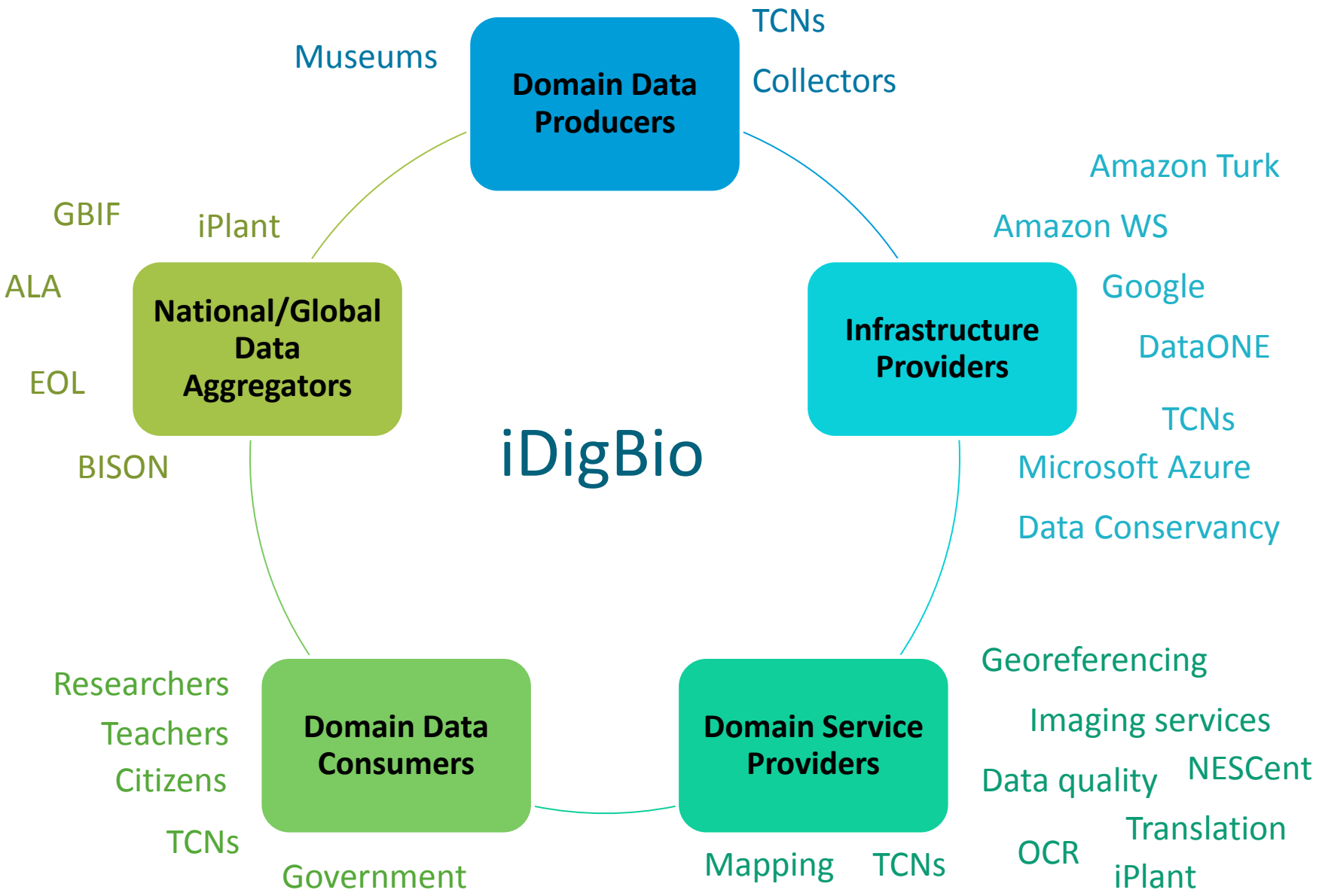
iDigBio Technology, Cloud and Appliances

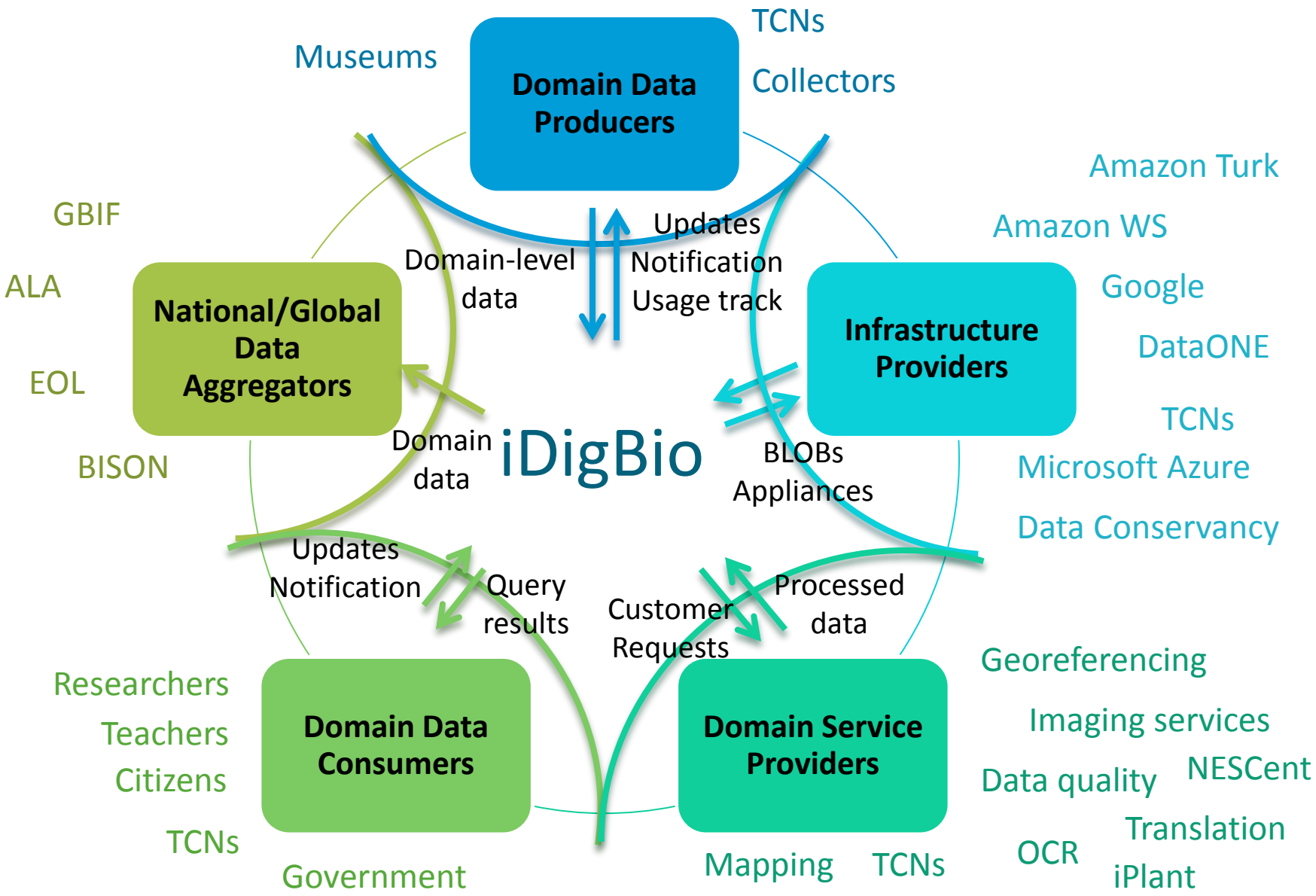
Jose Fortes
(on behalf of the iDigBio IT team)

