An Asset Management Approach to Continuous Integration of Heterogeneous Biomedical Data

Robert E Schuler, Carl Kesselman,
Karl Czajkowski

Information Sciences Institute,
University of Southern California

A Software as A Service ased Approach to Digital Asset Management for Complex Big-Data.

Carl Kesselman

Dept. Industrial and Systems Engineering

Information Sciences Institute, University of Southern California

BigData Landscape

- Increasingly need to combine multiple data in cross cutting analytic methods
  - E.G. in biomedical: genetics, multiple imaging modalities, proteomics, and clinical elements,
- Must integrate into a formal, standard, clean, consistent, accessible, and linked representation
- “data wrangling” is often the most resource intensive activity in data analysis
  - 50% time overheads reported

Current Tools…

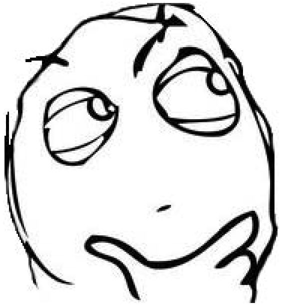
- Shared file systems with data organized in directory hierarchies and with metadata coded into “meaningful” file names
- Idiosyncratic methods used to capture and unify pertinent metadata such as phenotype, experiment details, preparation methods, and quality control flags.
- Manifests in spreadsheets, often the preferred means of describing and tracking data
- One off databases

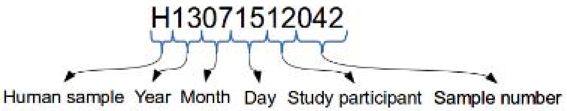
E.G. **A Physical Sciences Network Approach to Understanding Cancer**

[1]Department of Physics, Arizona State University, Tempe, AZ 85287, [2]Center for Biosignatures Discovery Automation, Biodesign Institute, Arizona State University, Tempe, AZ 85287, [3]Clinical Research Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, [4]Beyond Center for Fundamental Concepts in Science, Arizona State University, Tempe, AZ 85287, [5]Department of Medicine, University of Washington, Seattle, WA 98195, [6]Department of Biomedical Engineering, Cornell University, Ithaca, NY 14853, [7]School of Chemical and Biomolecular Engineering, Cornell University, Ithaca, NY 14853, [8]Department of Medicine, Weill Cornell Medical College, New York, NY 10065, [9]School of Mathematics, University of Minnesota Twin Cities, Minneapolis, MN 55455, [10]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, and Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, [11]Cancer Biology and Genetics Program, Department of Neurosurgery, Memorial Sloan-Kettering Cancer Center, New York, NY 10065, [12]Departments of Radiology and Integrated Mathematical Oncology, H. Lee Moffitt Cancer Center and Research Institute, 12902 Magnolia Drive, Tampa, FL 33612, [13]Department of Chemical and Biomolecular Engineering, The Johns Hopkins University, Baltimore, MD 21218, [14]Vascular Program, Institute of Cell Engineering and McKusick-Nathans Institute of Genetic Medicine and the Departments of Pediatrics, Medicine, Oncology, Radiation Oncology, and Biological Chemistry, The Johns Hopkins University School of Medicine, Baltimore, MD 21205, [15]Departments of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, [16]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, [17]Koch Institute for Integrative Cancer Research and Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, MA 02139, [18]Department of Nanomedicine, The Methodist Hospital Research Institute, Houston, TX 77030, [19]David H. Koch Center, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, [20]Department of Experimental Diagnostic Imaging, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, [21]Department of Surgical Oncology, The University of Texas M. D. Anderson Cancer Center, Houston, TX 77030, [22]Biomedical Engineering Department, Northwestern University, Evanston, IL 60208, [23]Division of Hematology/Oncology, Robert H. Lurie Comprehensive Cancer Center, Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, [24]Department of Biochemistry, Molecular Biology, and Cell Biology, Northwestern University, Evanston, IL 60208, [25]Department of Chemistry, Chemistry of Life Processes Institute, Northwestern University, Evanston, IL 60208, [26]Department of Pathology and UCSF Helen Diller Family Comprehensive Cancer Center, University of California at San Francisco, San Francisco, CA, 94143, [27]Department of Physics, Princeton University, Princeton, NJ 08544, [28]Department of Biomedical Engineering, School of Medicine, Oregon Health & Science University, Portland, OR 97239, [29]Department of Cell and Developmental Biology, School of Medicine, Oregon Health & Science University, Portland, OR 97239,

[30]Department of Cell Biology, The Scripps Research Institute, La Jolla, CA 92037, [31]Department of Pathology, Scripps Clinic, La Jolla, CA 92037, [32]Department of Surgery and Center for Bioengineering and Tissue Regeneration, University of California at San Francisco, San Francisco, CA 94143, [33]Department of Physics, Biophysics Graduate Group, University of California - Berkeley, Berkeley, CA 94720, [34]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, [35]Department of Anatomy, Department of Bioengineering and Therapeutic Sciences, Eli and Edythe Broad Center of Regeneration Medicine and Stem Cell Research at UCSF, and Helen Diller Comprehensive Cancer Center, University of California at San Francisco, San Francisco, CA 94143, [36]Center for Applied Molecular Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90211, [37]Department of Radiology, School of Medicine, Stanford University, Stanford, CA 94305, [38]Department of Biology, Department of Computer Science, New York University, New York, NY 10003, [39]Information Sciences Institute, University of Southern California, Los Angeles, 90292, [40]Applied Minds, Inc. Glendale, CA 91201

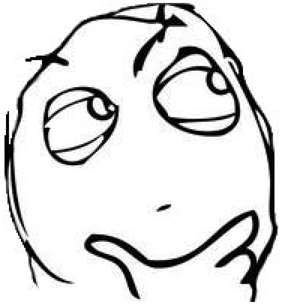
USC

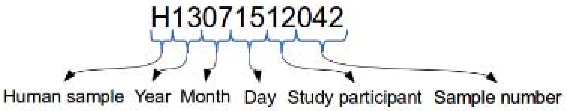




H13071512042

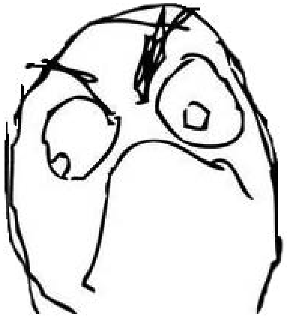
Human sample Year Month Day Study participant Sample number

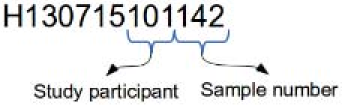




H13071512042

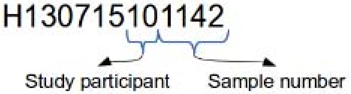
Human sample Year Month Day Study participant Sample number





