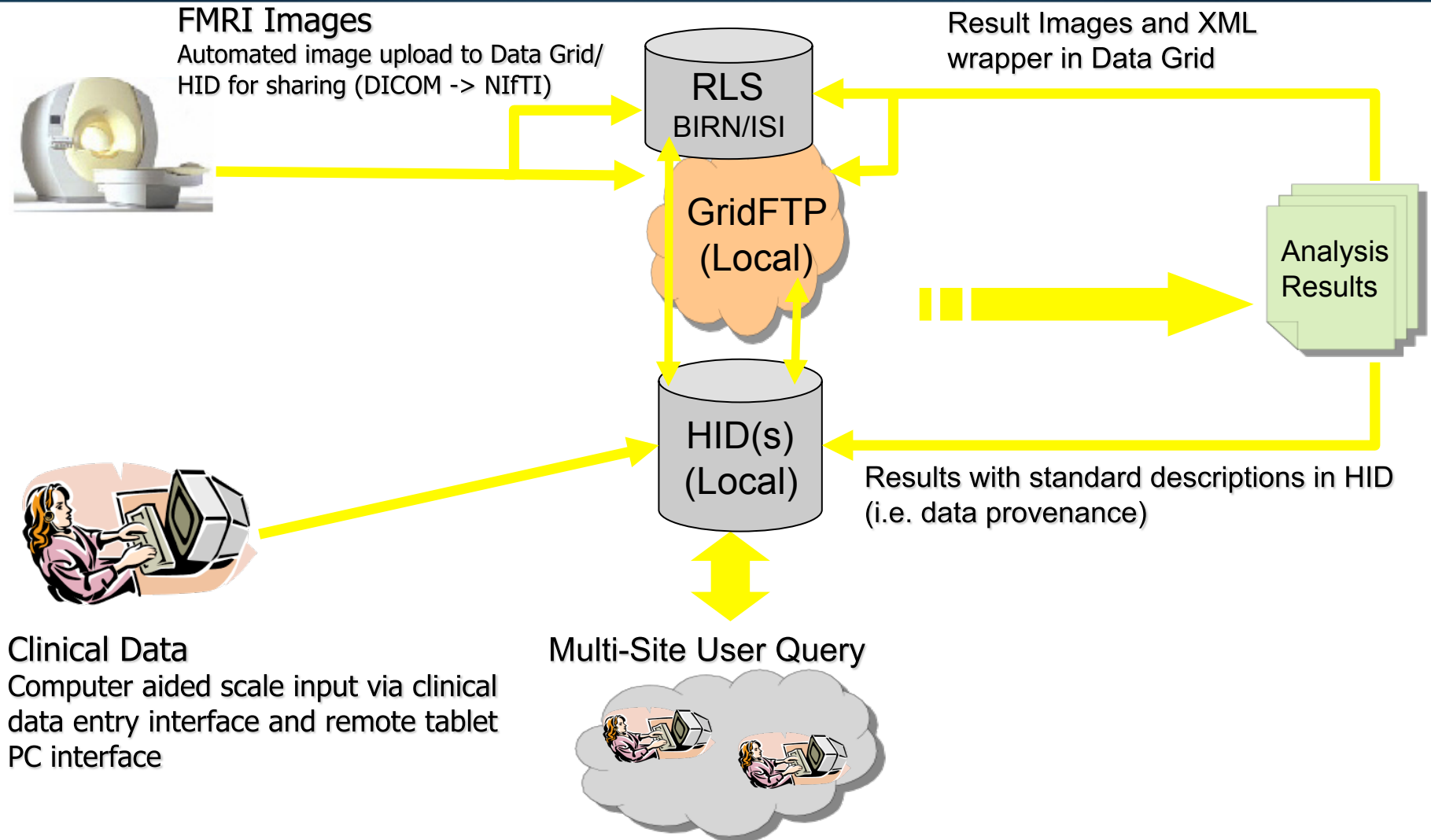
Human Imaging Database

- Open-source, extensible neuroimaging and clinical data management system.
- Native support for federated HIDs (multi-site query/retrieve)
- Graphical form designer for rapid clinical assessment deployment
- Support for GridFTP and SRB distributed file systems and NFS mounted filesystems for data linkage
- Support for Globus grid security management
- Support for derived data storage/query/retrieval (beta)
- Support for genetic data storage/query/retrieval (alpha)
- Software: www.nitrc.org/projects/hid

FBIRN: RLS and GridFTP features at work

- Replica Location Service:
 - Enables *location independent* data placement and retrieval
 - Supports *fast lookup* across distributed sites
- GridFTP: A secure FTP service
 - Strong cryptographic authentication (X509)
 - Strong cryptographic data privacy (TLS)
 - Recursive directory copy between systems
 - “Tar streaming” of small files to improve performance and reliability

FBIRN: HID integration with RLS and GridFTP



(F)BIRN Technology Summary

- High Performance Transfer Protocol
 - Strong security
 - High performance, proven technology
- Location Registry
 - Distributed registry to track file locations
 - Low overhead, supports 100's ops/s
- User Management
 - Registration and credentials
- Group Management
 - Centralized group management
 - Vetting of users for group admission workflows
 - Caching at local servers
- Policy Enforcement
 - Flexible access control policies
 - User and group ownership rights

Features & Benefits

Features

- Support for multiple authentication protocols and security levels
- Users/Groups own data, not the system
- Flexible access control permissions
- Supports small file streaming
- Parallel streams and transfer restarts

Benefits

- Adjust to user and project preferences, relax for > performance
- Fewer vulnerabilities for data security
- Collaborative project data sharing
- Improve small file transfer performance
- Improve large file transfer performance

FBIRN Results and Next Steps

- Results
 - Deployed across 10+ Function BIRN sites
 - Reliably moving 40+ GBs in production use
 - Performance gains of ~3x SSH Copy
- Next Steps
 - Simplify user/group administration (In progress)
 - New utility for data synchronization
 - Support for additional transfer protocols

Specific BIRN Infrastructure Improvements for FBIRN

- On-the-fly tarring (grouping) of small files to improve transfer performance
- Avoiding control channel disconnects using the TCP keep-alive mechanism
- Continuous monitoring of FBIRN infrastructure