iDigBio External Advisory Board Meeting
2012 (Project Year 1)
Supported by NSF Award EF-1115210

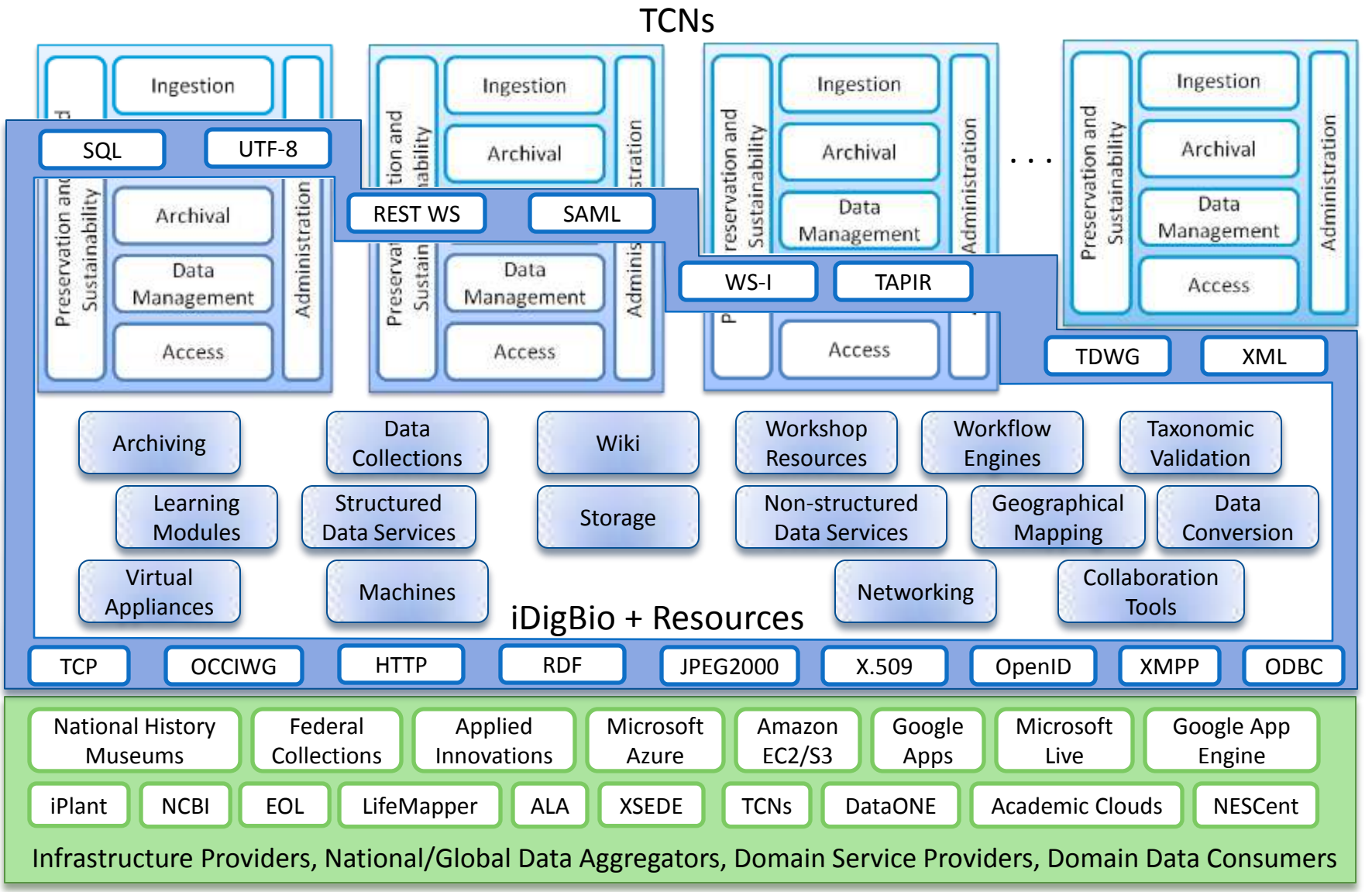
CI Stakeholders



Stakeholders APIs

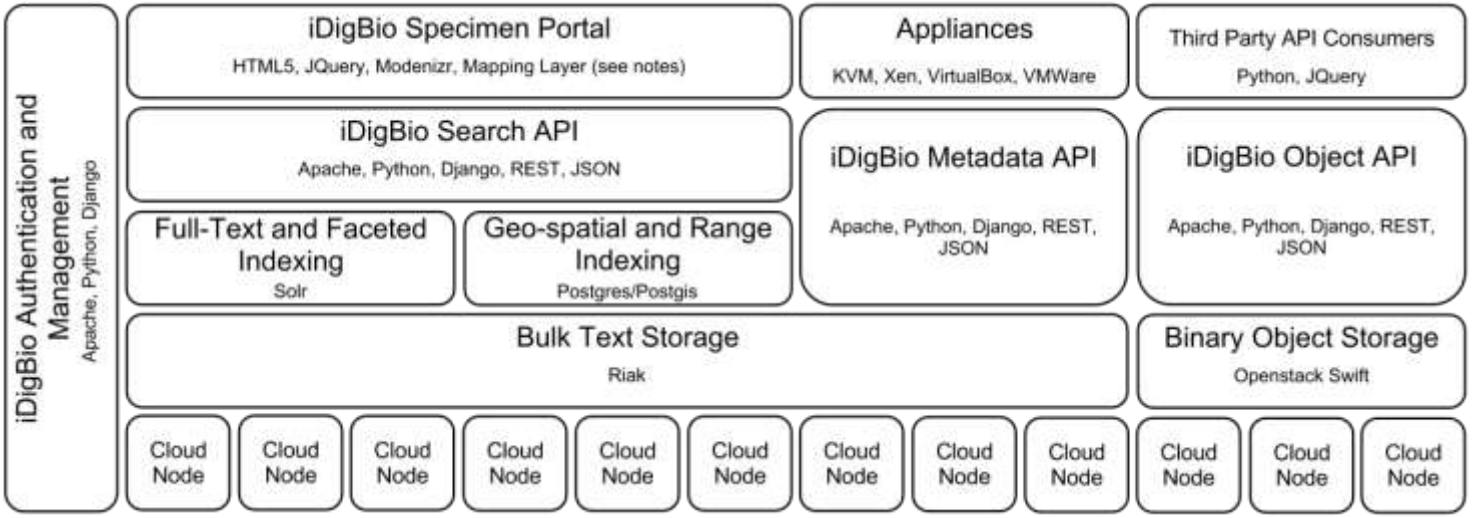


Interface Model for iDigBio and TCNs



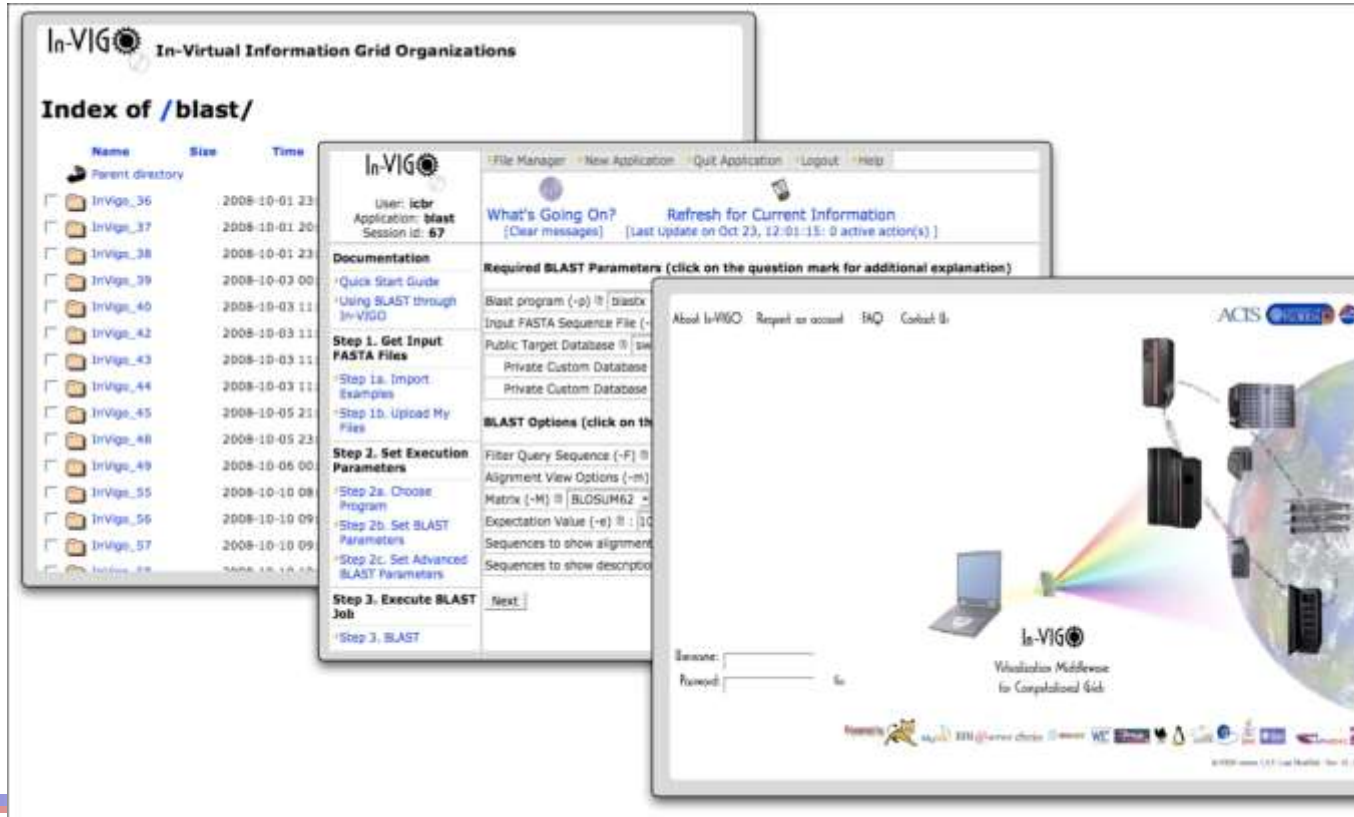
Building the iDigBio Cloud

- Cloud-based strategy
  - Providing useful services/APIs (programmatic and web-based)
  - Federated scalable object storage and information processing
  - Digitization-oriented virtual appliances
  - Reliance on standards, proven solutions and sustainable software
- Continuous consultation with stakeholders
  - Surveys, workgroups, summit/workshops, person-to-person …