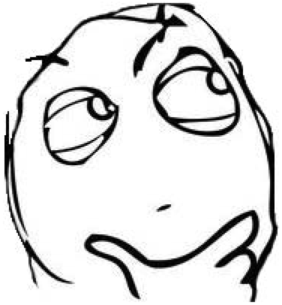
H130715101142

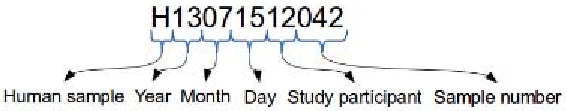
Study participant Sample number



H130715101142

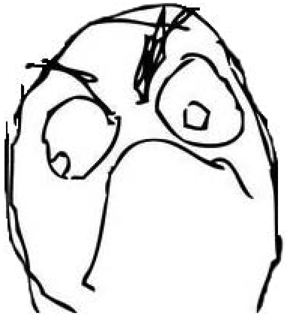
Study participant Sample number

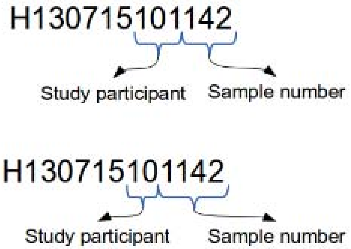




H13071512042

Human sample Year Month Day Study participant Sample number



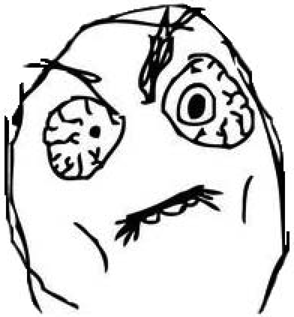


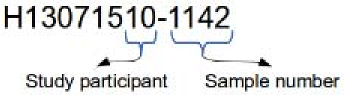
H130715101142

Study participant Sample number

H130715101142

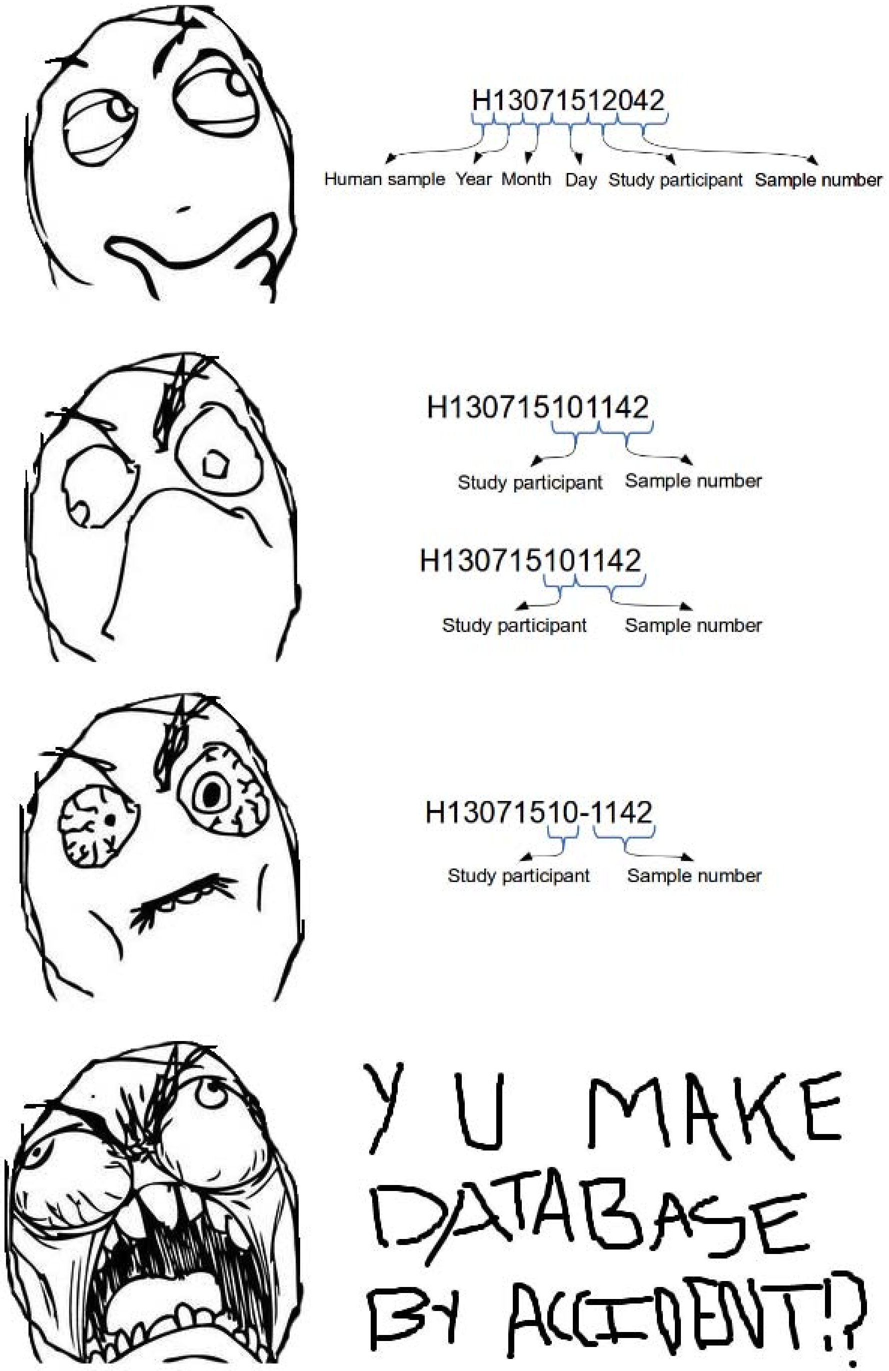
Study participant Sample number





H13071510-1142

Study participant Sample number



H13071512042

Human sample Year Month Day Study participant Sample number

H130715101142

Study participant Sample number

H130715101142

Study participant Sample number

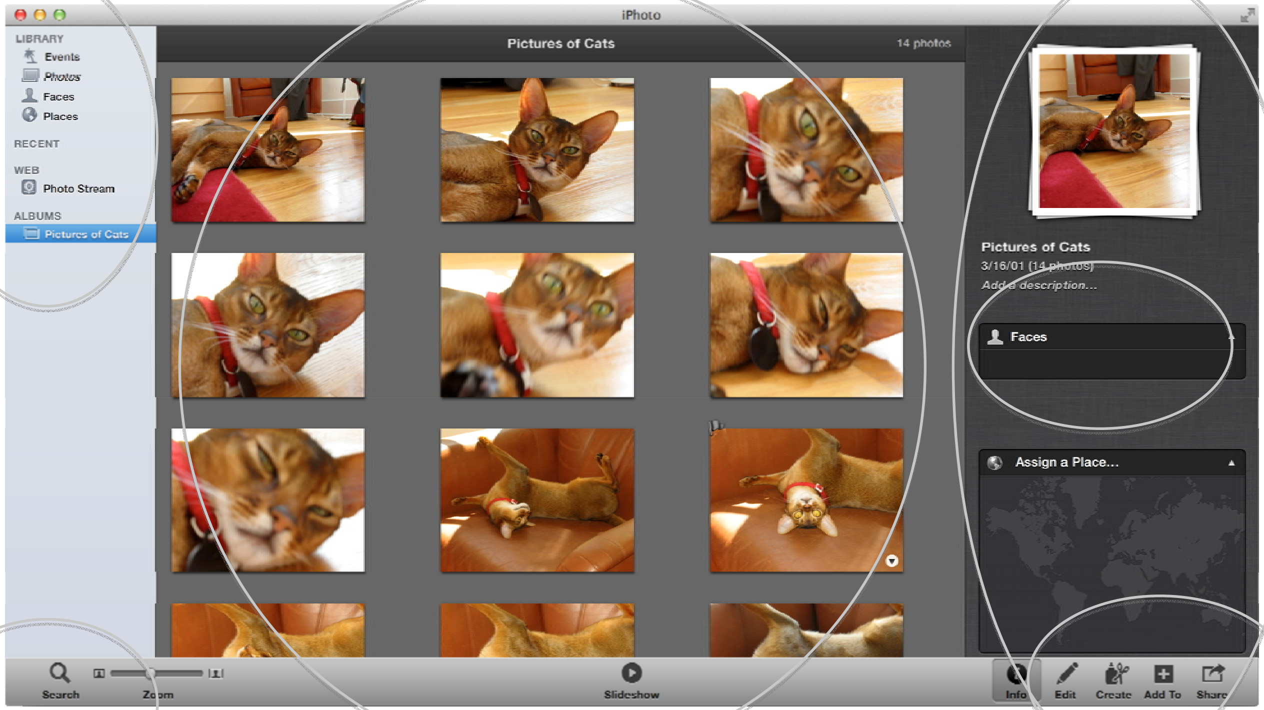
H13071510-1142

Study participant Sample number

Data Management

Why don’t we have tools for managing data sets of cancer and kidneys that are as good as the tools we have for managing data sets of cats and kids?

Flexible data organization



Editable attributes and metadata

Automatic analysis

Edit and share

Full text search

Data browsing

Apple iPhoto

Digital Asset Management

- "management tasks and decisions surrounding the ingestion, annotation, cataloguing, storage, retrieval and distribution of digital assets"
- streamline free-form "creative" processes rather than enforce predefined business processes.
- Many commercial DAM offerings, but not well suited to biomedical data
