iDigBio Technology, Cloud and Appliances

Jose Fortes
(on behalf of the iDigBio IT team)

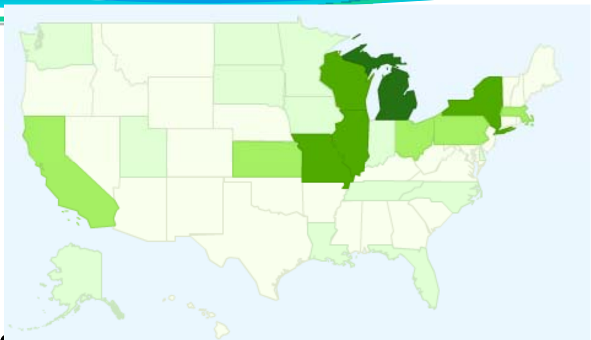
iDigBio
Integrated Digitized Biocollections

NSF

Biodiversity Workshop
Melbourne, Australia
April 16, 2012

Supported by NSF Award EF-1115210

iDigBio (idigbio.org)



- **Goal**: making data and images for millions of biological specimens available in electronic format for the biological research community, agencies, students, educators, and public

- **Mission:** leadership, coordination, and outreach in digitization of collections by implementing resources for communication, use of technology, access to data, research and education.
  - The “Hub” part of the NSF ADBC program aggregating TCNs and PENs
- A **resource**: permanent cloud computing infrastructure
  - to link biological data from collections across the USA
  - to use search and analytics tools to mine and reference data

BIODIVERSITY
HOTSPOTS
(Use the menu to see all the Hotspots by name.)
HOTSPOTS MENU
(Move your mouse over each Hotspot to learn more about that region.)
HOTSPOTS EXPLORER

Research Questions

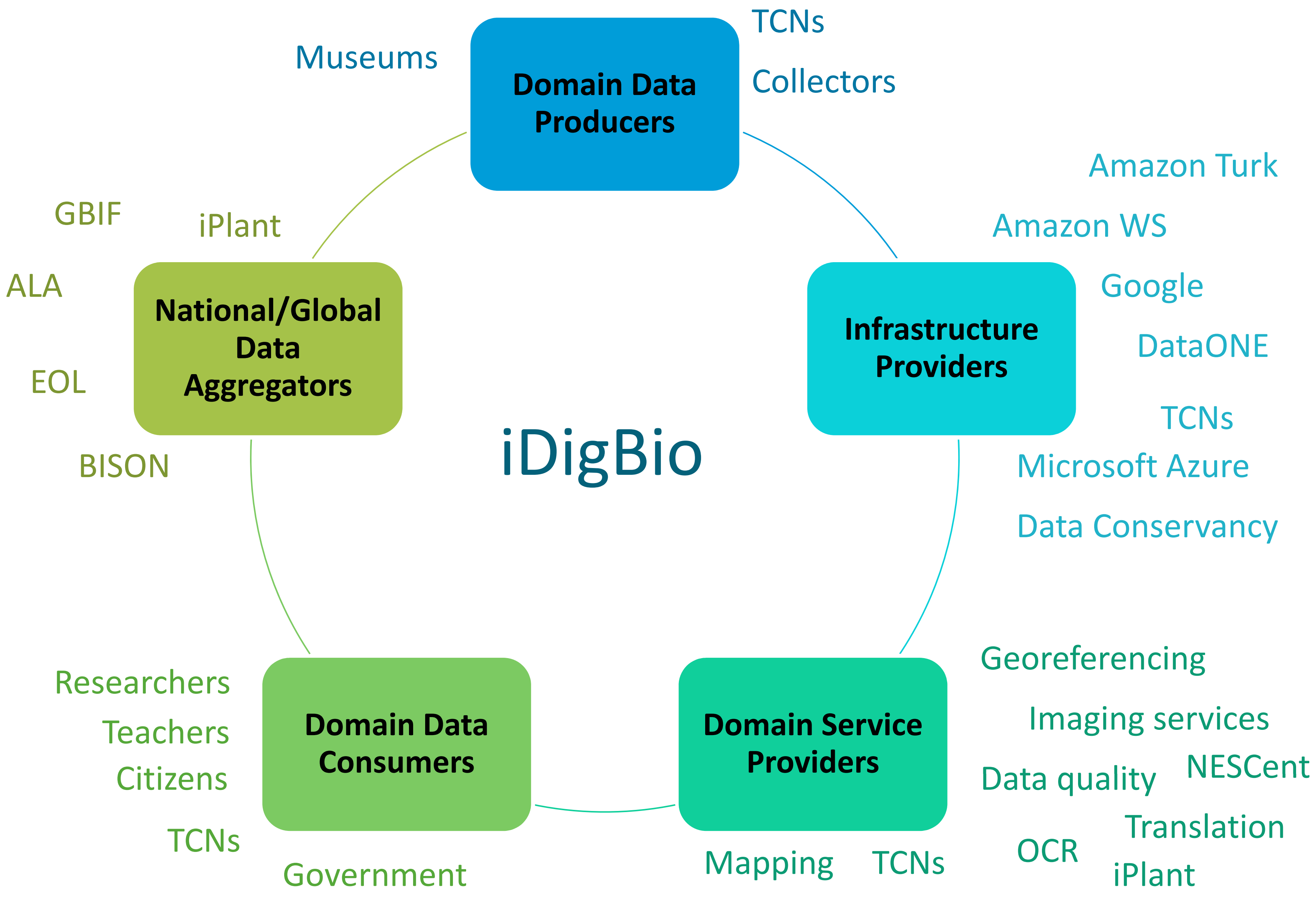
- How are species distributed in geographical and ecological space?
- What is the history of life on Earth?
- What factors lead to speciation, dispersal, and extinction?
- What are the impacts of climate change likely to be?
- What information is needed for effective conservation strategies?