Unique UF+FSU record

- Track record of building cyberinfrastructure
  - PUNCH and In-VIGO
    - Nanohub, Netcare, In-VIGOBlast …
  - Morphbank
  - AFRESH
  - Telecenter
  - Archer

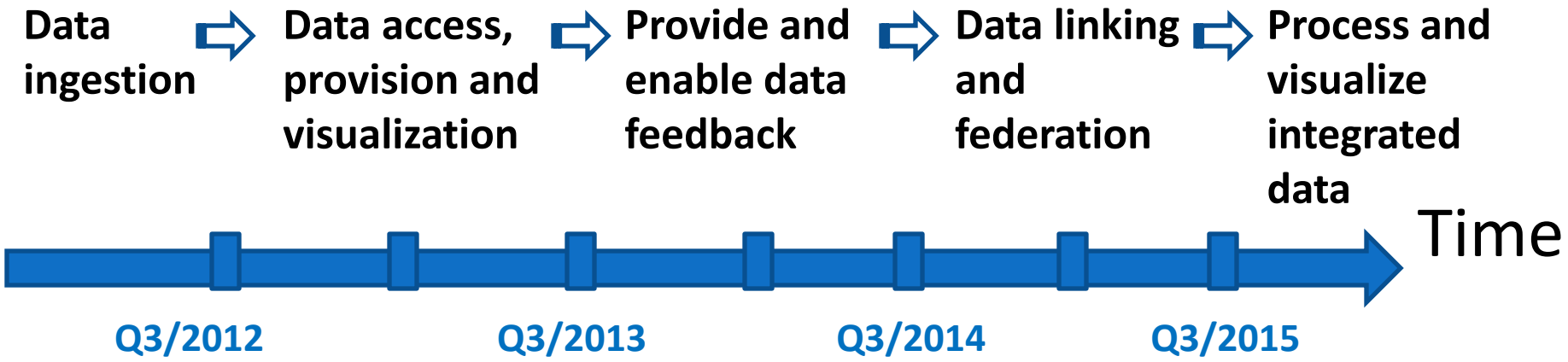


Keeping our eyes on the ball

Common/frequent needs: archival storage, server hosting, feedback on the data, data intensive transformations …

10-year tsunami of requirements: from being on Facebook to multilingual search-and-compute across multiple data sets…

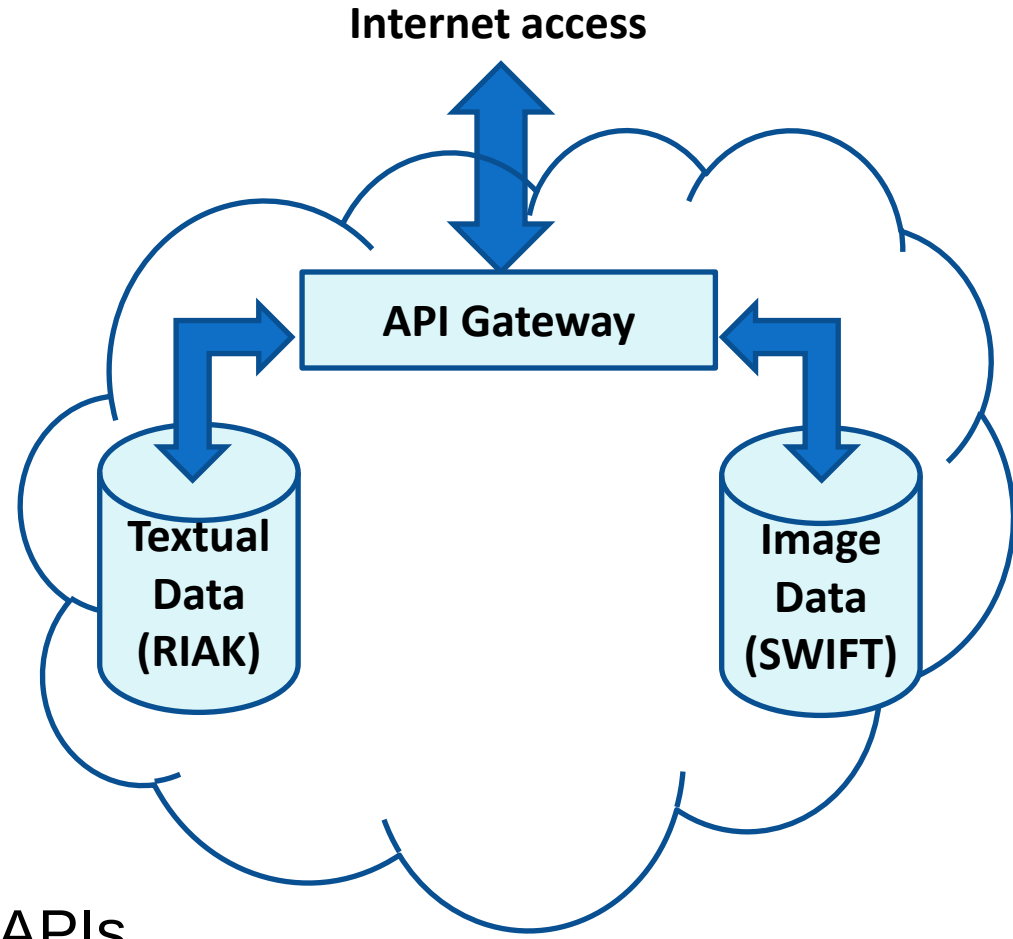
Evolution of iDigBio capabilities



Increasing storage and server hosting in support of the above
Increasing number of appliances in support of the above
Web site for interaction with public, community, education and above