  - Complex and diverse data types
  - Specialized data ingest requirements
  - Data size (big data)

Biomedical Digital Asset Management

- Digital Asset Management approach applied to to biomedical data.
- “pay-as-you-go” approach provides continuous integration throughout the research and discovery lifecycle.
  - influenced by the concept of Dataspaces,
  - incremental refinement promotes flexible, use-case driven data integration,
- Focus on usability
  - Low barrier to entry
  - Mimic cloud based tools and services investigators are already familiar with

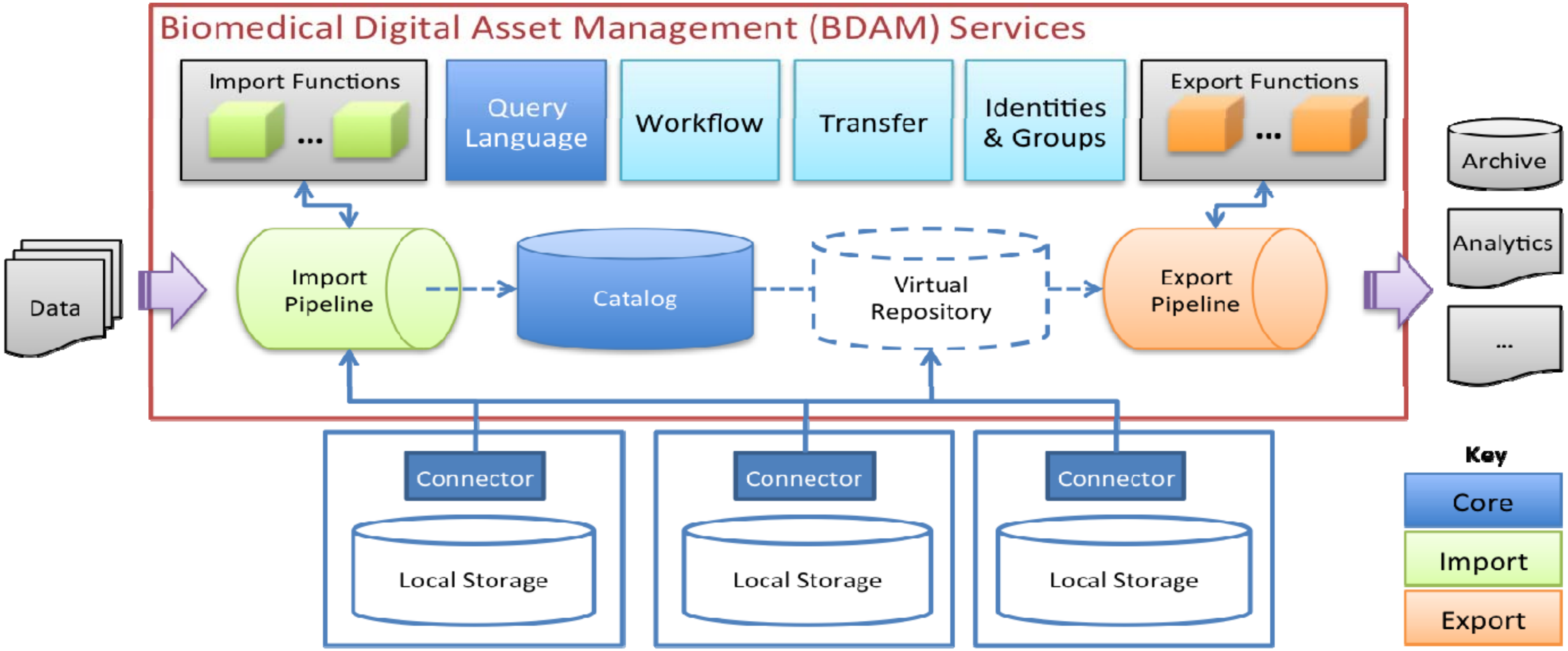
Design Requirements (1)

- Integrate early and often
  - Learn from data spaces and MAD
  - ETL and query mediation approach can require too much upfront schema structure
- BDAM implications
  - *relaxed consistency*: data evolves, rather than entering the system fully formed
  - *incremental refinement*: content and schema, throughout discovery
  - *schema introspection*: interfaces must be able to inspect the catalog’s schema and present interfaces for the user to query and manipulate

Design Requirements (2)

- loose coupling
  - in which core components are operated in a software-as-a-service (SaaS) platform while user data may reside in local storage
- fine-grain access control
  - restrict access to metadata about specific assets attributes of the any asset, or collection of metadata
- multi-tenancy
  - to allow each scientific application to operate at its own pace and with its own content and access policies.