Transfer characteristics of tar-enabled GridFTP

Concurrency	Size (GB)	Speed (MB/s)	Success Count	Failure Count	Success Rate(%)
1	2	15.44	88	0	100
1	21	10.41	22	0	100
3	2	9.12	174	0	100
3	21	6.77	27	0	100
5	2	6.43	85	0	100
5	21	5.41	45	0	100
7	2	5.48	84	0	100
7	21	4.59	49	0	100
9	2	3.62	90	0	100
9	21	4.40	77	4	95
11	2	2.96	100	10	91
11	21	2.78	60	17	78

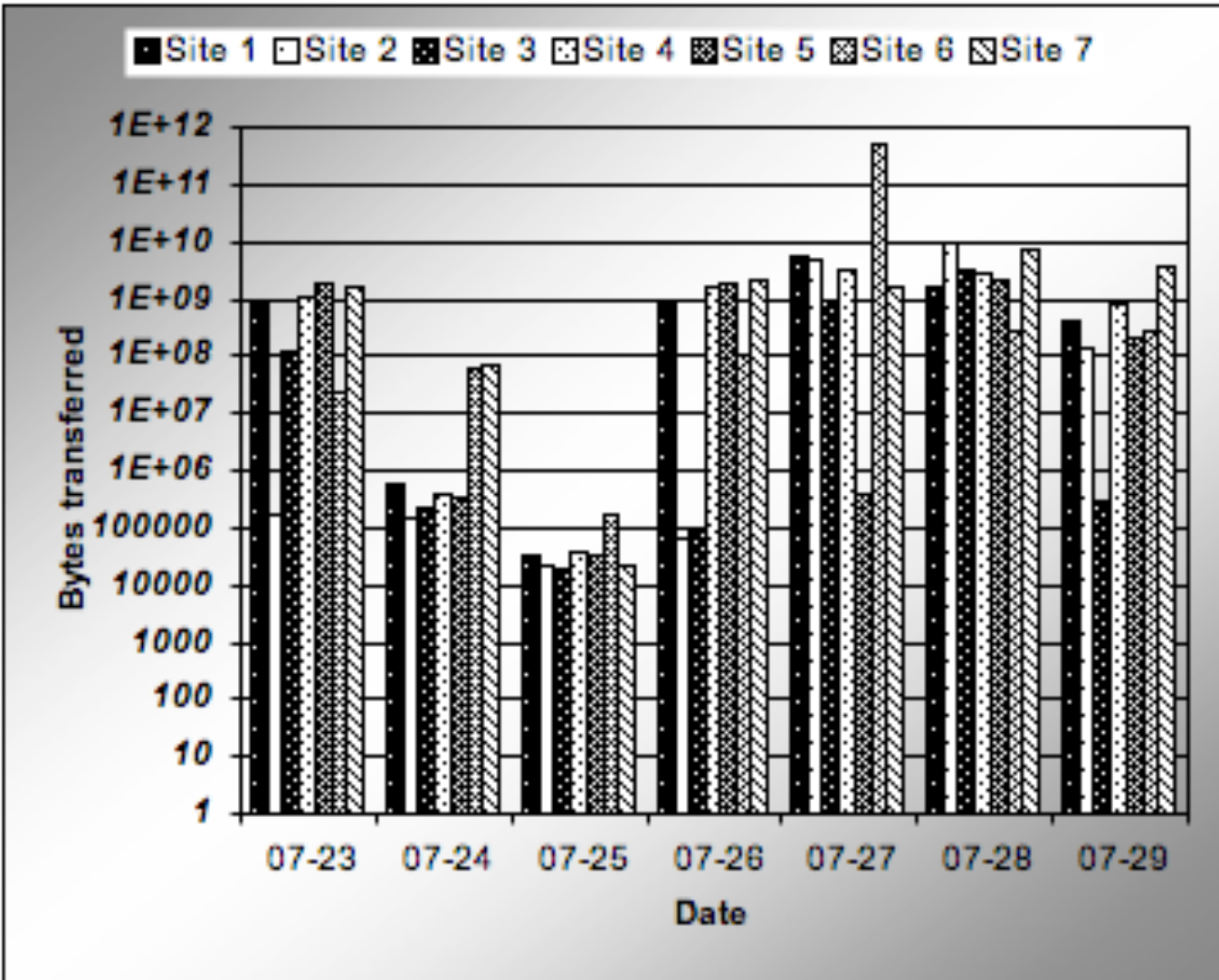
Control Channel Disconnect

- No data over the control channel after the connection has been established and while data are being sent
 - Only after the data is transferred - control channel used to complete the transfer
- Control channel could be idle for long time
 - When clients or servers behind a NAT proxy or a firewall, connections get disconnected
 - Caused problems with long-running transfers, control channel connection was dropped silently - transfers hung and failed.

Avoid Connection Dropping with TCP Keepalive Mechanism

- Solution - we leverage the TCP keepalive mechanism to avoid connection dropping
 - “probes” are transmitted to network peers at a configurable time interval
- Linux has built-in support for TCP keepalive
 - Default settings exceed the time limit used by proxies and firewalls
- With TCP keepalive configured at frequent interval, connection dropping was avoided.
 - All tests listed in Table 1 succeeded

FBIRN Data Movement Volume



Continuous Monitoring

- FBIRN scientists need the data transfer services to behave predictably:
 - Performance should meet expectations based on past behavior.
- This requirement led us to two goals:
 - Stabilize the system's behavior
 - Measure the system's behavior and publish results to set reasonable expectations among the users.

Automated Test Mechanism

- Set of tests across all of the FBIRN GridFTP servers on a continuous basis
- Two types of tests are performed
 - A relatively short data transfer between each pair of data centers is performed frequently (multiple times per day)
 - A relatively long data transfer is performed infrequently (once per day or every other day) to measure the performance of the system.

