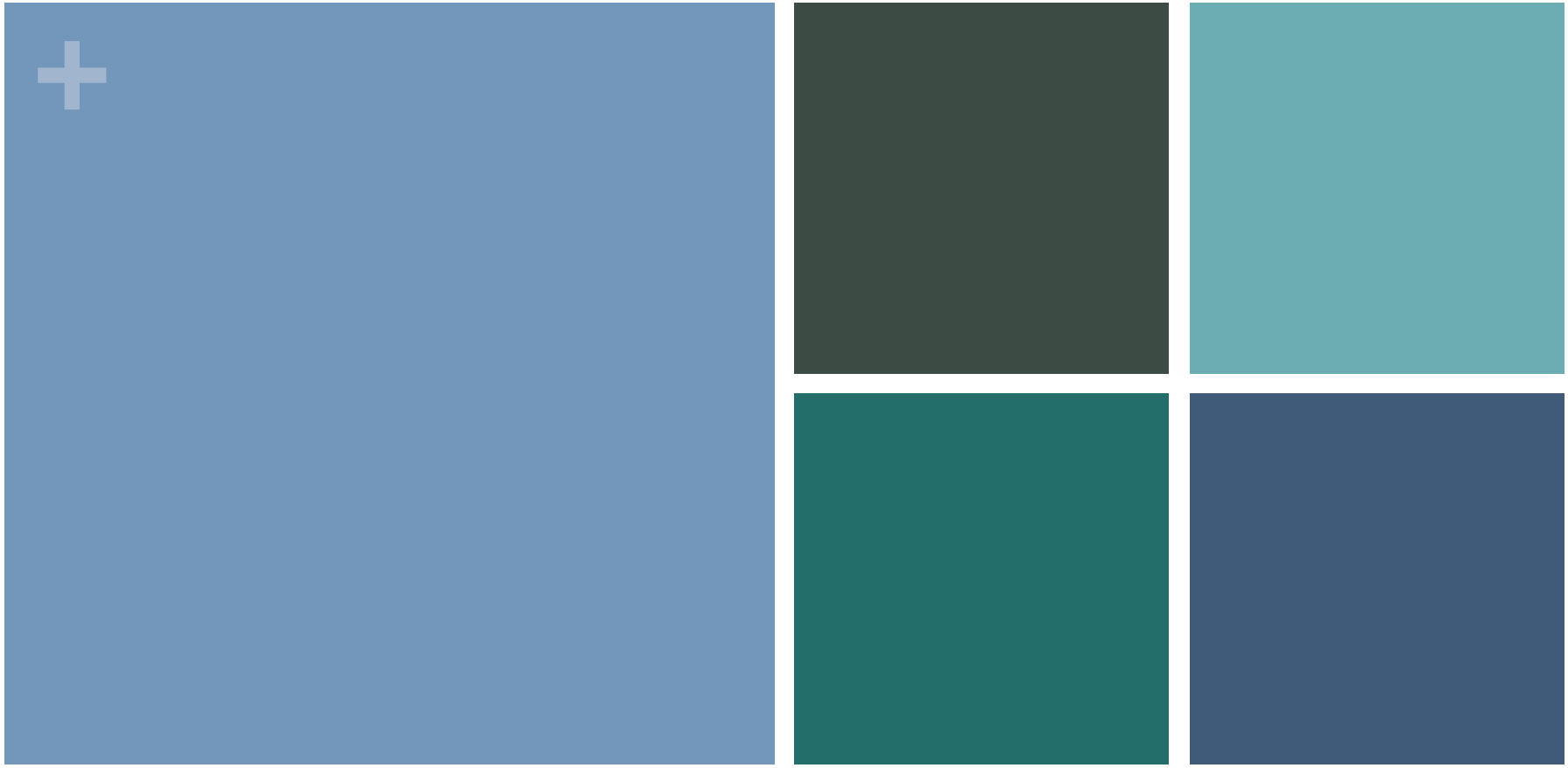


# Archive-It & DuraCloud Integration: Preserving Web Collections

Carissa Smith
DuraCloud Product Manager, DuraSpace

July 24, 2013

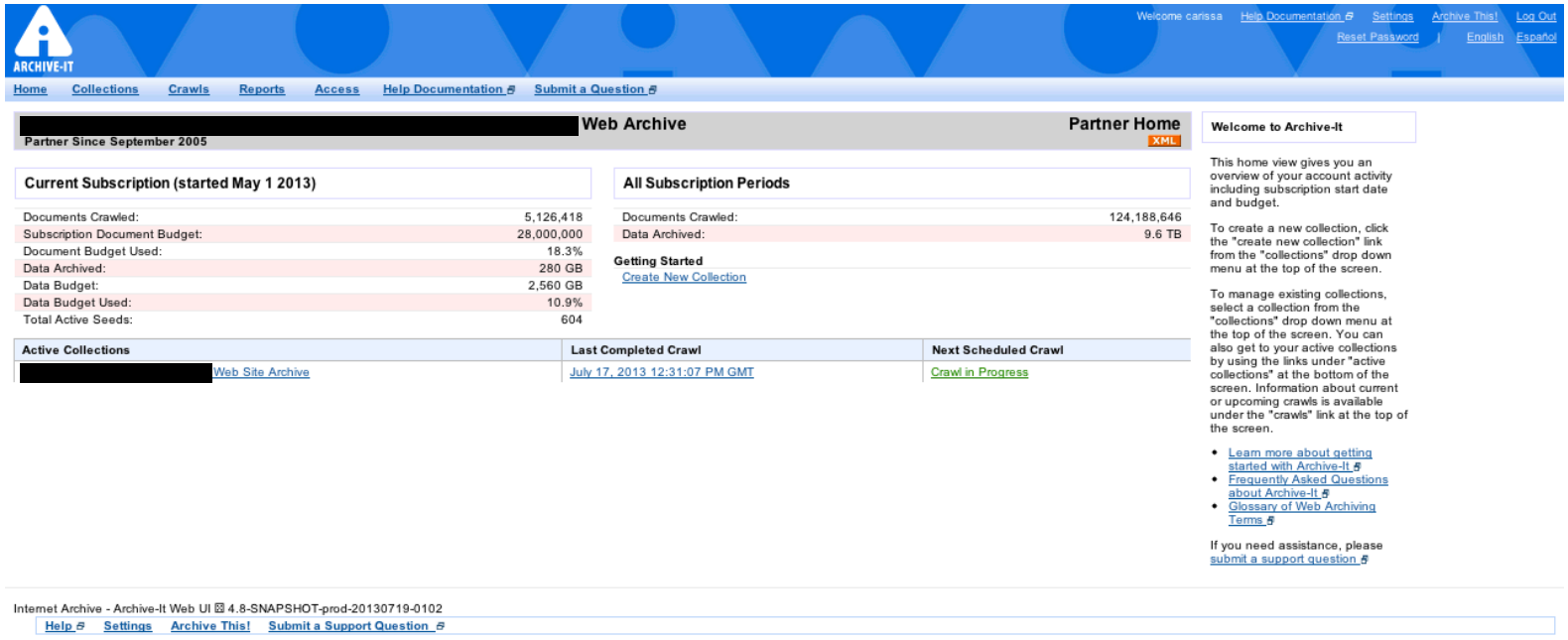
Digital Preservation 2013

# Archive-It

- Description
  - a web archiving service from the Internet Archive created for organizations to capture, build, and manage collections of web content
- History
  - first deployed in February 2006
- Current Status
  - over **275** partner organizations in 46 US states and 16 countries