BDAM System Architecture



Data Catalog

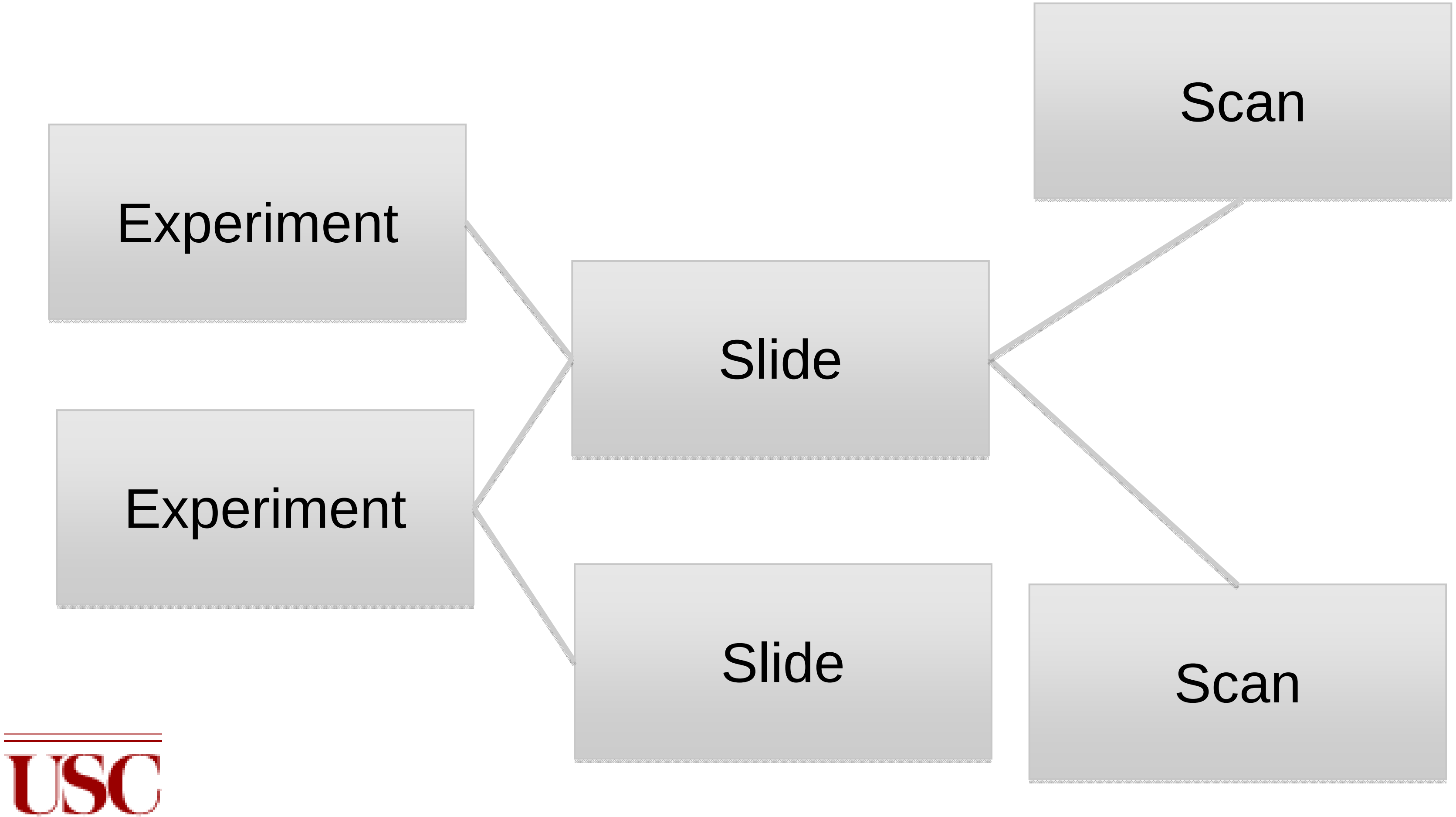
- Records individual data assets along with meaningful metadata
- Can be browsed or searched to find assets matching certain criteria.
- Catalog schema can be queried and modified
- RESTful web services APIs
  - functions for retrieve and edit the metadata schema;
  - create, destroy, update and retrieve whole metadata records;
  - updating or retrieving metadata properties for specific records;
  - queries by metadata criteria and associations to other contextual records.
- Evaluated two catalog implementations
  - Triple based tagging model
  - Table based model
  - Both implemented as web services layer on Postgres

Tagfiler

- Compromise evolving understanding and benefits of strict schema
  - Subject, tag, value triples (like RDF)
  - type of a metadata record is determined by the properties it has (i.e. “duck-typed”)
- Assets identified based on patterns constraining arbitrary sets of attributes.
  - tuned for arbitrary graph-query patterns
- Tuned for sparse data
  - Decompositional Storage Model (DSM) to store triples in property-specific tables and to generate complex joining queries when searching.