Slide provided by Pam Soltis

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

CI Stakeholders



iDigBio

Domain Data Producers

Museums

TCNs

Collectors

Infrastructure Providers

Amazon Turk

Amazon WS

Google

DataONE

TCNs

Microsoft Azure

Data Conservancy

Domain Service Providers

Georeferencing

Imaging services

Data quality

NESCent

Translation

OCR

iPlant

Mapping

TCNs

Domain Data Consumers

Researchers

Teachers

Citizens

TCNs

Government

National/Global Data Aggregators

GBIF

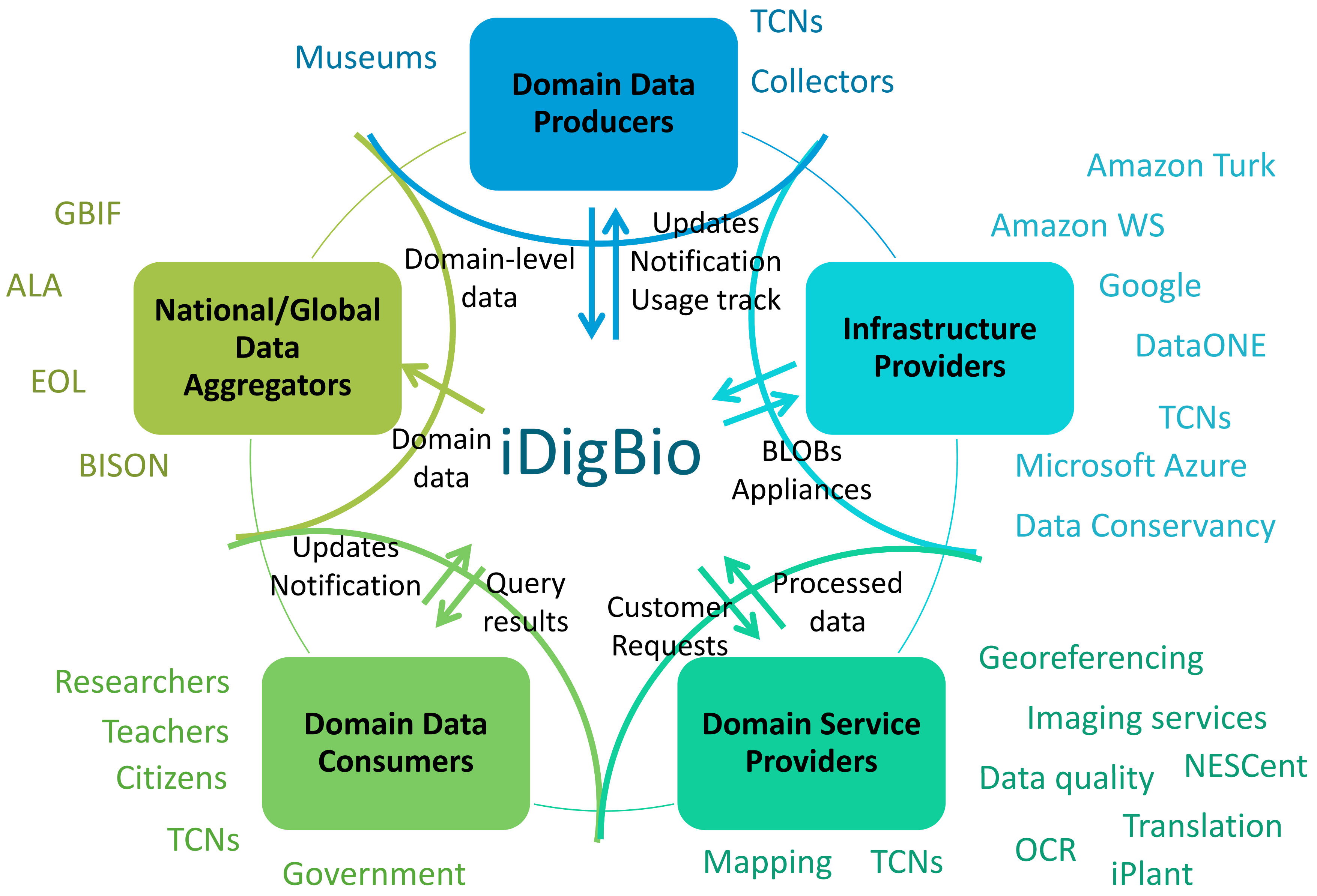
iPlant

ALA

EOL

BISON

Stakeholders APIs



Museums

Domain Data Producers

TCNs

Collectors

Domain-level data

Updates
Notification
Usage track

Amazon Turk

Amazon WS

Google

DataONE

Infrastructure Providers

TCNs

Microsoft Azure

Data Conservancy

BLOBs
Appliances

iDigBio

GBIF

ALA

EOL

BISON

National/Global Data Aggregators

Domain data

Updates
Notification

Query results

Customer Requests

Processed data

Researchers

Teachers

Citizens

TCNs

Domain Data Consumers

Government

Domain Service Providers

Georeferencing

Imaging services

Data quality

NESCent

Translation

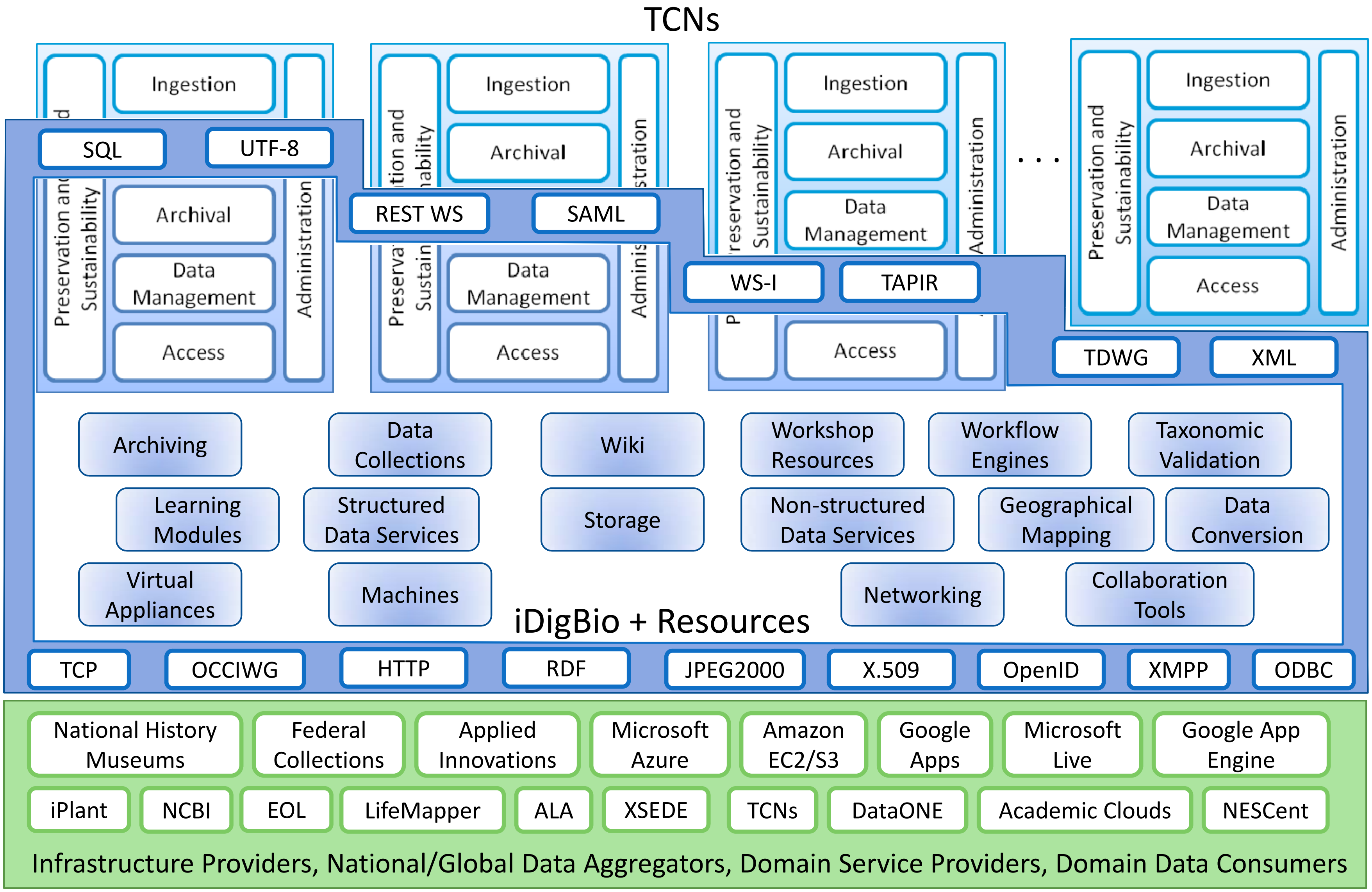
OCR

iPlant

Mapping

TCNs

Interface Model for iDigBio and TCNs



TCNs

Ingestion
Archival