Continuous Monitoring

		Throughput (MB/s)								
From	To	5/21	5/22	5/23	5/24	5/25	5/26	5/27	5/28	5/29
Site A	Site C	*	X	X	X	X	X	X	*	*
	Site D	**	**	X	X	X	X	X	**	*
	Site E	O	*	O	*	*	*	*	*	O
	Site F	**	**	**	X	**	**	**	**	**
	Site G	**	**	**	X	X	**	**	**	**
	Site B	*	X	X	X	X	X	X	*	*
	Site H	**	**	**	**	**	**	**	**	*
	Site I	**	X	X	X	X	X	X	**	**
	Site J	**	**	**	**	**	**	**	**	**
Site B	Site A	*	*	*	*	*	*	*	*	*
	Site C	*	*	*	*	*	*	*	*	*
	Site D	***	***	***	***	***	***	***	***	***
	Site E	*	*	*	*	*	*	*	*	*
	Site F	**	**	**	**	**	**	**	**	*
	Site G	**	**	**	**	*	**	**	**	**
	Site H	***	***	***	***	***	***	***	***	**
	Site I	***	***	***	***	***	***	**	**	***
	Site J	***	***	***	***	***	***	***	***	***

Legend:

* Complete <25 MB/s

** Complete 25-50 MB/s

*** Complete >50 MB/s

X Failed

O Expired

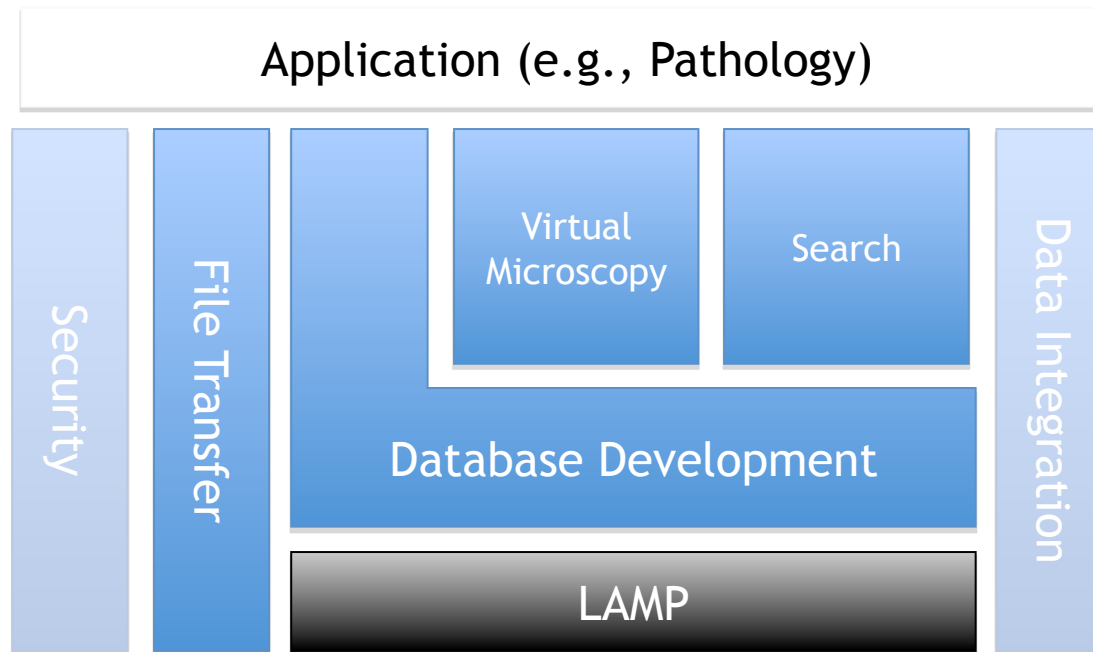
BIRN PATHOLOGY



Project History

- The BIRN Pathology project began in early 2010
- Objective was to create a new shared resource for the research community: an online collection of pathology data, including very high resolution microscopy images
- Initial user group consists of pathologists
- ISI developing the infrastructure
- Early phases: rapid prototyping and iteration with users to build out the core data model, user workflows and user experience, and investigation of microscopy tools
- January 2011: packaged (RPM), documented (wiki), released (YUM repo) the version 1.0 of the system
- Ongoing enhancements to data model, microscopy image support, refactoring of system into reusable components

BIRN Pathology Technology ‘Stack’



key

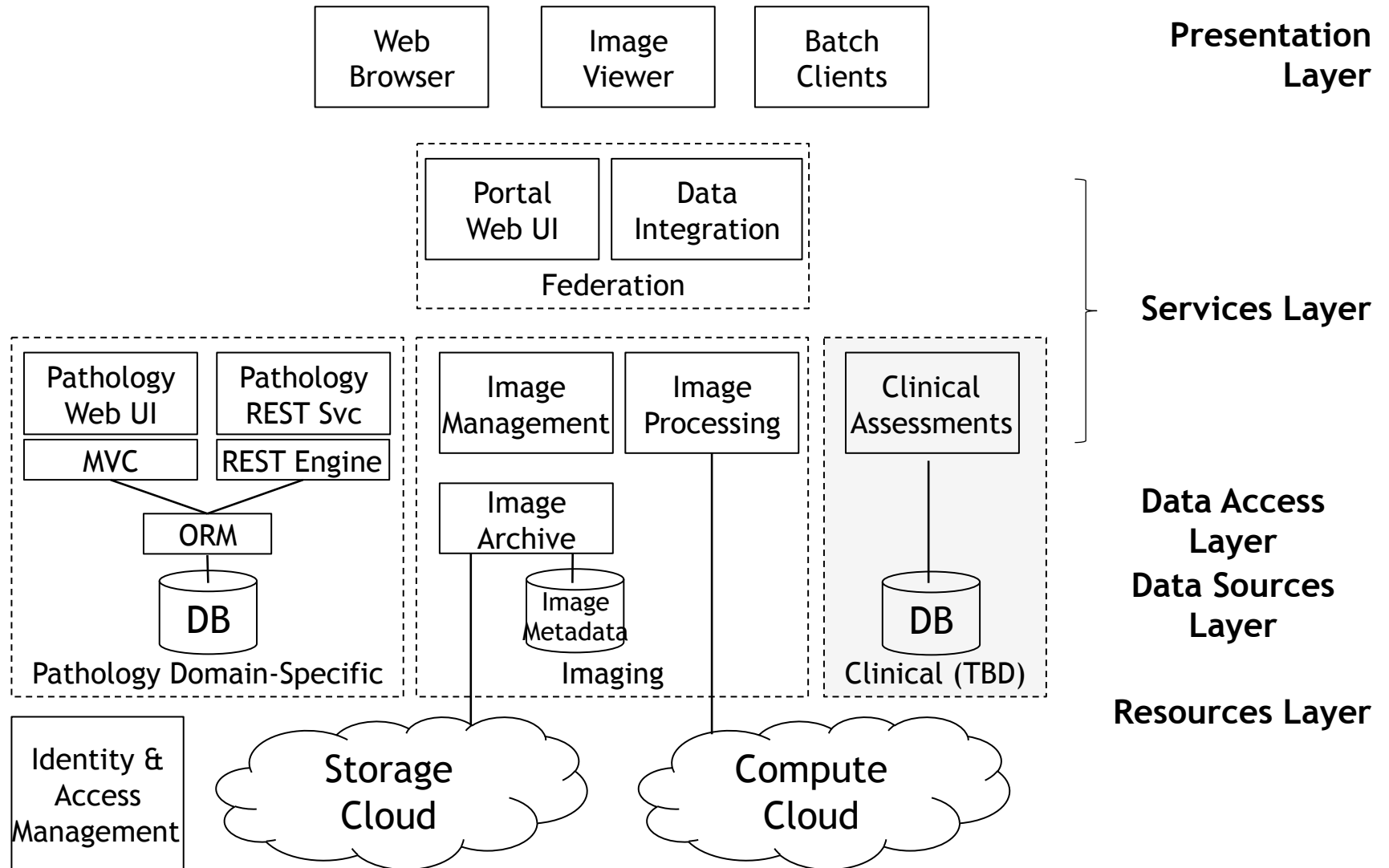
Data Management
"Capabilities"