Use case example: imaging data from a slide scanner

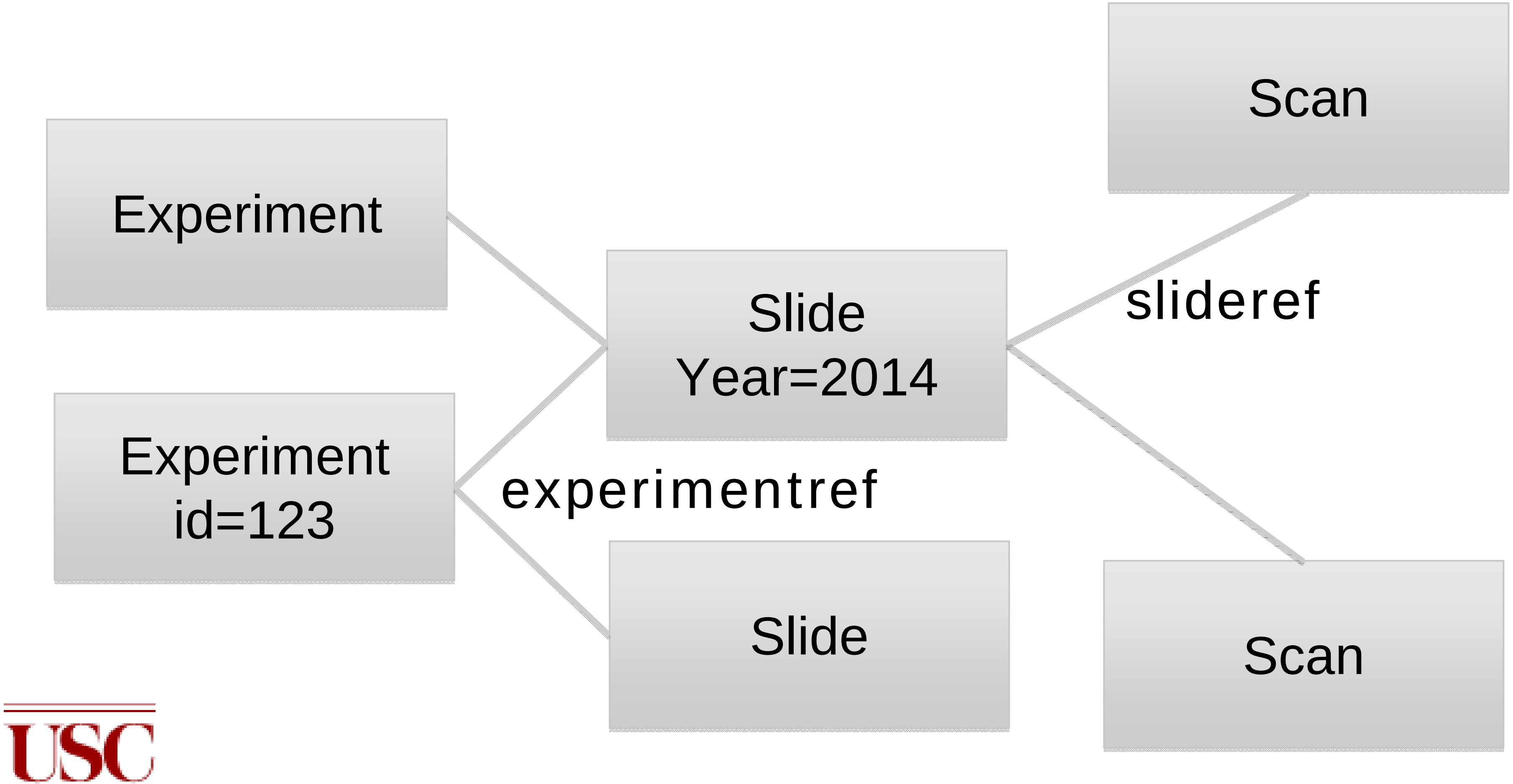
- Managing data from slide scanner:



USC

Use case example: imaging data from a slide scanner

- Managing data from slide scanner:



Tagfiler Query

```
https://bdam.example.org/tagfiler/catalog/42/subject
/slideref=@(/year=2014;experimentref=@(/id=123))
(id;year;fileurl;slideref)
```

Tagfiler Query

https://bdam.example.org/tagfiler/catalog/42/subject/slideref=@(/year=2014;experimentref=@(/id=123))(id;year;fileurl;slideref)

- select catalog 42

Tagfiler Query

https://bdam.example.org/tagfiler/catalog/42/subject/slideref=@(/year=2014;experimentref=@(/id=123))(id;year;fileurl;slideref)

- select catalog 42
- find subjects tagged with “slideref” referencing other subjects that have a year tag with value 2014 and “experimentref” referencing other subjects with an identifier equal to 123.

Tagfiler Query

https://bdam.example.org/tagfiler/catalog/42/subject /slideref=@(/year=2014;experimentref=@(/id=123)) (id;year;fileurl;slideref)

- select catalog 42
- find subjects tagged with “slideref” referencing other subjects that have a year tag with value 2014 and “experimentref” referencing other subjects with an identifier equal to 123.
- return properties (id, year, fileurl, and slideref).

ERMrest

- Table of typed entities with type-specific properties
  - Scientists can think in tables
    - Capture entities and relations
  - supports schema evolution and introspection by being schema agnostic.
    - Similar in phylosiphy to SQLShare (Howe)
- Compact URI naming scheme to traverse linked
  - A URI denoting one set of typed entities can be extended with the name of another linked entity type to denote a set of related entities of that other type.
  - Either URI may also be extended with filter expressions to denote a subset of entities of the same type.

ERMrest Query

- For same model:

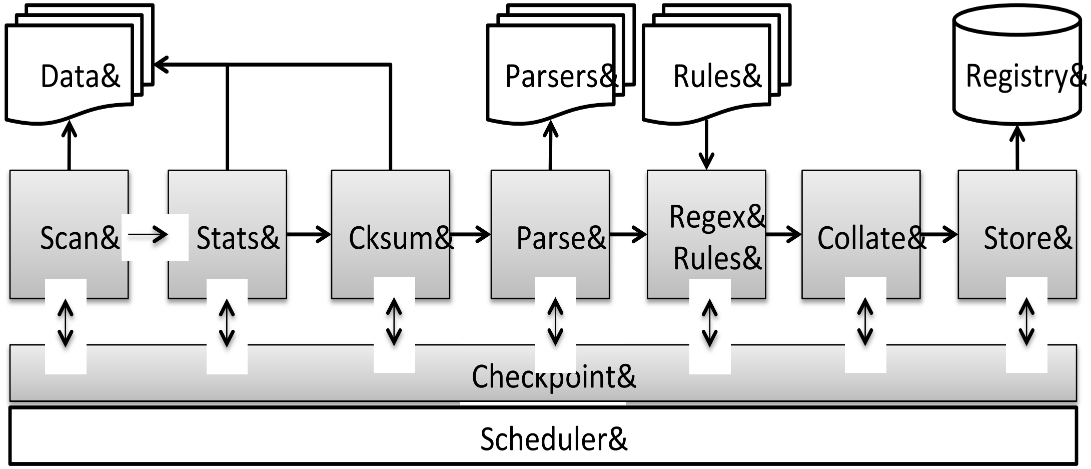
  https://bdam.example.org/ermrest/catalog/42/entity/experiment/id=123/slide/year=2014/scan

- can also follow references in either direction:

  https://bdam.example.org/ermrest/catalog/42/entity/scan/id=456/slide/experiment.