Data Management
Access
Preservation and Sustainability
Administration

Ingestion
Archival
Data Management
Access
Preservation and Sustainability
Administration

Ingestion
Archival
Data Management
Access
Preservation and Sustainability
Administration

. . .

Ingestion
Archival
Data Management
Access
Preservation and Sustainability
Administration

SQL
UTF-8
REST WS
SAML
WS-I
TAPIR
TDWG
XML

Archiving
Data Collections
Wiki
Workshop Resources
Workflow Engines
Taxonomic Validation
Learning Modules
Structured Data Services
Storage
Non-structured Data Services
Geographical Mapping
Data Conversion
Virtual Appliances
Machines
Networking
Collaboration Tools

iDigBio + Resources

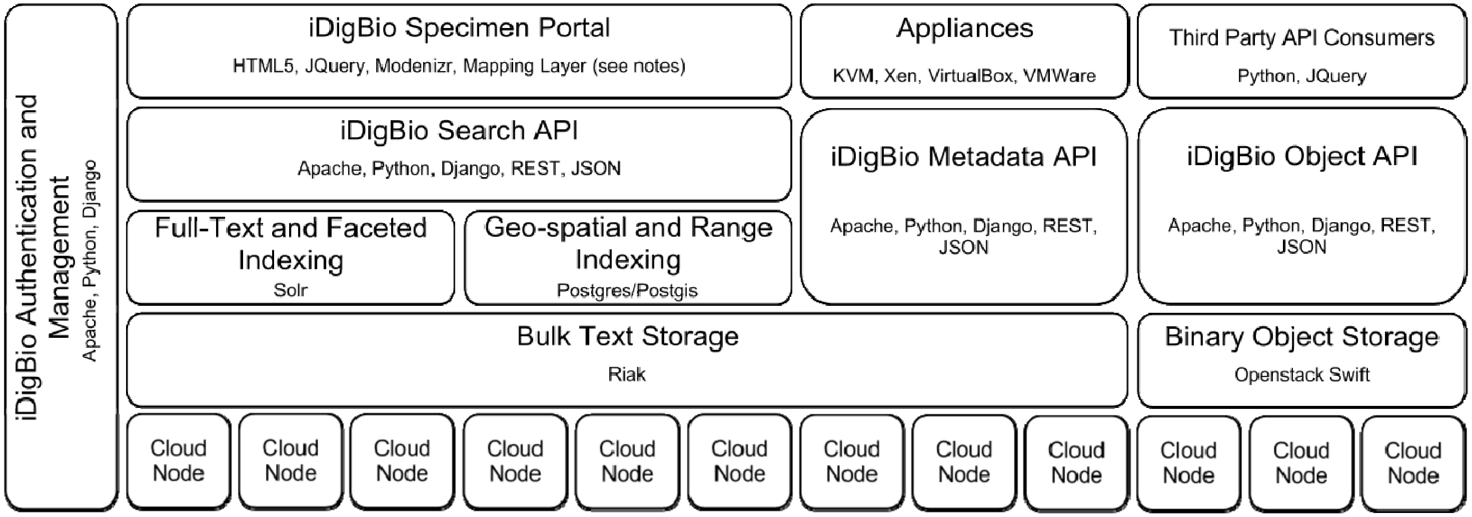
TCP
OCCIWG
HTTP
RDF
JPEG2000
X.509
OpenID
XMPP
ODBC

National History Museums
Federal Collections
Applied Innovations
Microsoft Azure
Amazon EC2/S3
Google Apps
Microsoft Live
Google App Engine
iPlant
NCBI
EOL
LifeMapper
ALA
XSEDE
TCNs
DataONE
Academic Clouds
NESCent

Infrastructure Providers, National/Global Data Aggregators, Domain Service Providers, Domain Data Consumers

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 …



iDigBio Authentication and Management
Apache, Python, Django

iDigBio Specimen Portal
HTML5, JQuery, Modenizr, Mapping Layer (see notes)