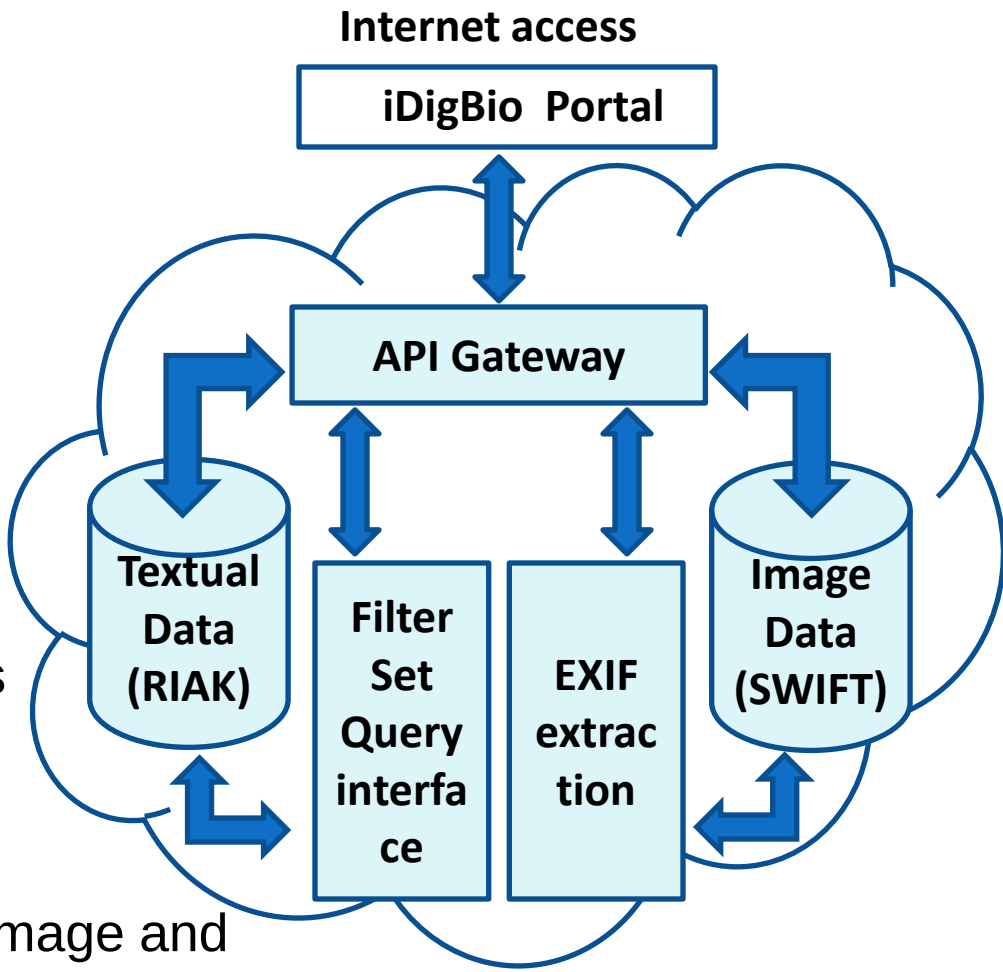
Near-term goals: ingest data

- Textual data
  - JSON document database
  - Data ingestion via DwC-a files
  - Get / Set API
- Image Data
  - Internet-accessible object storage
  - Upload appliance
  - Limited access to low-level APIs