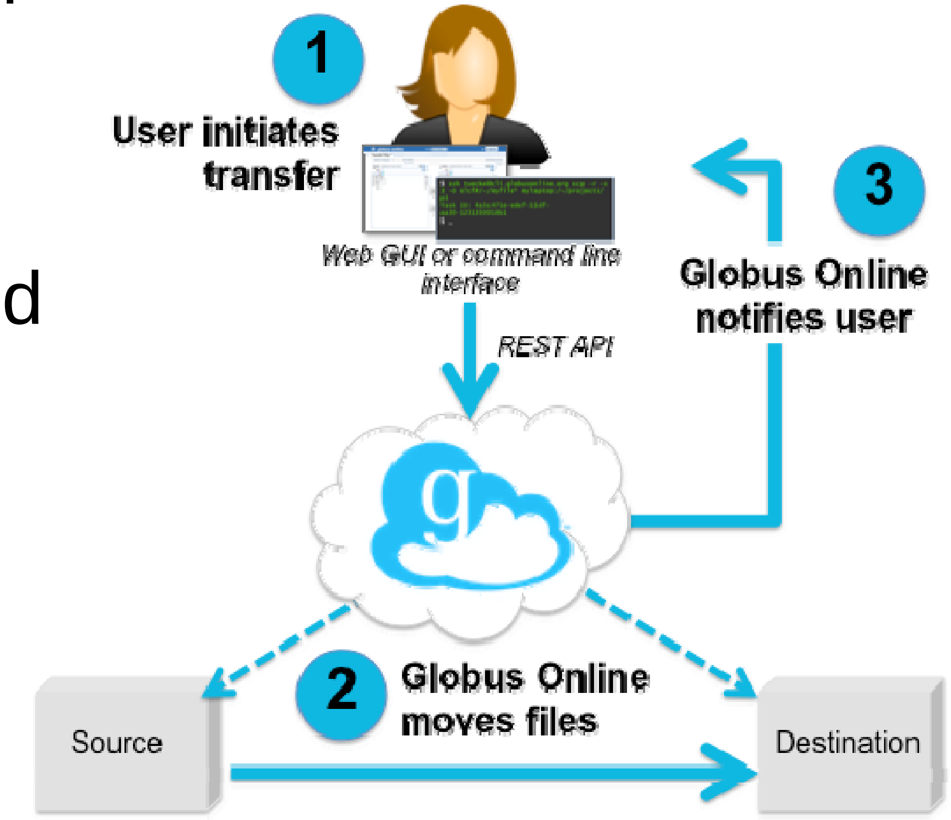
IOBox Ingest pipeline

- File scanning stage to generate a manifest
- Extraction of basic statistics
- Format specific *extraction*
  - HDF5, NetCDF, DICOM, NIfTI, Excel, Olympus SVI, Aperio SVS, and Hamamatsu NDPI, OME-XML and OME-TIFF, and SAM, VCF, and CSV files.
- User-defined rules to *harvest* additional metadata
- Entry into catalog

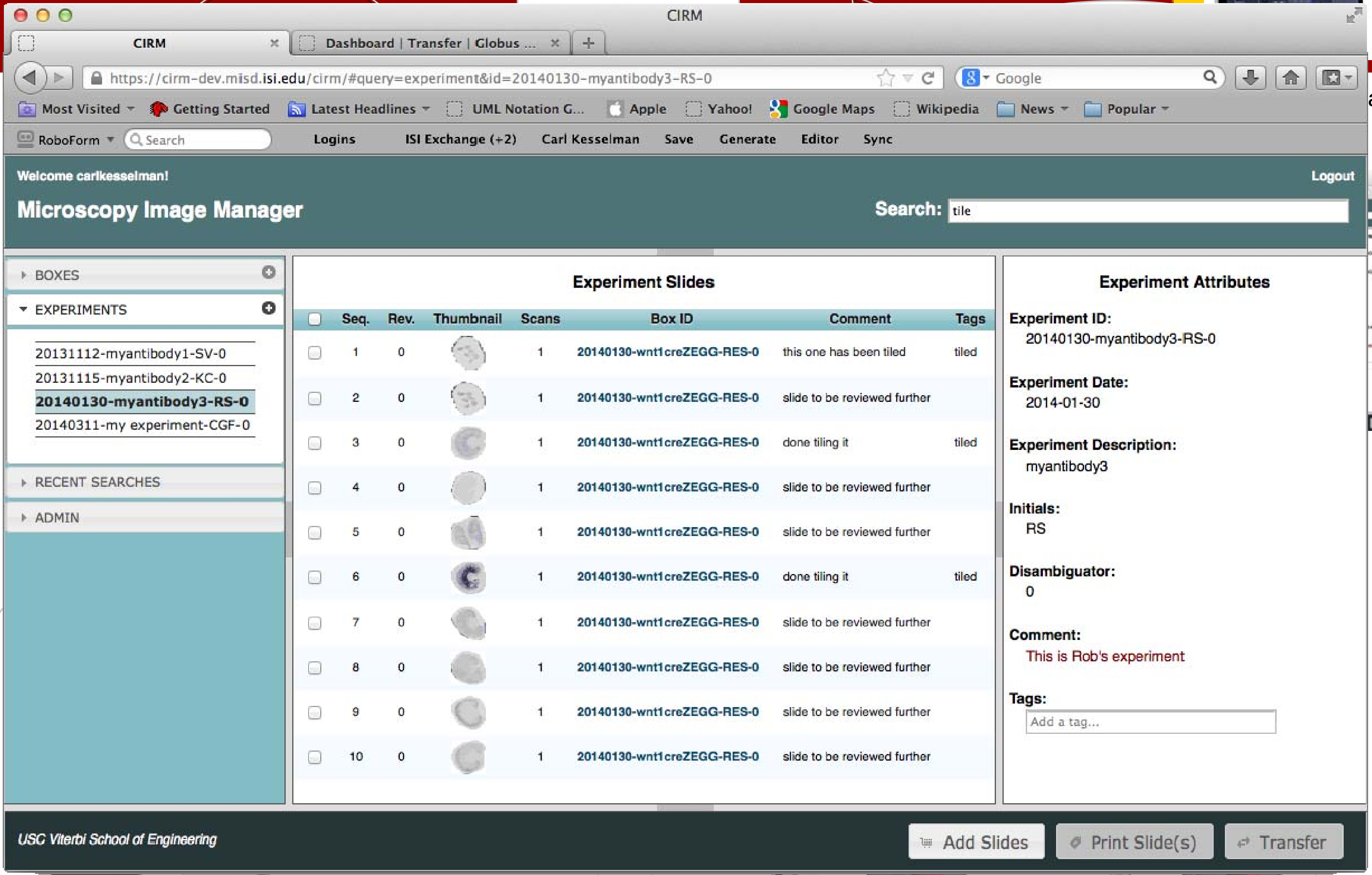


Data Movement

- Pure cloud model
  - Dropbox as a transfer service
- Hybrid cloud model
  - Globus online in cloud combined with local storage services


1
User initiates transfer
Web GUI or command line interface
REST API
3
Globus Online notifies user
2
Globus Online moves files
Source
Destination

ation



Welcome carlkesselman!