Other BIRN
"Capabilities"

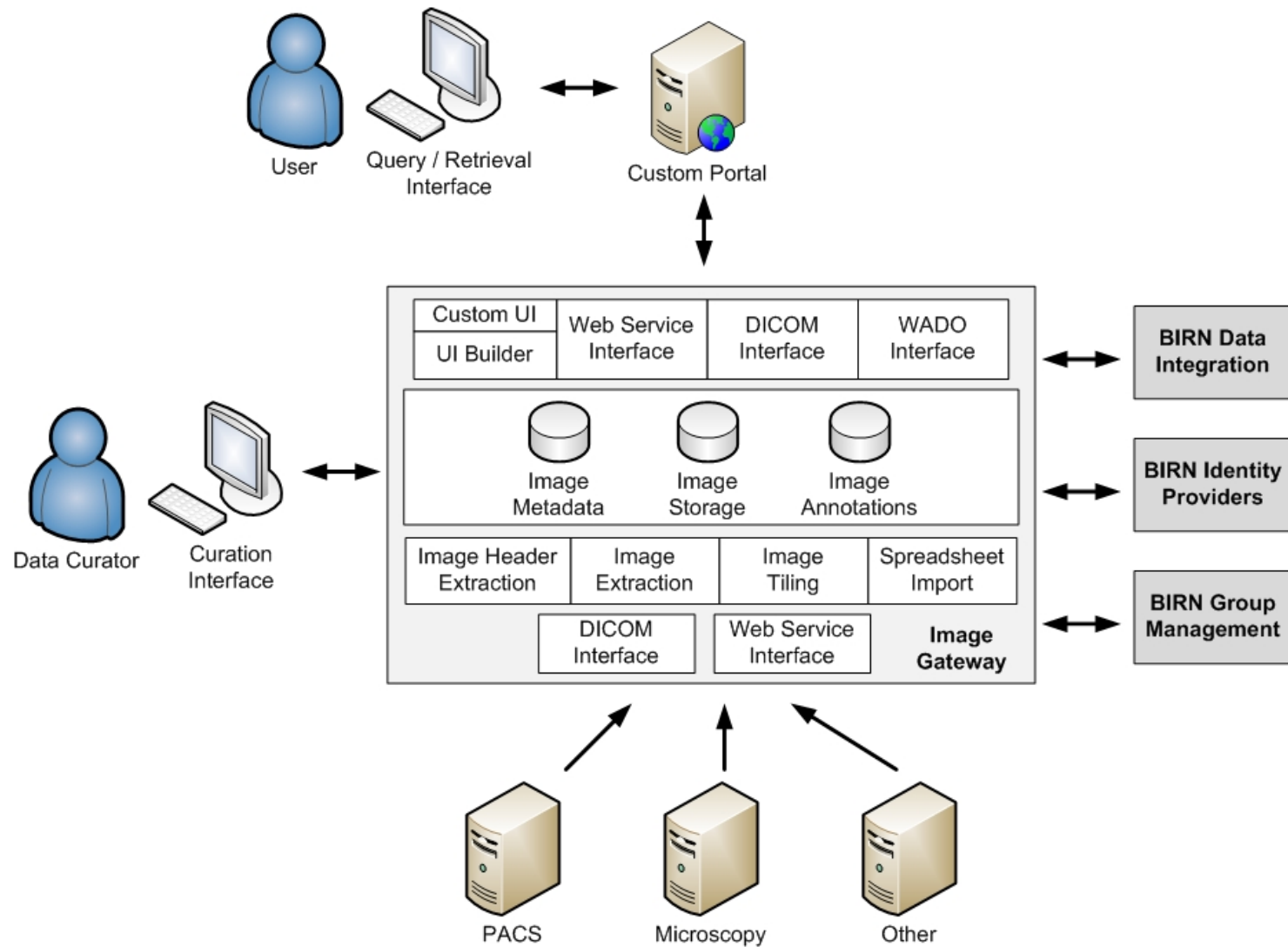
3rd Party Tools

Application Specific

Pathology Systems Architecture



Imaging Gateway



Database Development

Use Cases

Scientists need to create custom web-accessible databases to share data within and between research groups.