Appliances
KVM, Xen, VirtualBox, VMWare

Third Party API Consumers
Python, JQuery

iDigBio Search API
Apache, Python, Django, REST, JSON

iDigBio Metadata API
Apache, Python, Django, REST, JSON

iDigBio Object API
Apache, Python, Django, REST, JSON

Full-Text and Faceted Indexing
Solr

Geo-spatial and Range Indexing
Postgres/Postgis

Bulk Text Storage
Riak

Binary Object Storage
Openstack Swift

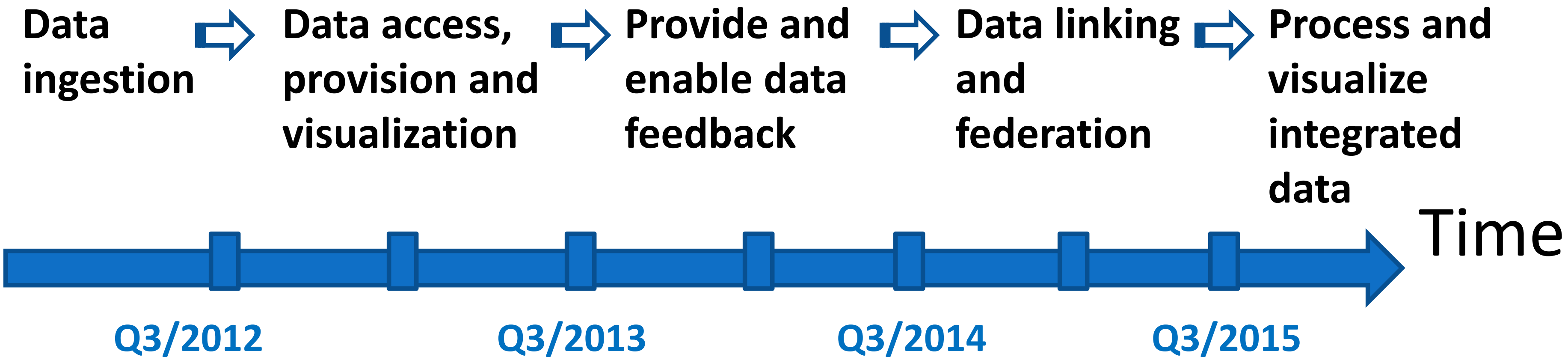
Cloud Node
Cloud Node
Cloud Node
Cloud Node
Cloud Node
Cloud Node
Cloud Node
Cloud Node
Cloud Node
Cloud Node
Cloud Node
Cloud Node

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



Data ingestion ⇨ Data access, provision and visualization ⇨ Provide and enable data feedback ⇨ Data linking and federation ⇨ Process and visualize integrated data