Logout

Microscopy Image Manager

Search: tile

- BOXES
- EXPERIMENTS
  - 20131112-myantibody1-SV-0
  - 20131115-myantibody2-KC-0
  - **20140130-myantibody3-RS-0**
  - 20140311-my experiment-CGF-0
- RECENT SEARCHES
- ADMIN

Experiment Slides

□	Seq.	Rev.	Thumbnail	Scans	Box ID	Comment	Tags
□	1	0		1	20140130-wnt1creZEGG-RES-0	this one has been tiled	tiled
□	2	0		1	20140130-wnt1creZEGG-RES-0	slide to be reviewed further	
□	3	0		1	20140130-wnt1creZEGG-RES-0	done tiling it	tiled
□	4	0		1	20140130-wnt1creZEGG-RES-0	slide to be reviewed further	
□	5	0		1	20140130-wnt1creZEGG-RES-0	slide to be reviewed further	
□	6	0		1	20140130-wnt1creZEGG-RES-0	done tiling it	tiled
□	7	0		1	20140130-wnt1creZEGG-RES-0	slide to be reviewed further	
□	8	0		1	20140130-wnt1creZEGG-RES-0	slide to be reviewed further	
□	9	0		1	20140130-wnt1creZEGG-RES-0	slide to be reviewed further	
□	10	0		1	20140130-wnt1creZEGG-RES-0	slide to be reviewed further	

Experiment Attributes

Experiment ID:
20140130-myantibody3-RS-0

Experiment Date:
2014-01-30

Experiment Description:
myantibody3

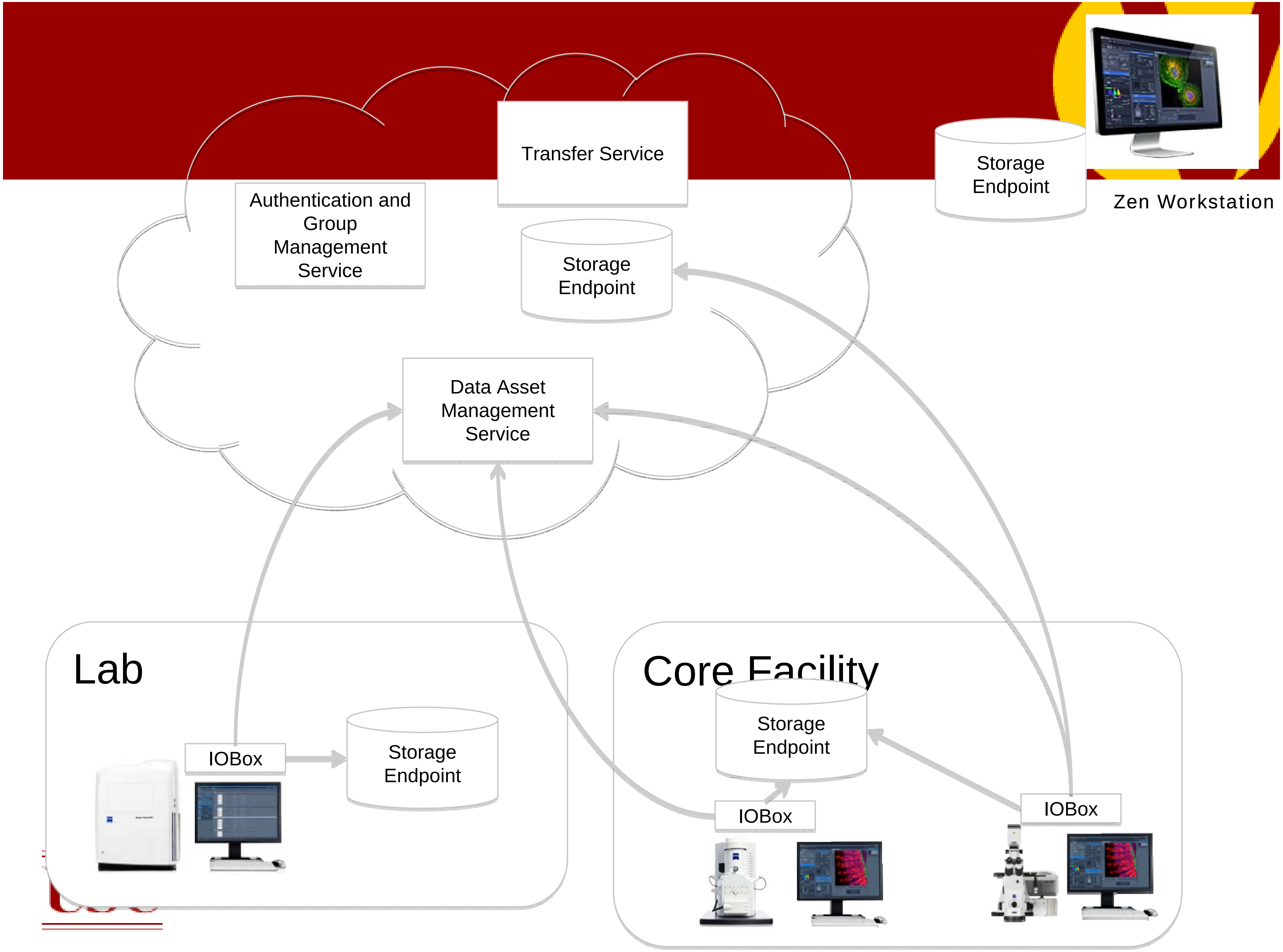
Initials:
RS

Disambiguator:
0

Comment:
This is Rob's experiment

Tags:
Add a tag...