Scientists store data on spreadsheets, need to manage this data (DBMS), and expose it programmatically (WS REST)

- In a nutshell:
 - **Django + Extensions**
- Features
 - Django Web Framework
 - MVC, ORM, HTML Templates
 - Auto-generated data forms!
 - RESTful interface routing
 - Spreadsheet batch importer
 - Crowd integration (users/groups/SSO)
 - Data object access control (ACLs)
 - Javascript UI widgets
 - Scheduled database backup
- System Requirements
 - Linux, Apache, MySQL or Postgres, Python, Django

Virtual Microscopy

Use Cases

Scientists need to share high resolution (100X) images over relatively low bandwidth networks.

Scientists need to annotate and share annotations on images.

- **Image Processing Tools**
 - Accepts images in numerous proprietary microscopy formats (Bio-Formats)
 - Creates pyramidal, tiled image sets in open format (JPEG) used by Zoomify image viewer
 - Extracts proprietary graphical annotation formats and converts them to Zoomify annotation format
- **Image Server + Database**
 - (Built on our Database Development tools)
 - Monitors an image upload directory, automatically processes images, and updates the image database
 - Stores and serves annotations according to associated ACLs

Search

Use Case

Scientists need to search their databases using full-text queries and field-based “advanced” search queries.

- ### Features

- Single-site, full-text indexing of databases and text files.
- Search results returned as Python objects based on Django ORM mappings.
- ‘Advanced search’; i.e., multiple fields, and logical expressions.
- Easy to synchronize index with changes to Django-based object model.

- ### Technology

- Based on Xapian and Djapian

Pathology Application

Use Cases

Pathologists create images in proprietary microscopy formats, annotate the images, and need to upload them to a web-accessible database for sharing among collaborating sites.

Pathologists associate metadata and supplementary files with their microscopy images.

• Features

- Pathology Data Model
 - Subject, Case, Specimen, Tissue, Image entities
 - Species, Disease, Etiology domain tables
- Pathology data entry workflows
 - UI forms tailored to the data entry workflow of pathologists.
- Microscopy support for proprietary formats:
 - Aperio SVS (*supported*)
 - Hamamatsu NDPI (*partially supported*)
 - Olympus VSI (*under development*)
- And, of course, all of the features from the underlying software and services
 - Database Development + Virtual Microscopy + Search
 - Data Integration using BIRN Mediator
 - Capability to run federated queries across sites

Screenshots

BIRN**BIRN Pathology Image Database**Welcome, David [Change Password](#) / [Log Out](#)

Home > Edit > Animal A-111-22-33334 > Case A2145-B

Change Case and Clinical Diagnosis

Case ID	A2145-B
Subject	A-111-22-33334
Case Date	2008-06-04
Age (in days)	
Age (in years)	26.0
Age Category	Adult ▼
Body Weight (kg)	72.72
Comments	Sample case for demo

Procedure

Available Procedures

CSF COLLECTION (P-1C118)
ABORTION, INDUCED, NOS (P-17550)
ACID FAST (P-YY131)
ACID FAST BACILLI (P-YY128)
ACID-ETHER CONCENTRATION (P-34411)
ACUPUNCTURE (P-1X900)
ADMINISTRATION OF DRUG OR SUBSTANCE, VIA ENDOTRACHEAL (P-N2132)
Administer alcohol by IV (P-YY900)

Chosen Procedures

Select your choice(s) and click

LOCAL ANESTHESIA, NOS (P-1X400)

Case History
Comments

Diseases

[CARDIOMYOPATHY, NOS \(D-71000\)](#) ☐ [Delete](#)

[Add Disease](#)

Attachments

[Add New Attachment](#)

[Delete](#)

[Save](#) [Save and assign images](#)

Screenshots

Case Details

Case ID	A2145-B
Subject	A-111-22-33334
Case Date	June 4, 2008
Age (in days)	None
Age (in years)	26.0
Age Category	Adult

Clinical Diagnosis

Case History Comments		
Diseases	Disease	CARDIOMYOPATHY, NOS (D-71000)
	Etiology	ANTIBIOTIC, NOS (E-72000)
	Etiology Description	
	Disease Process	Other
	Disease Process Description	
	Morphology Code	CARDIOVASCULAR MALFORMATION, CONGENITAL, NOS (M-24900)
	Morphology Code Description	
	Histo Diagnosis Code	None
	Histo Diagnosis Code Description	
	Topographical Location	HEART, NOS (T-32000)
	Transmission	

Images

No images are available.

Attachments

No attachments are available.