Time

Q3/2012

Q3/2013

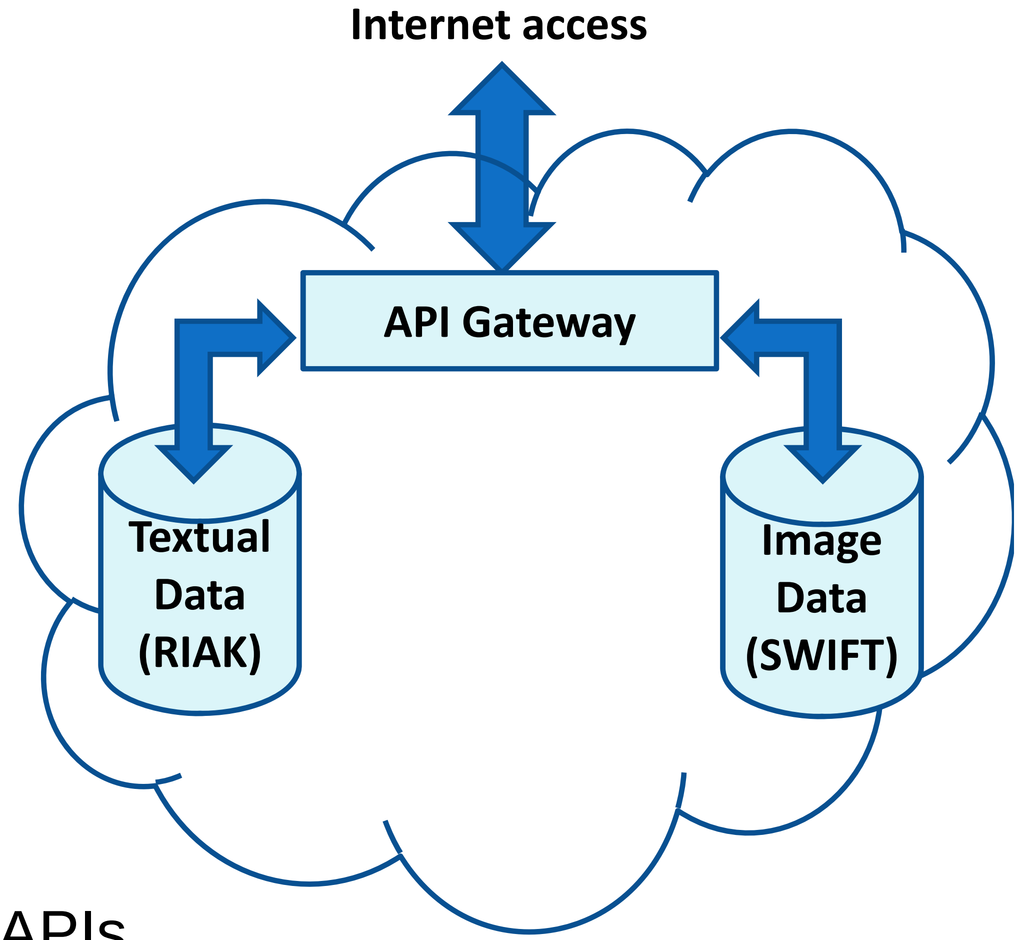
Q3/2014

Q3/2015

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

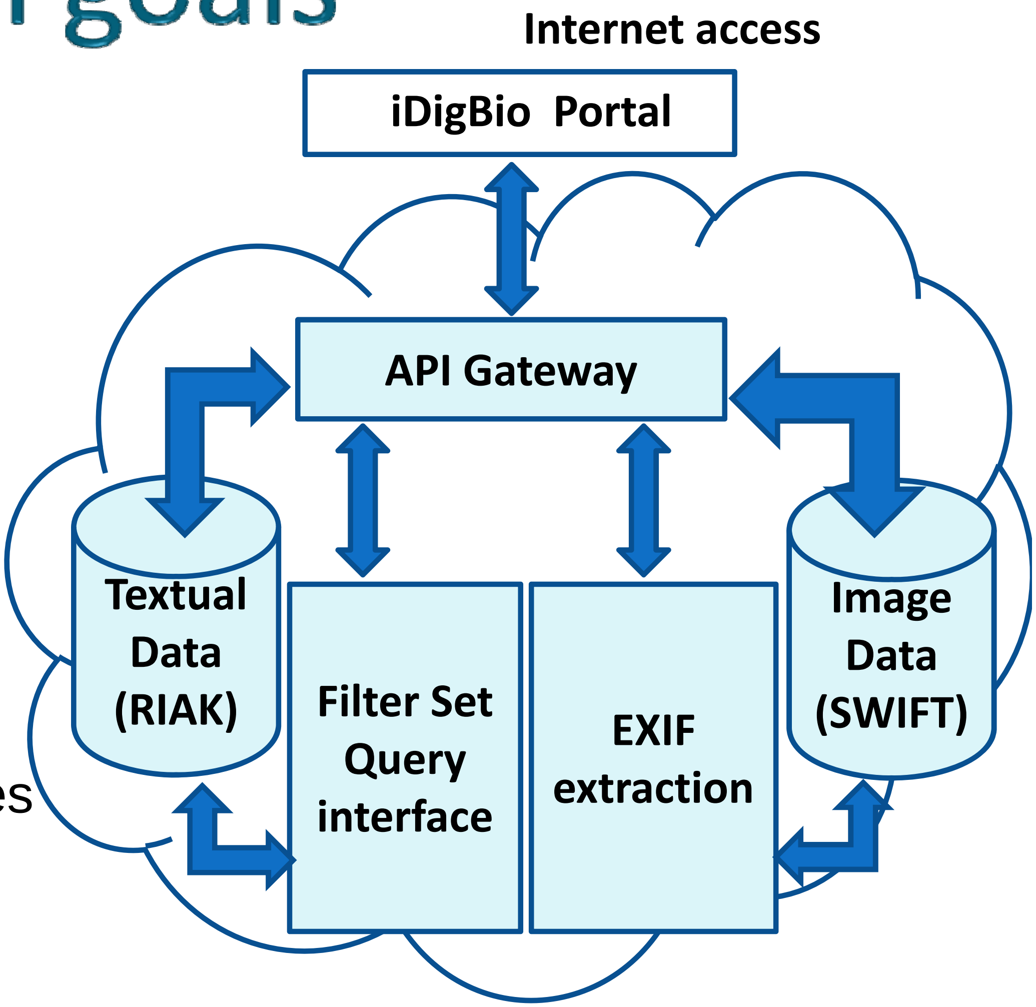
Near-term goals: ingest data

- 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

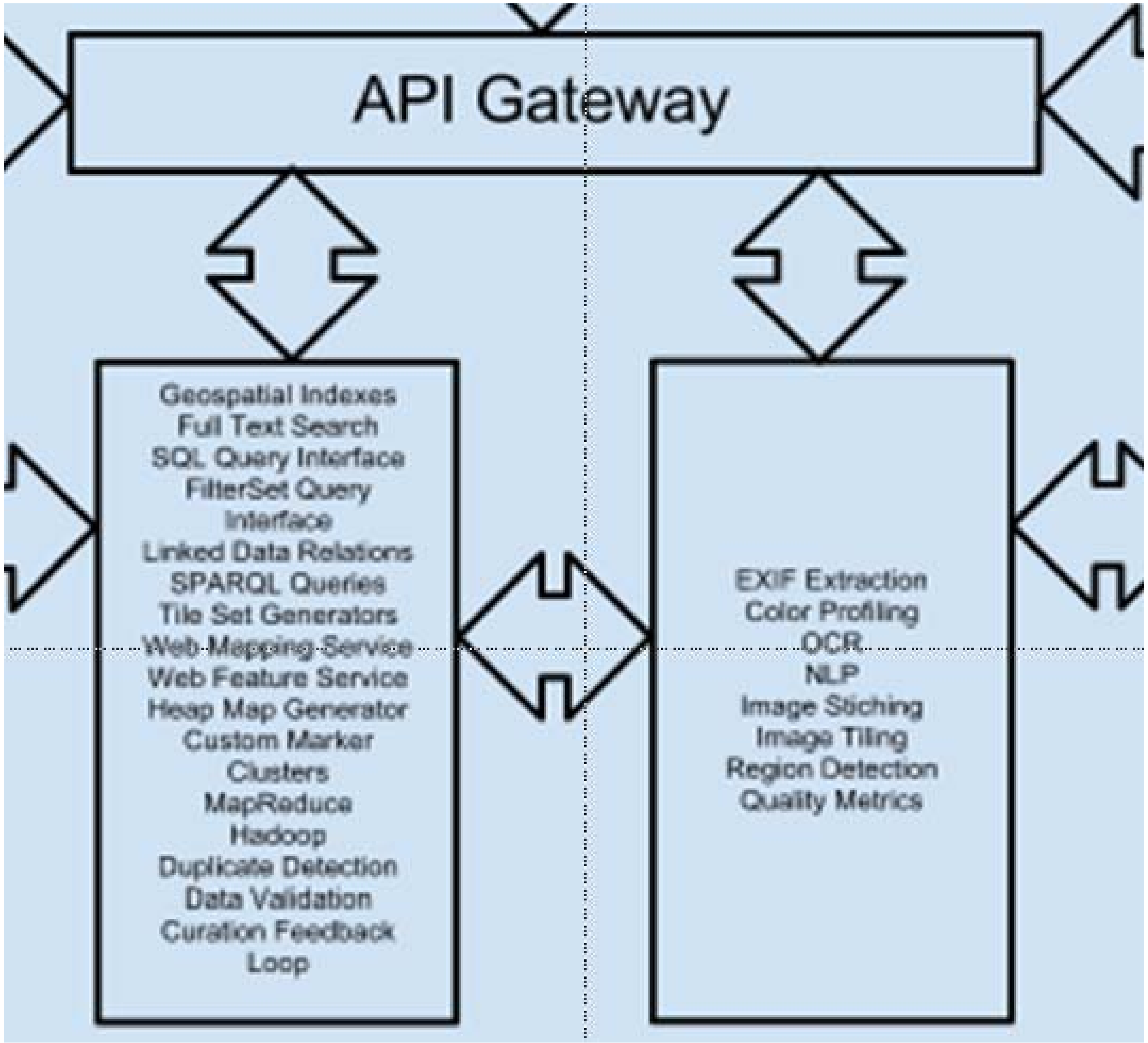


Medium-term goals

- 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.



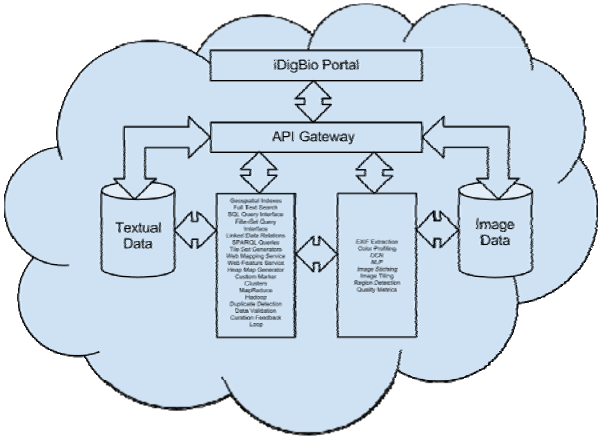
(Very) Long-term Goals



API Gateway

Geospatial Indexes
Full Text Search
SQL Query Interface
FilterSet Query
Interface
Linked Data Relations
SPARQL Queries
Tile Set Generators
Web Mapping Service
Web Feature Service
Heap Map Generator
Custom Marker
Clusters
MapReduce
Hadoop
Duplicate Detection
Data Validation
Curation Feedback
Loop