USC Viterbi School of Engineering



Authentication and Group Management Service

Transfer Service

Storage Endpoint

Data Asset Management Service

Storage Endpoint

Zen Workstation

Lab

IOBox

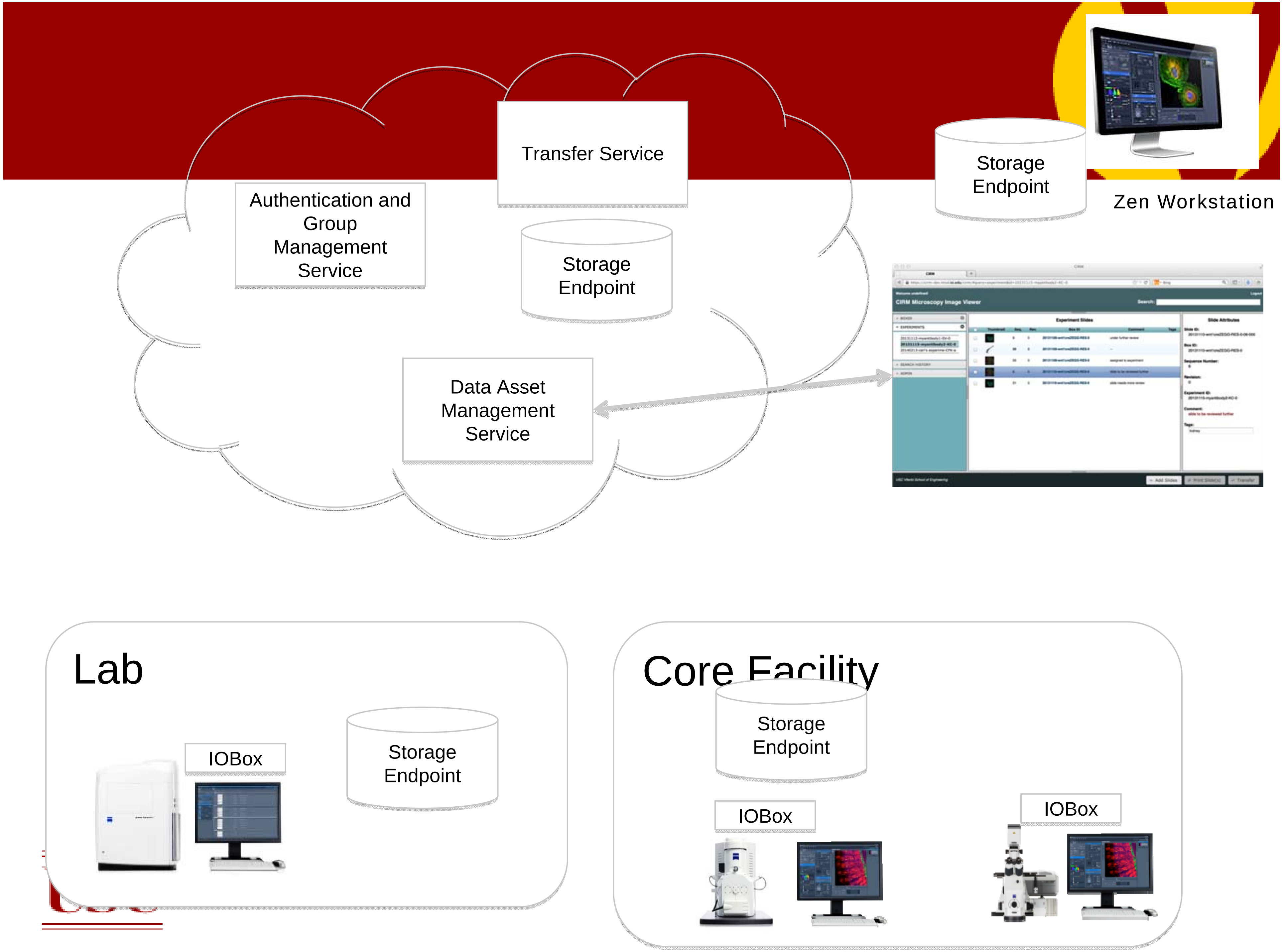
Storage Endpoint

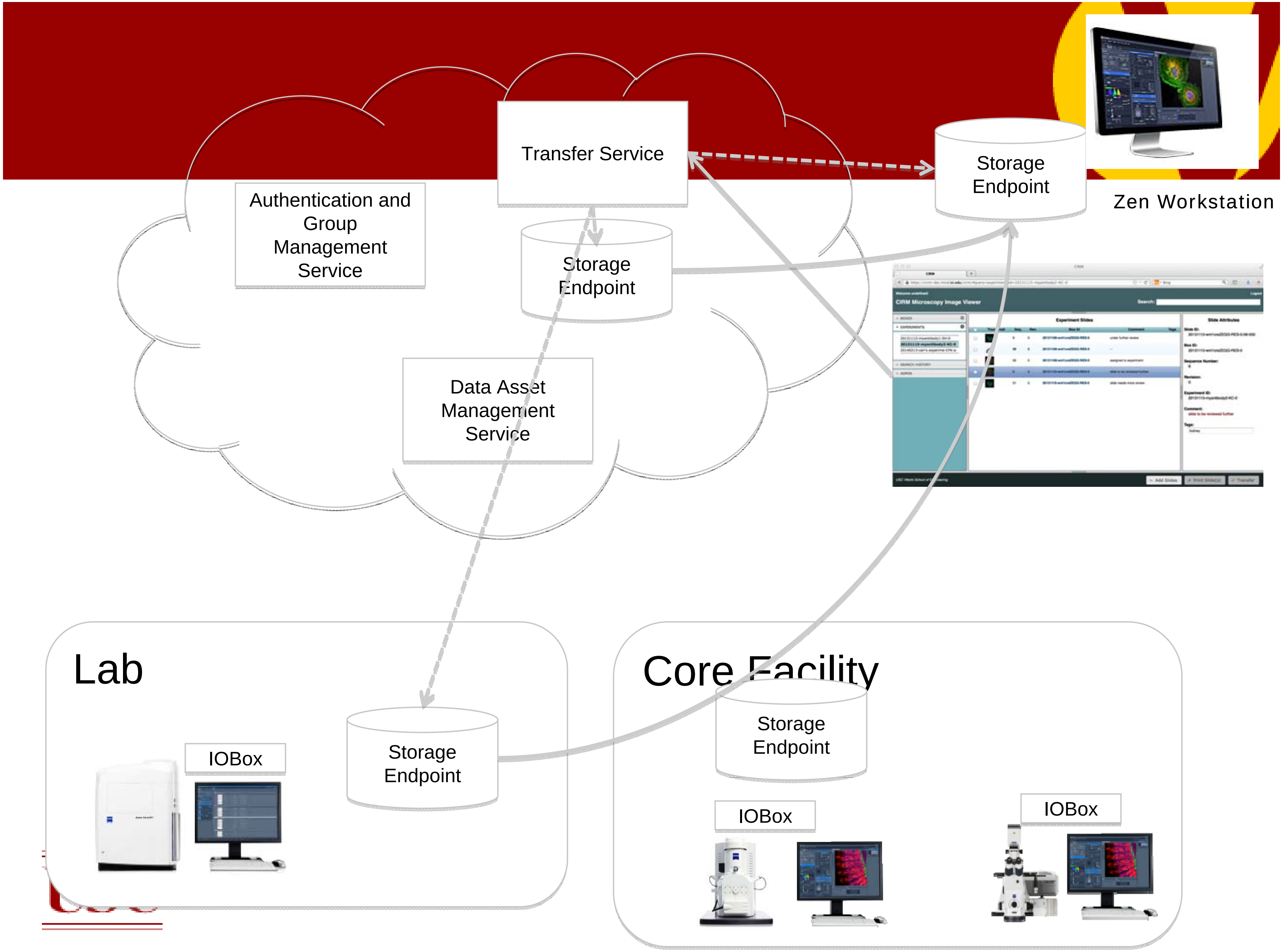
Core Facility

Storage Endpoint

IOBox

IOBox


Transfer Service
Authentication and Group Management Service
Storage Endpoint
Data Asset Management Service
Storage Endpoint
Zen Workstation
CIRM Microscopy Image Viewer
Search:
Experiment Slides
Slide Attributes
Add Slides
Transfer
Lab
IOBox
Storage Endpoint
Core Facility
Storage Endpoint
IOBox
IOBox



Transfer Service

Authentication and Group Management Service

Storage Endpoint

Data Asset Management Service

Storage Endpoint

Zen Workstation

CIRM Microscopy Image Viewer

Lab

IOBox

Storage Endpoint

Core Facility

Storage Endpoint

IOBox

IOBox

Summary

- Continuous evolution and integration is necessary to address the realities of how data is used
- Data-wrangling and data management process with little tooling to support it
- Digital asset management techniques can be effectively applied to biomedical data