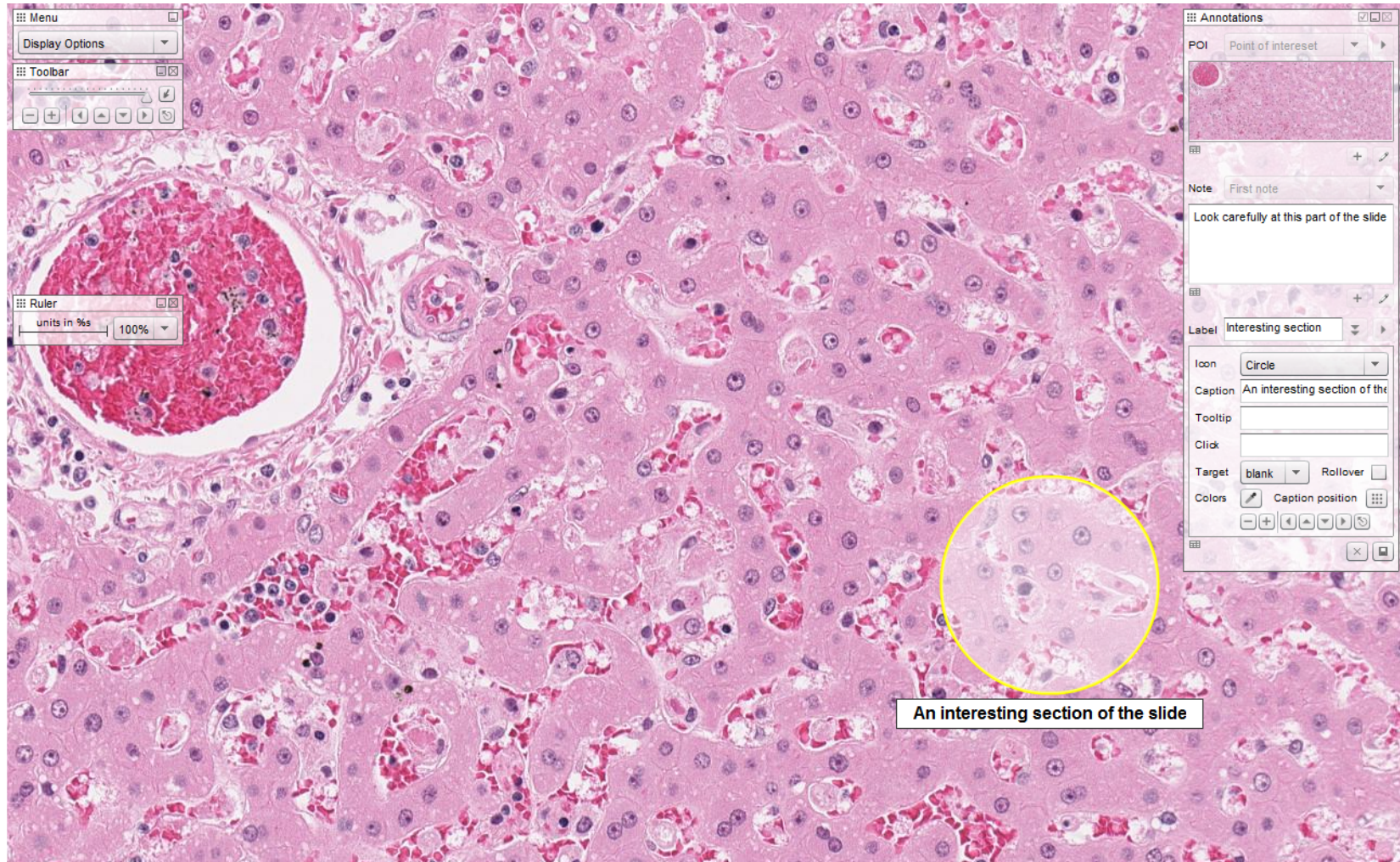
Specimens

Specimen ID	Organs	Tissue
555-1212	heart	vascular
555-1213	heart, aorta	vascular

Procedures

Procedure Name
LOCAL ANESTHESIA, NOS (P-1X400)

Screenshots



BIRN Pipelines for Visual Informatics and Computational Genomics



Ivo Dinov, Fabio Macciardi, Federica Torri,
Jim Knowles, Andy Clark, Joe Ames, Carl
Kesselman, and Arthur Toga

<http://pipeline.loni.ucla.edu>
[https://wiki.birncommunity.org/x/
wIBWAQ](https://wiki.birncommunity.org/x/wIBWAQ)

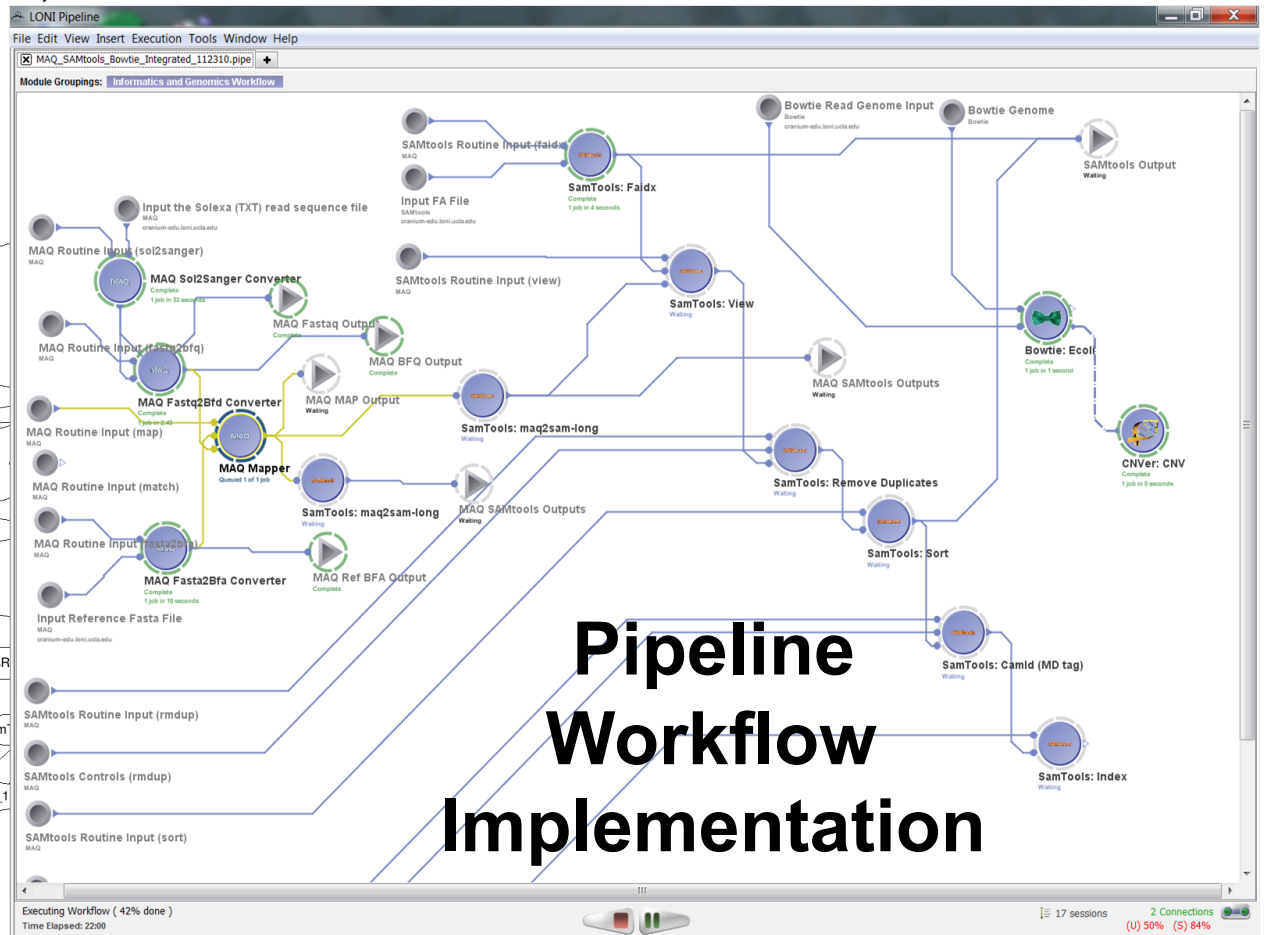
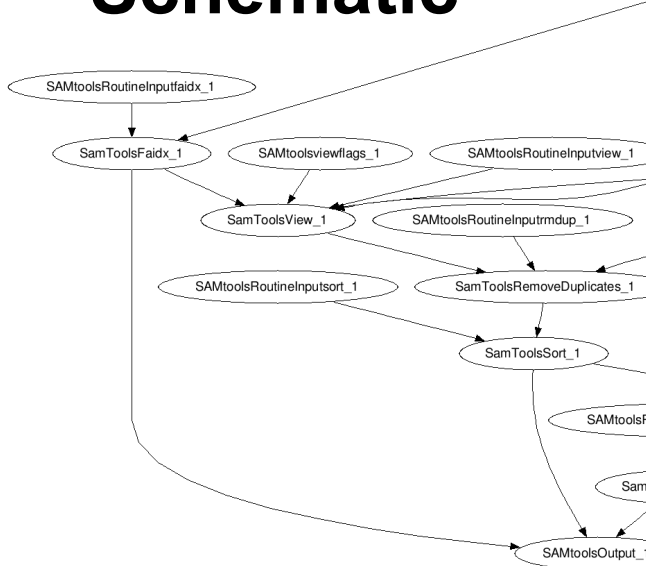
BIRN Approach