# Archive-It Interface

## Screenshot

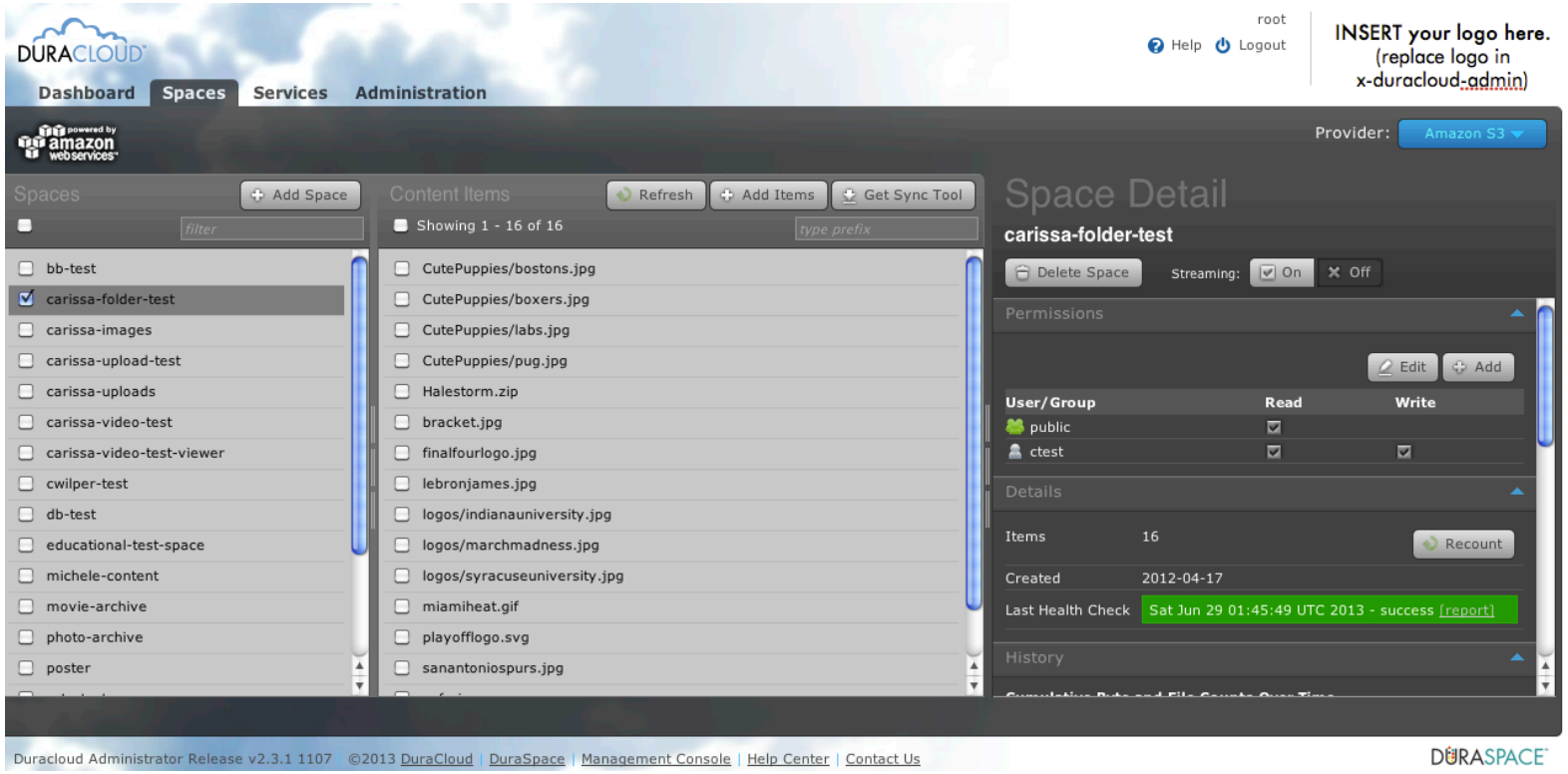


# DuraCloud

- Description
  - a subscription service from DuraSpace that offers organizations a way to easily and cost effectively archive, share, and manage content in the cloud
  - with one click multiple copies are created in the cloud in different locations with several providers all while ensuring the health of the content through DuraCloud's automated content health checking services
- History
  - first deployed in 2009
- Current Status
  - over **25** partner organizations preserving **40+** TBs of content and **8,000,000+** files

# DuraCloud Interface

## Screenshot



# Collaboration

- Opportunity
  - provide Archive-It partner organizations with additional options to back-up their collections of archived content
- Initial Requirements
  - quickly and easily back-up an Archive-It account in DuraCloud
  - view web collections in both web service interfaces
- Goal
  - an additional copy of Archive-It content stored in DuraCloud that can take advantage of DuraCloud's preservation services