EXIF Extraction
Color Profiling
OCR
NLP
Image Stiching
Image Tiling
Region Detection
Quality Metrics



iDigBio Portal

API Gateway

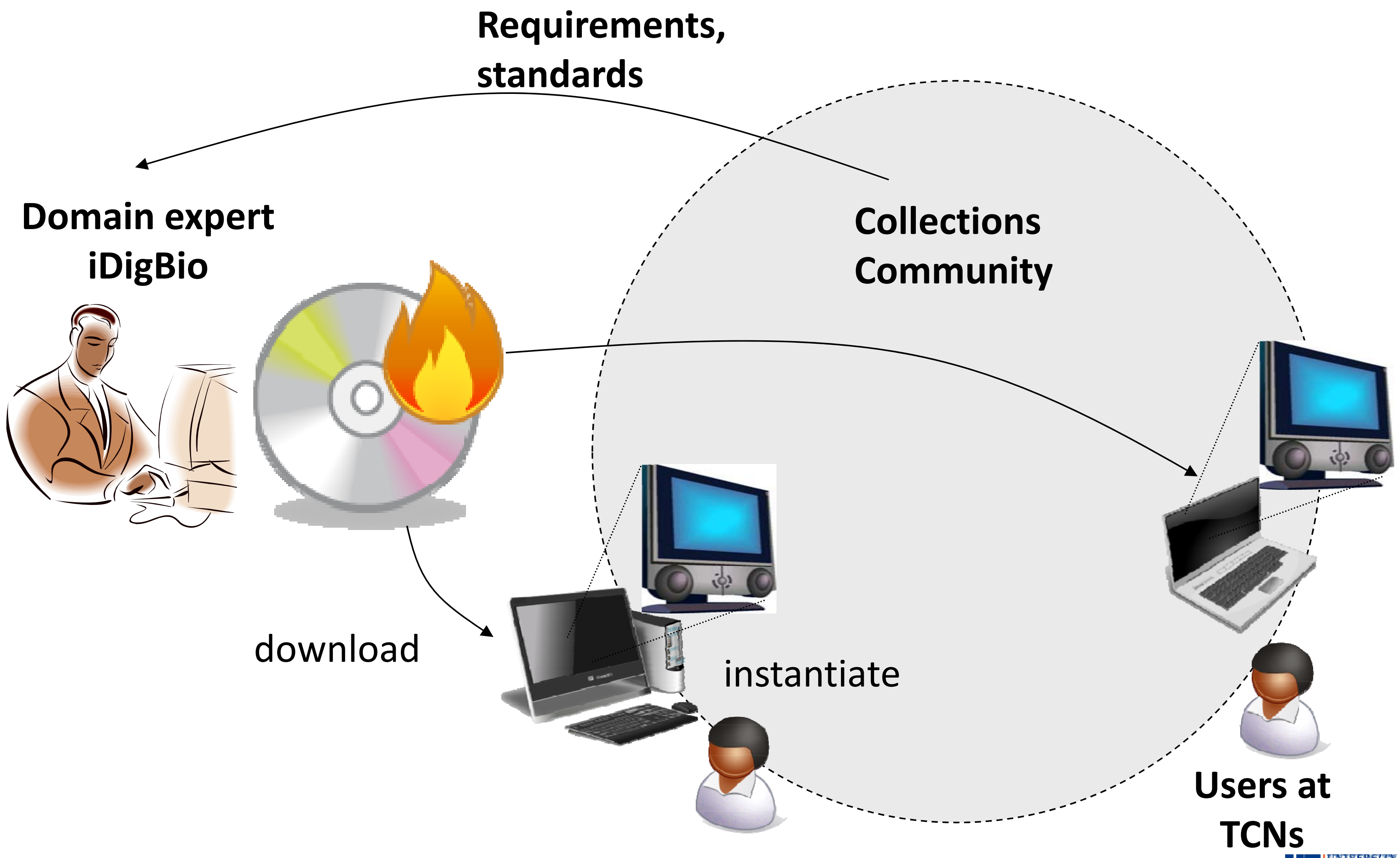
Textual
Data

Geospatial Indexes
Full Text Search
SQL Query Interface
Filter-Set Query
Interface
Linked Data Relations
SPARQL Queries
Tile Set Generators
Web Mapping Service
Web Feature Service
Heap Map Generator
Custom Marker
Clusters
MapReduce
Hadoop
Duplicate Detection
Data Validation
Curation Feedback
Loop

EXIF Extraction
Color Profiling
OCR
NLP
Image Stiching
Image Tiling
Region Detection
Quality Metrics