- Provide a distributed graphical interface for advanced sequence data processing, integration of diverse datasets, computational tools and web-services
- Promote community-based protocol validation, and open sharing of knowledge, tools and resources
- Utilize the LONI Pipeline, a Distributed graphical workflow environment, to provide:
 - a mechanism for interoperability of heterogeneous informatics and genomics tools
 - a complete study design, execution and validation infrastructure, supporting community result replication

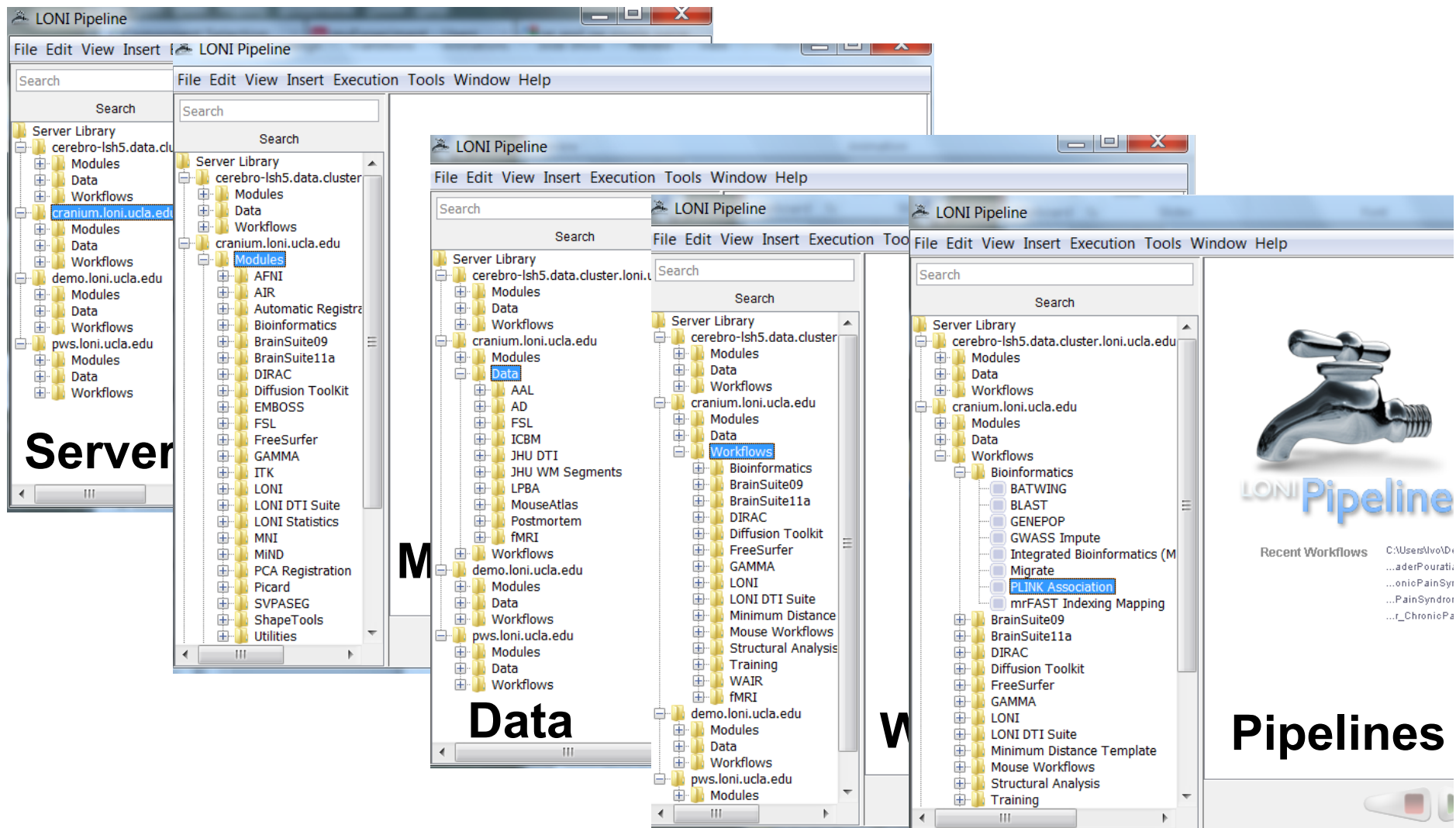
Next Generation Sequence Analysis

- **Step I Analysis: Mapping and Assembly with Qualities (MAQ), SAMtools, Bowtie, CNVer**

