globus

DataSet Services in Globus Online

Carl Kesselman

University of Southern California

How Do We Use Data?

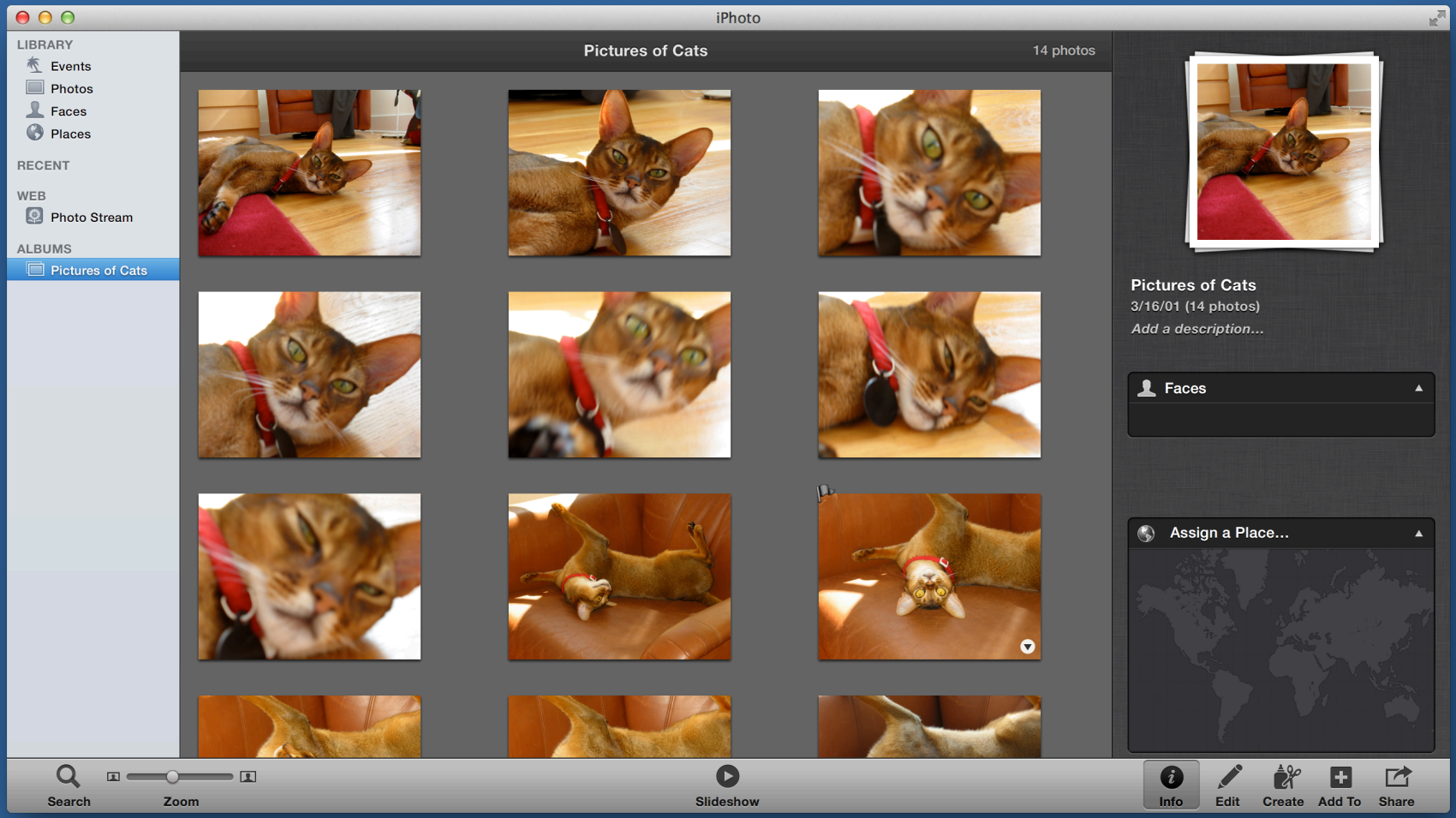
- Researcher's questions often require many pieces of data to answer
- Data elements are spread across:
  - Different locations, e.g. GO endpoints
  - Different file types (excel, txt, CDF, HDF, DICOM)
  - Ad hoc types and locations
- Appropriate grouping of data can vary during investigation
- Data need to be operated on as a unit
  - Shared, processed, copied, …

Organizing Data Is a Challenge

- Structure into directories using file and directory naming conventions
  - Hard to change once its done
  - Globus Transfer can help once we have done this
- Capture data status in README files, …

  Managing complex heterogeneous data using ad hoc methods is too time-consuming, complex and error prone
  - Why can't we manage our data like we manage our pictures and music?

Photos as DataSets



Introducing the DataSet

- Provide a simple method for grouping together files based on use
  - Logical grouping to organize, reorganize, search and describe closer to how the data is used
- Tag the data with characteristics that reflect its content
  - Capture as much existing information as we can
- Tag data to reflect current status in investigation
  - Stage of processing, provenance, validation, ..
- Share data sets for collaboration
- Provide methods that operate on data sets
  - Copy, export, analyze, …

Core Ideas

- A dataset represents a collection of files
  - Datasets are first class citizens
- Datasets are annotated with tags
  - Can be predefined or created on the fly
- Files in GO endpoints can be added to dataset
  - Content remains in the endpoint
  - Basic characteristics are automatically extracted
- Datasets are stored in per-project catalogs
  - Can define conventions for community
- Fine grain access control for sharing and collaboration

Globus DataSet Services

- Builds on current GO services and new hosted services
  - tagging service (tagfiler)
  - Extensible metadata extraction service (InBox)
- Accessible via UI or REST interface

Demonstration in Genomics

- Exome analysis pipelines used by the Onel lab at U. Chicago
  - Compare gene sequence data from leukemia patients against the human genome project sequence data to investigate rare variants
  - More detailed description of science in next talk
- Data from human and animal subjects
- Phenotypic, genotypic and imaging data
  - DICOM, FASTQ, and VCF formatted files
- Create datasets around patients and analyze
  - Create, annotate, share, and export

Demonstration

DataSet Services Wrapup

- Paradigm shift in how researchers work with their data
  - Allow researcher to organize data in ways that make sense for what they are trying to do
- Enhances sharing and collaboration
- Data management for the small labs
- Thanks to: Karl Czajkowski, Rob Schuler, Bryce Allen, Rachana Ananthakrishnan, Steve Tuecke

Questions and Open Discussion