Image
Data

Virtual appliance cycle



Requirements, standards

Domain expert iDigBio

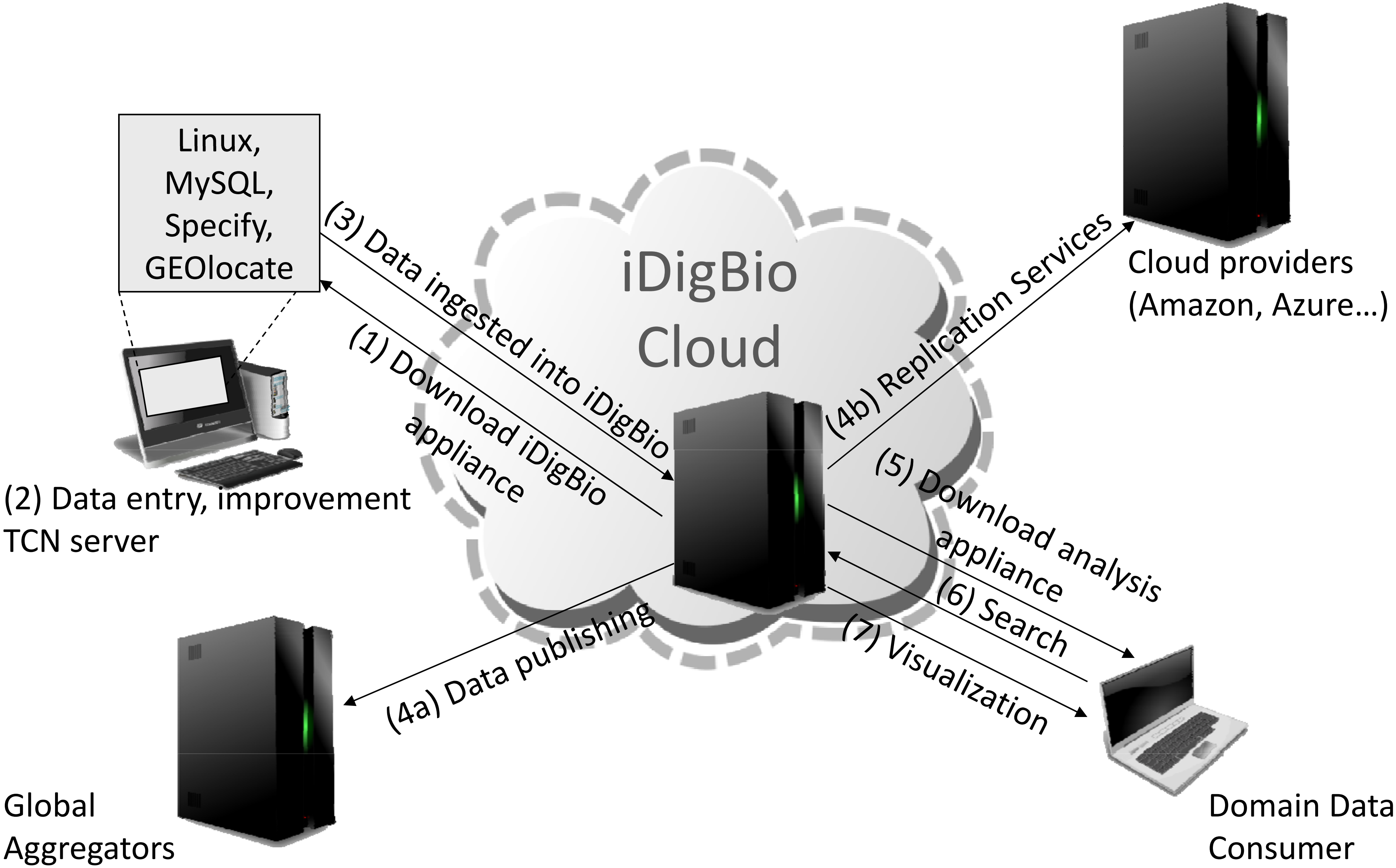
Collections Community

download

instantiate

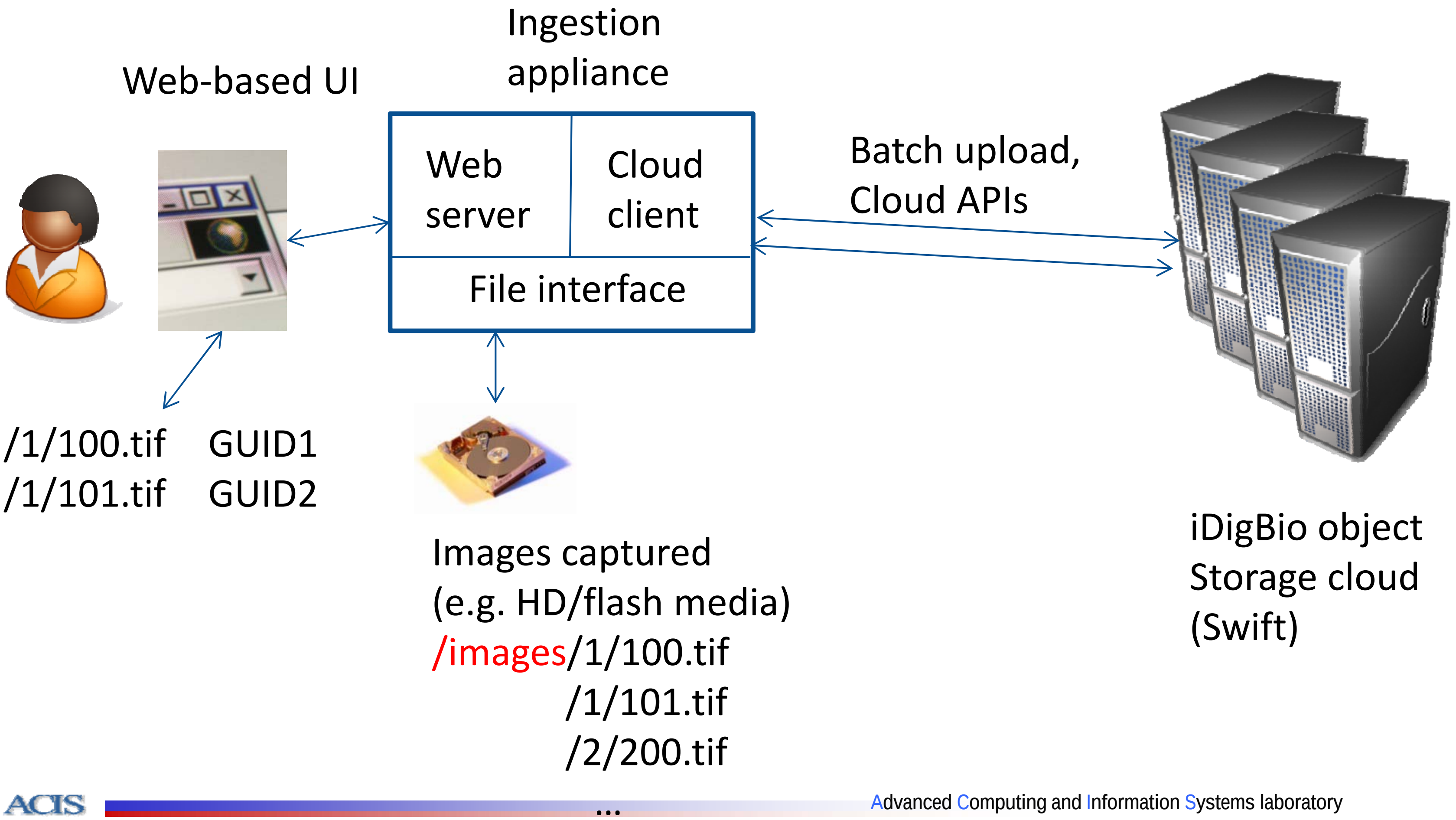
Users at TCNs

Toolbox Workflow Example



Short term

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



Web-based UI

Ingestion appliance