# Integration

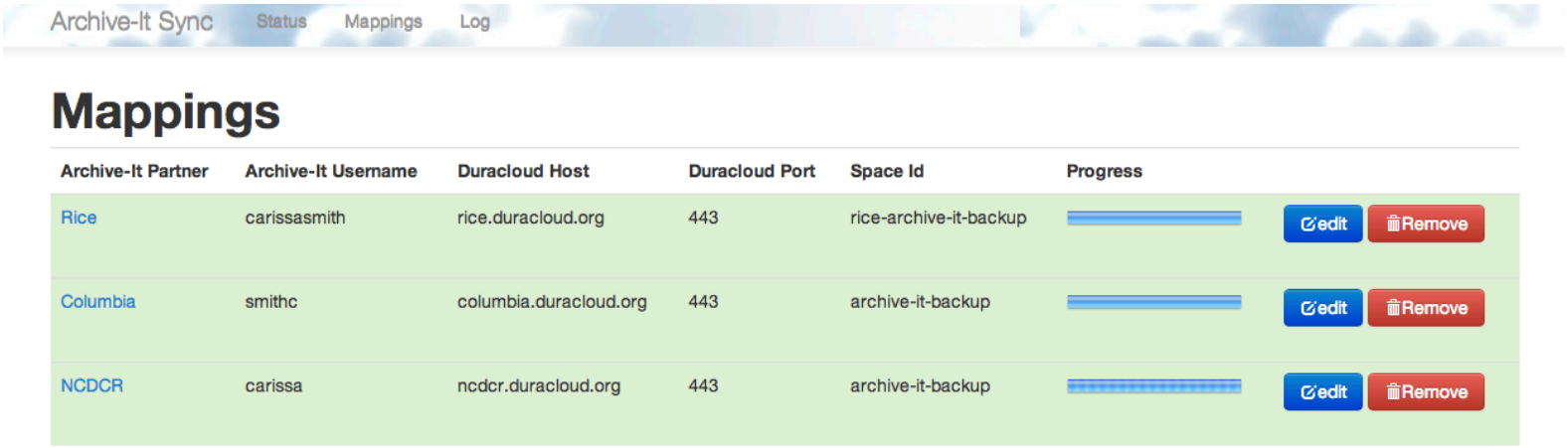
- Summary
  - how about another web application!?!
  - thus was born Archive-It Sync
- How It Works
  - Archive-It Sync performs an initial content ***pull*** from an Archive-It account and stores WARCS in DuraCloud
  - Archive-It Sync then watches Archive-It account for any additions to the collection over time and synchronizes those to DuraCloud
- The Best Part
  - Archive-It partner simply says back-up my content to DuraCloud and it happens
  - No need for partner action!

# Archive-It Sync

- Technical Details
  - java/spring mvc web application
  - calls into Archive-It's LOCKSS API
  - communicates to DuraCloud via DuraCloud REST API using DuraCloud java client libraries
  - continuously monitors Archive-It for new WARCS and queues them for synchronization
  - all Archive-It WARCS are synchronized simultaneously, see sync activity in your DuraCloud account within minutes
  - optionally sync all web archives across collections or constrain by collection and date range
  - dynamically and easily increase/decrease scope of backups to meet your current preservation needs

# Archive-It Sync Interface

## Screenshot



# Timeline

- April 2013: Begin Pilot Project
- May – July 2013: Test Archive-It Sync
- August 2013: Verify content transfers
- August – September 2013: Review Archive-It Sync feature list generated through course of pilot
- ***Fall/Winter 2013: Launch service to public!***

# Partners

- Columbia University
- Rice University
- North Carolina Department of Cultural Resources: State Library and Archives

# Project Status

- Archive-It Sync
  - successfully synchronized **50,000+ WARCS**
  - transfer completed quickly
  - continues to watch partner Archive-It accounts
- Keep track and add to the development of new features here: http://jira.duraspace.org/browse/AITSYNC

# Lessons Learned

- The following capabilities were added to Archive-It Sync
  - "chunk" large WARCS into smaller files to store in DuraCloud (2GB is chunk size)
  - select specific collection(s) within an Archive-It account
  - define specific time range within a collection
- Other features on the immediate roadmap for Archive-It Sync
  - ability to reconfigure account mappings and then choose whether to:
    - keep all files
    - reset account to delete files that do not meet new configuration options

# Future Work

- Additional capabilities for Archive-It Sync
  - run Wayback Machine in DuraCloud to view WARCS
  - collection specific services run over content in DuraCloud
  - ***your idea here!***

# Find Out More

- Archive-It
  - http://www.archive-it.org
- DuraCloud
  - http://www.duracloud.org
- Email me
  - csmith@duraspace.org
- Interested in ***learning more?*** Simply contact me!