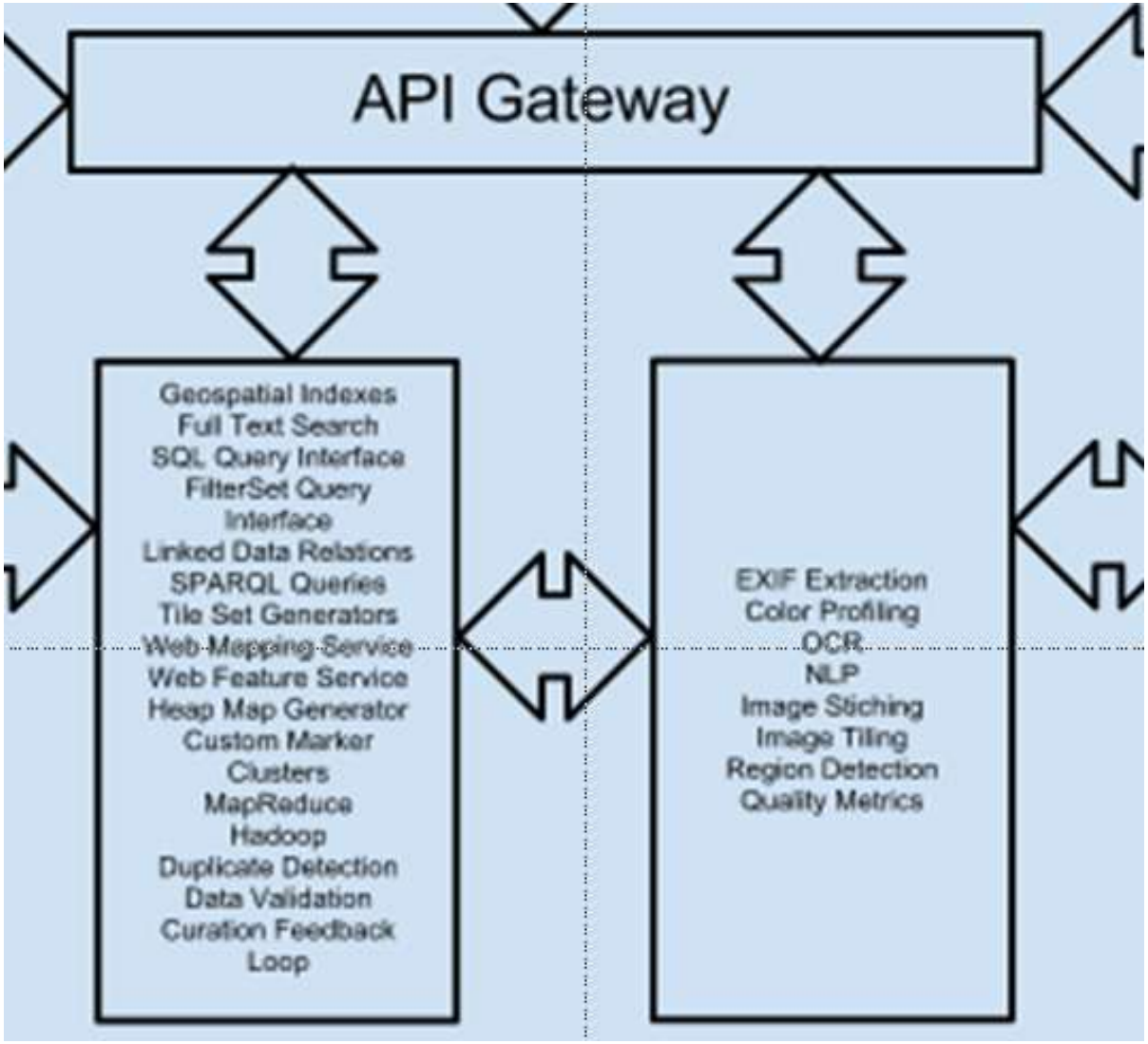


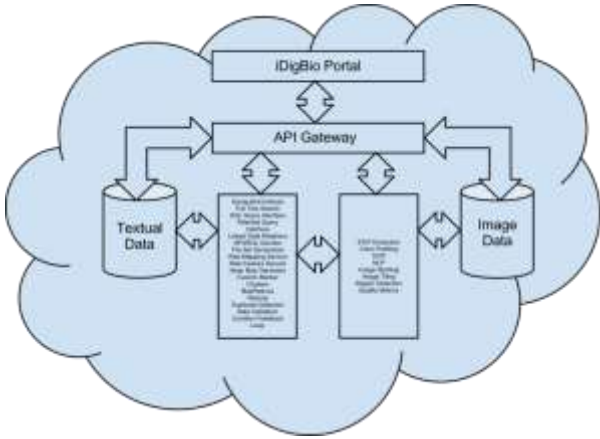
Medium-term goals

- Textual Data
  - JSON document database
  - Data Ingestion via DwC-a files
  - Rich RESTful API
- Image Data
  - Web-accessible object storage
  - Upload appliance
  - Fully abstracted storage
- Indexing and Search
  - Extract EXIF data from images
  - Limited but useful set of indexes
  - Intuitive search UI
  - Search available via API
- Portal
  - Consumes and interfaces text, image and search APIs (minimal server side code)
  - Web-based mapping - client side javascript limits useable record count to about 50k records at a time.



(Very) Long-term Goals

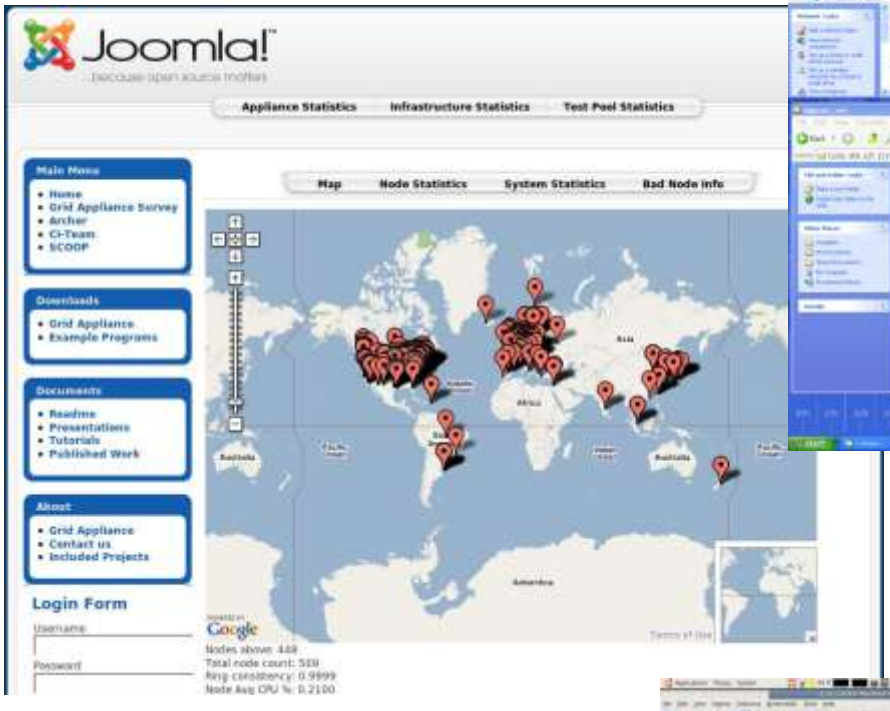




Virtual Appliances in iDigBio

- Packaging of software and dependences in virtual machines
  - End user/desktop (e.g. VMware, Virtualbox)
  - Infrastructure-as-a-Service clouds (e.g. OpenStack)
  - Enhance user experience, facilitate integration with cloud
- Image ingestion appliances (short term)
  - Batch upload of images from a local storage to cloud
  - Generate GUID/URLs for later processing
  - Reliable transfers using cloud APIs (e.g. Swift/iDigBio)
- Post-processing appliances (OCR tools; end-user or batch)
- Geo-referencing appliances (Training/verification)
- Research appliances (Data-intensive/batch workflows)