Web server

Cloud client

File interface

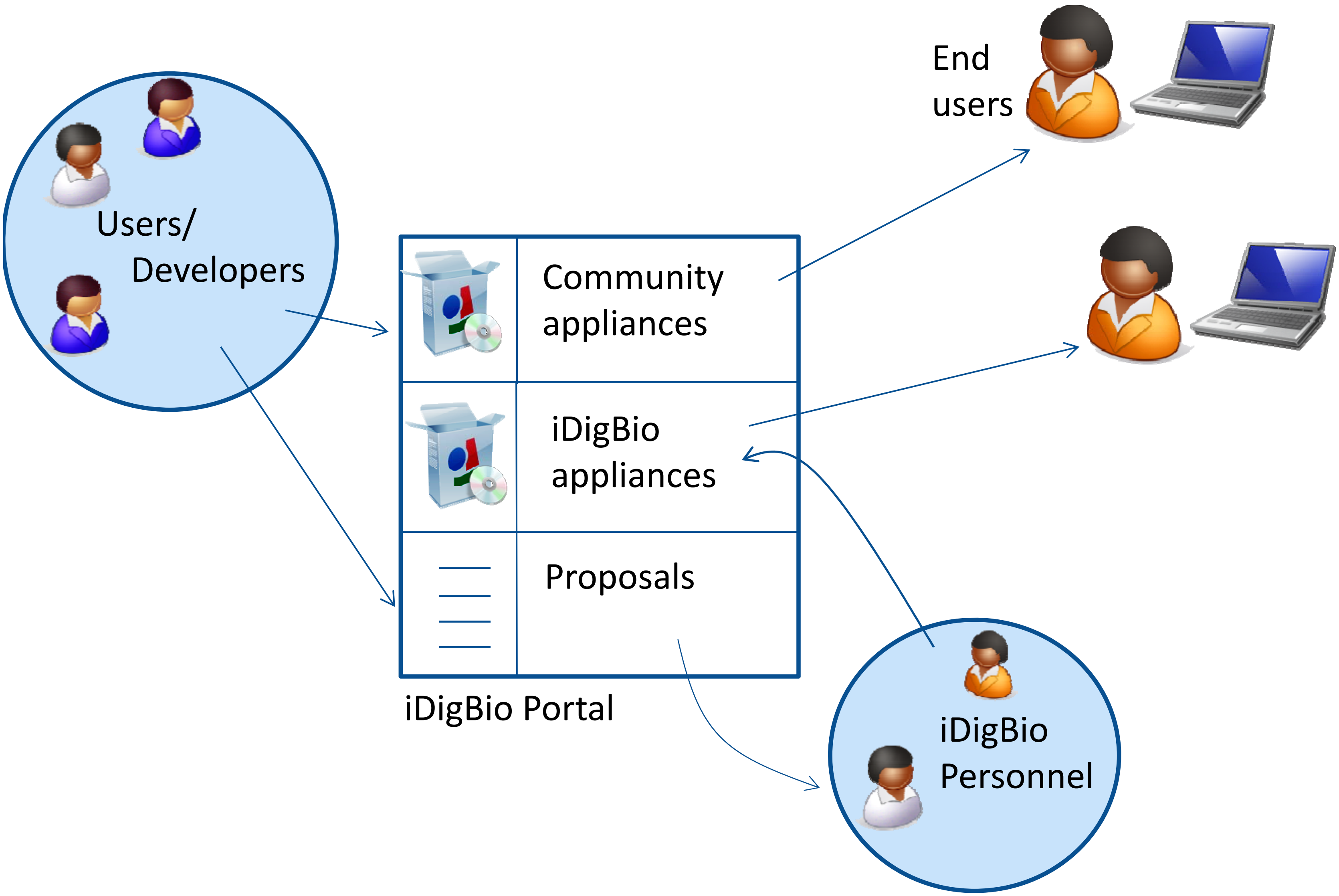
Batch upload, Cloud APIs

/1/100.tif GUID1
/1/101.tif GUID2

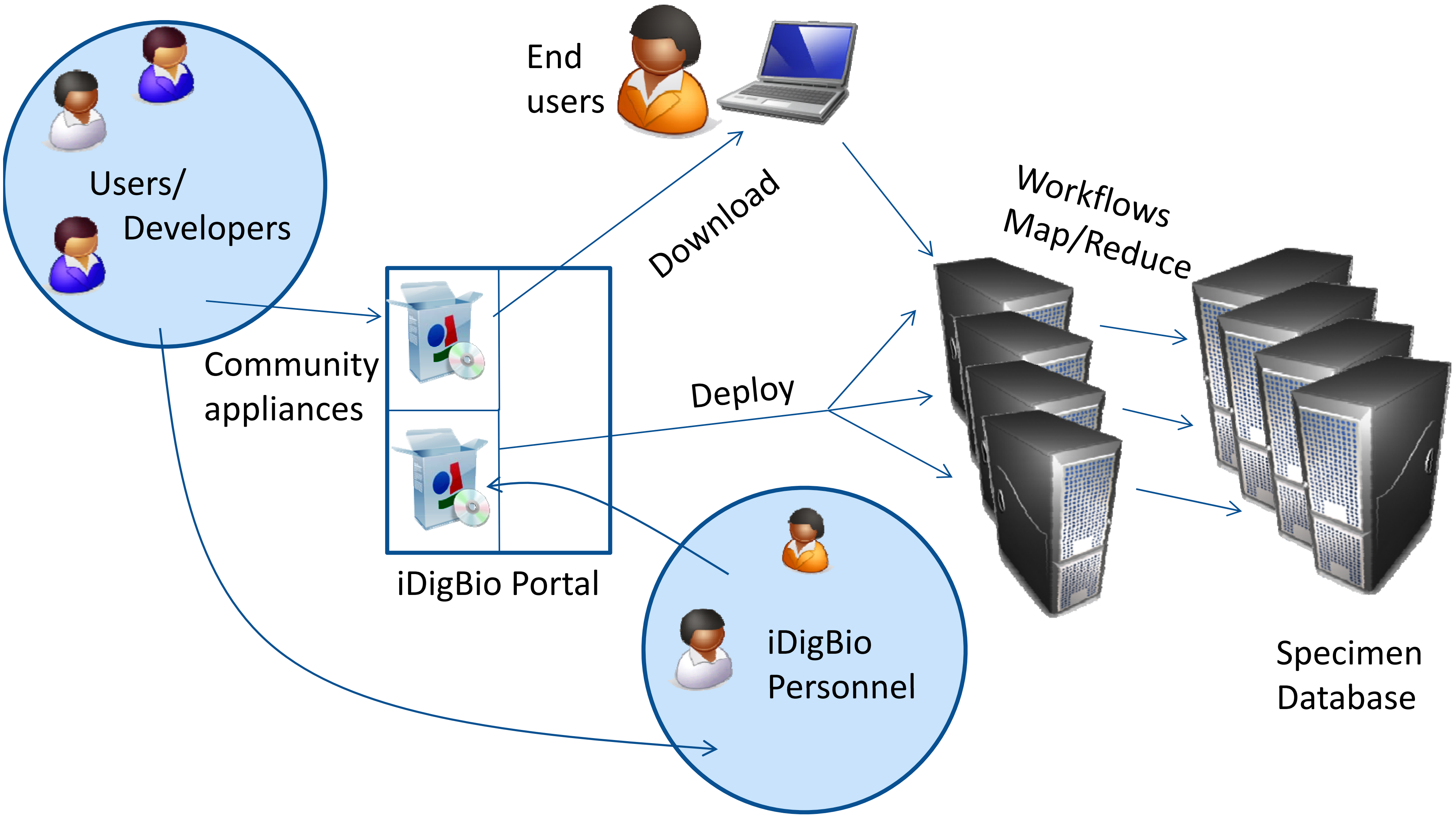
Images captured
(e.g. HD/flash media)
/images/1/100.tif
/1/101.tif
/2/200.tif
...

iDigBio object
Storage cloud
(Swift)

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