Workflow Schematic

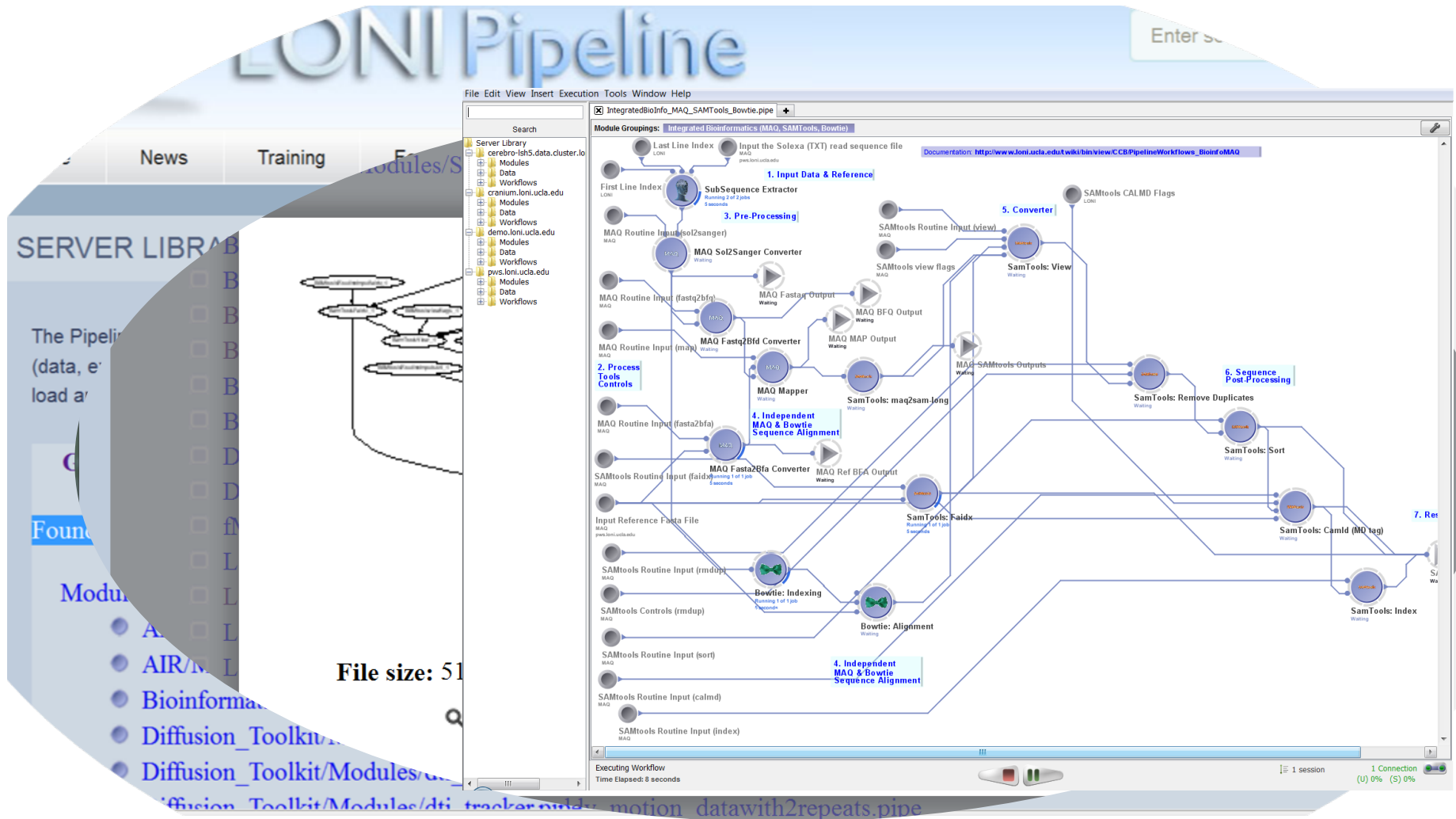


Pipeline Library of Solutions



Pipeline Web-Start (PWS)

<http://pipeline.loni.ucla.edu/PWS>



Community Pipeline Workflows

<http://www.MyExperiment.org/workflows>



[About](#) | [Mailing List](#) | [Publications](#)



Log in



Register



Give us Feedback



[Home](#)

[Users](#)

[Groups](#)

[Workflows](#)

[Files](#)

[Packs](#)

All

Home » Workflows

[+ BOOKMARK](#) [f](#) [t](#) [e](#) ...

New/Upload

Workflow

Log in / Register

Username or Email:

Password:

Remember me: ☐

OR

Use OpenID:



(eg: name.myopenid.com)

Need an account?
[Click here to register](#)

Workflows

Search filter terms

[« previous](#) [1](#) [2](#) [next »](#)

Sort by: Rank

Filter by type

- ☐ Taverna 2 583
- ☐ Taverna 1 559
- ☐ RapidMiner 146
- ☐ Bioclipse Script... 33
- ☐ GWorkflowDL 24
- ☒ LONI Pipeline 20
- ☐ Kepler 17
- ☐ BioExtract Server 16
- ☐ Trident (Packa... 10
- ☐ Chemistry Plan 7

Filter by tag

Showing 20 results. Use the filters on the left and the search box below to refine the results.

Type: LONI Pipeline ☒

[Remove all filters](#)

LONI Pipeline



EMBOSS-Matcher Pipeline Workflow (v1)

[View](#)

Created: 05/04/11 @ 21:34:49 | Last updated: 14/07/11 @ 01:52:02

[Download \(v1\)](#)

Credits: Pipeline

Original Uploader

License: Creative Commons Attribution-Share Alike 3.0 Unported License



This page contains a number of bioinformatics workflows based on the advanced EMBOSS informatics package. This page describes as Pipeline modules several bioinformatics tools from the EMBOSS suite and demonstrates the

Working with BIRN



Contact Information

Email address: info@birncommunity.org

BIRN Representatives

Joe Ames: jdames@uci.edu

Karl Helmer: helmer@nmr.mgh.harvard.edu

- Web: www.birncommunity.org
- Wiki: <https://wiki.birncommunity.org:8443/>

BIRN is supported by NCRR and NIH grants 1U24-RR025736, U24-RR021992, U24-RR021760 and by the Collaborative Tools Support Network Award 1U24-RR026057-01.