Archer cyber-infrastructure



Custom appliance image for computer architecture community

Hundreds of distributed compute/routers nodes
24/7 operation, 650+ cores

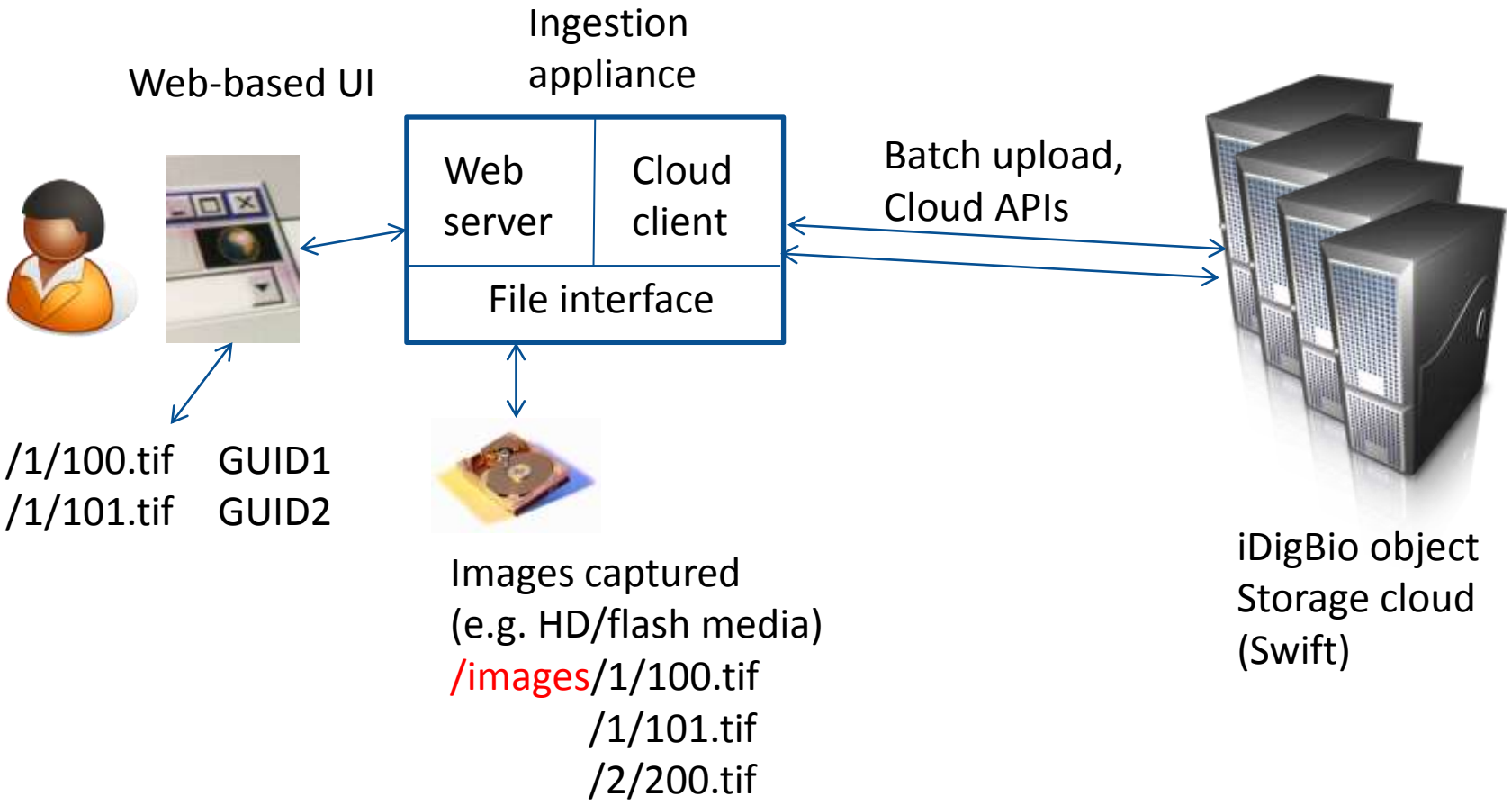
Job scheduling across participating institutions

Now: appliance proposal process

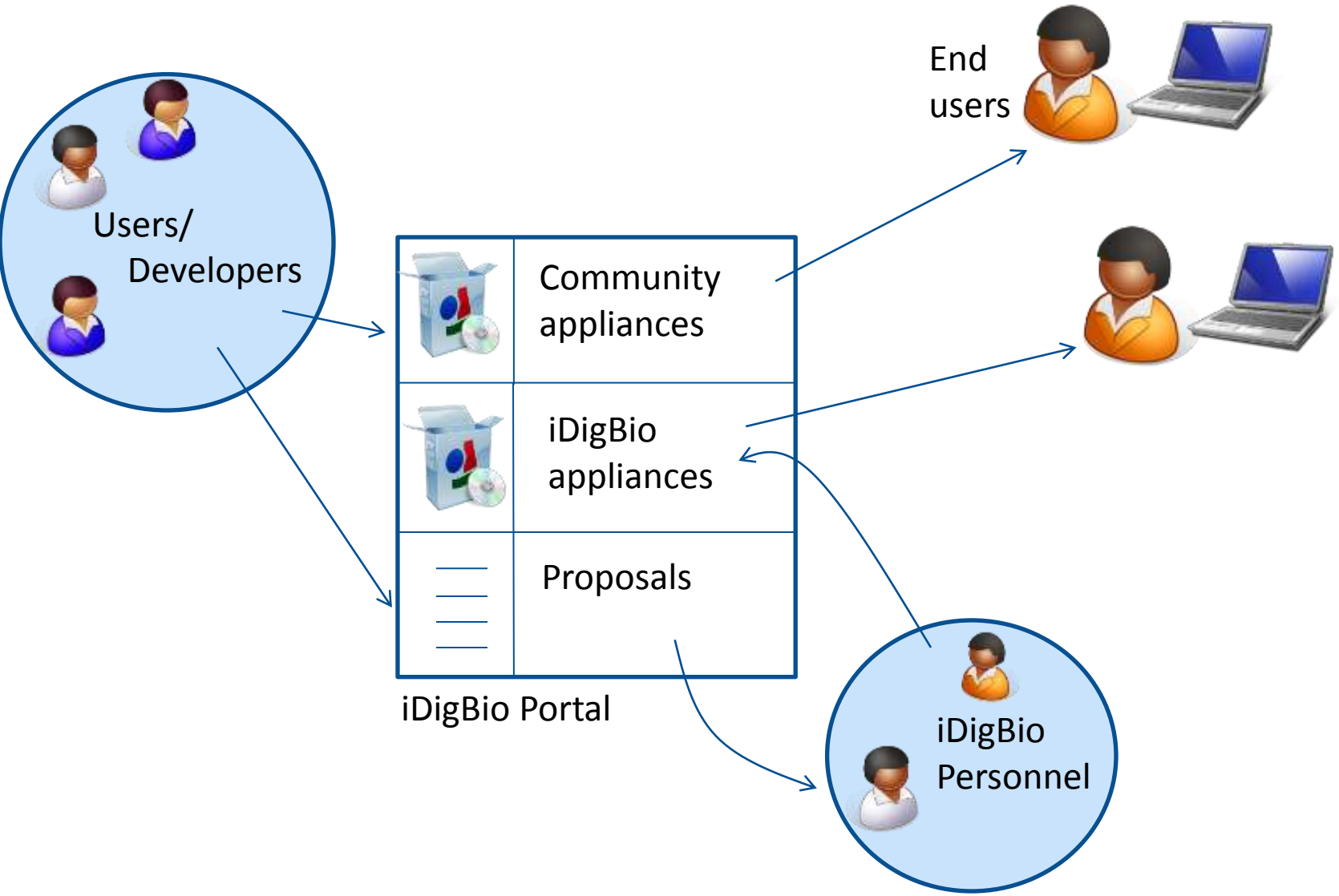
- By users/developers through the iDigBio Web portal
  - Requirements – demonstrates usage/buy-in, software license, documentation, etc
- Queue of appliances for integration
  - iDigBio will prioritize and work with developers
- Leverage expertise in appliance development
  - Focus on images that users can download and run on VMware, Virtualbox
  - Application, in addition to appliance, if applicable/desirable

Short term

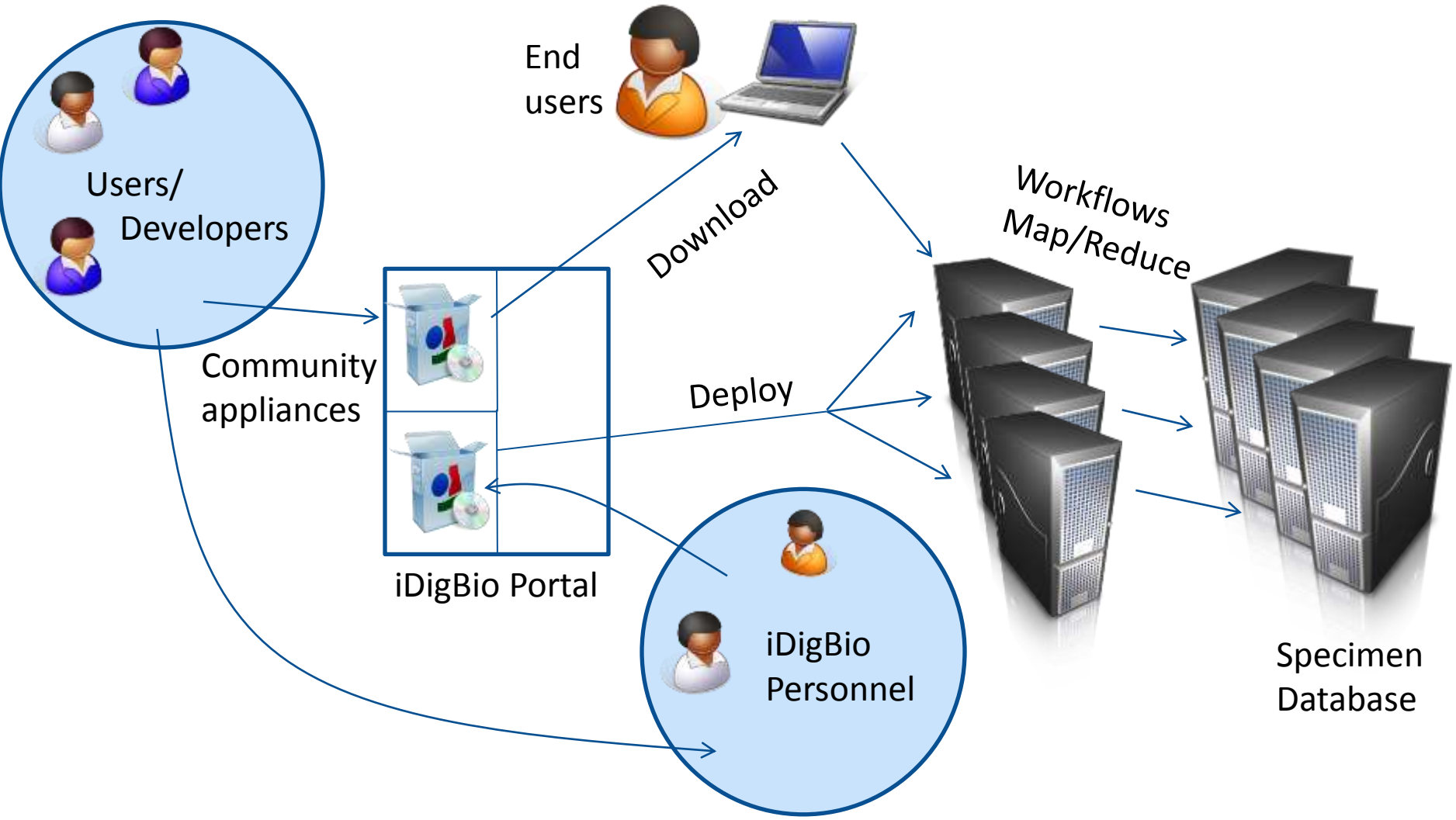
- Facilitate data ingestion, interface with iDigBio
- Tools identified by community in workshops/groups



Medium-term – “Marketplace”



Long-term – information processing



Summary

- iDigBio cloud
  - Service-oriented standards-based cyberinfrastructure focused on the ADBC community needs
  - Scalable data management and information processing using standard interfaces, data formats, protocols, tools
- Toolboxes as appliances
  - Evolving collection of community-selected tools
  - Built-in interfaces for effortless iDigBio integration
  - Embedded best practices and standards in biocollections work
- Software re-use when open-source, well maintained, manageable, sustainable and efficient to re-purpose
- Feedback and suggestions welcome
  - fortes@ufl.edu and “Contacts” at idigbio.org

Acknowledgments

- National Science Foundation
  - Judith Skog and Anne Maglia

- IDigBio team at University of Florida and Florida State University

Extras

iDigBio IT Vision

- Cyberinfrastructure to enable
  - the collaborative creation, integration and management of digitized biocollections,
  - their use in scientific research, education and outreach
- Visible as a collection of persistent Internet-accessible services, data and resources
  - For biocollection “producers”
  - For biocollection “consumers”
  - For biocollection service providers
  - For cyberinfrastructure providers
  - For national/global data aggregators