RESEARCH CENTER

HATHI TRUST
RESEARCH CENTER

Building Collections and Analyzing Data

Stacy Kowalczyk

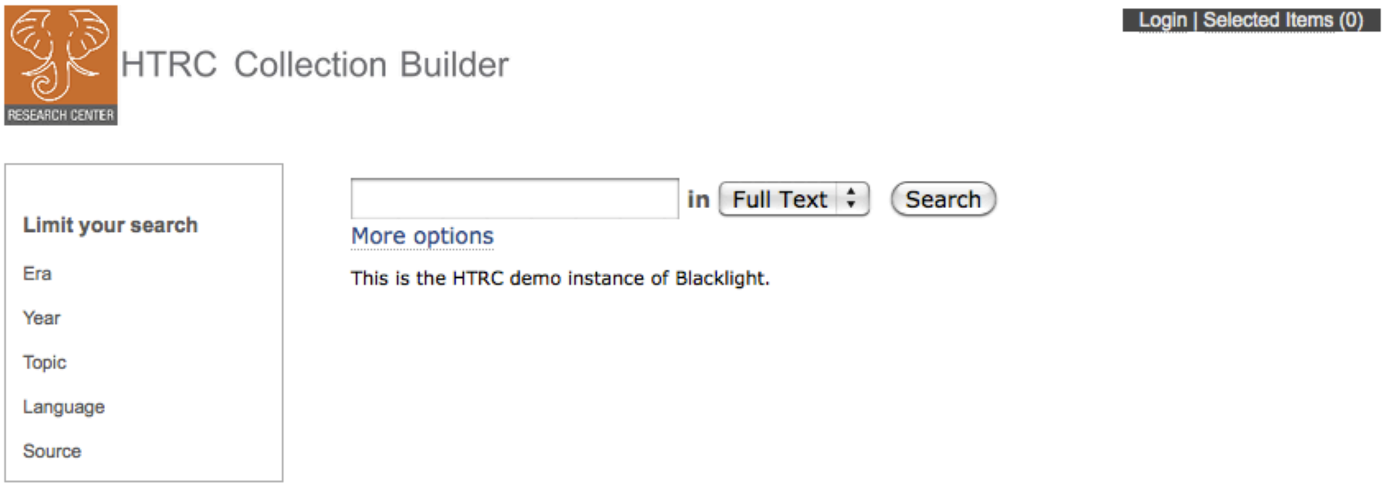
Goals

- To show the current capability of the HTRC
- To get feedback about additional functionality
- To get feedback on the usefulness of the applications
- To find out more about researchers needs for subcollections

Agenda

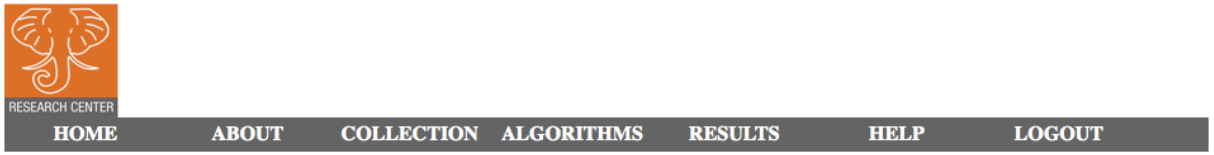
- Introduction
- Application demos
- Hands on collection building
- Hands on data analysis
- Discussion

HTRC Applications



HTRC Collection Builder

This is the HTRC demo instance of Blacklight.



Welcome to the HathiTrust Research Center!

The HathiTrust Research Center (HTRC) provides research access to the public domain text of the HathiTrust Digital Library. The HTRC is a collaborative research center launched jointly by Indiana University and the University of Illinois, along with the HathiTrust Digital Library, to help meet the technical challenges of dealing with massive amounts of digital text that researchers face by developing cutting-edge software tools and cyberinfrastructure to enable advanced computational access to the growing digital record of human knowledge.

The HTRC provides an infrastructure to search, collect, analyze, and visualize the full text of nearly 3 million public domain works and is intended for nonprofit and educational researchers. Click on the login link to begin. LOGIN

HTRC Collection Builder

- Provides a familiar search interface to the entire HTRC collection
- Allows for fielded and free text searching
- Interface for personalized sub-collections of the data for further processing
  - Creating new subcollection
  - Updating existing subcollections
- Based on the Blacklight open source search interface

HTRC Interactive Web Interface

- Provides simple interface to the HTRC infrastructure
  - Manage Collections
  - Submit Algorithms
  - Use Computational resources
  - Manage Jobs
  - View Results

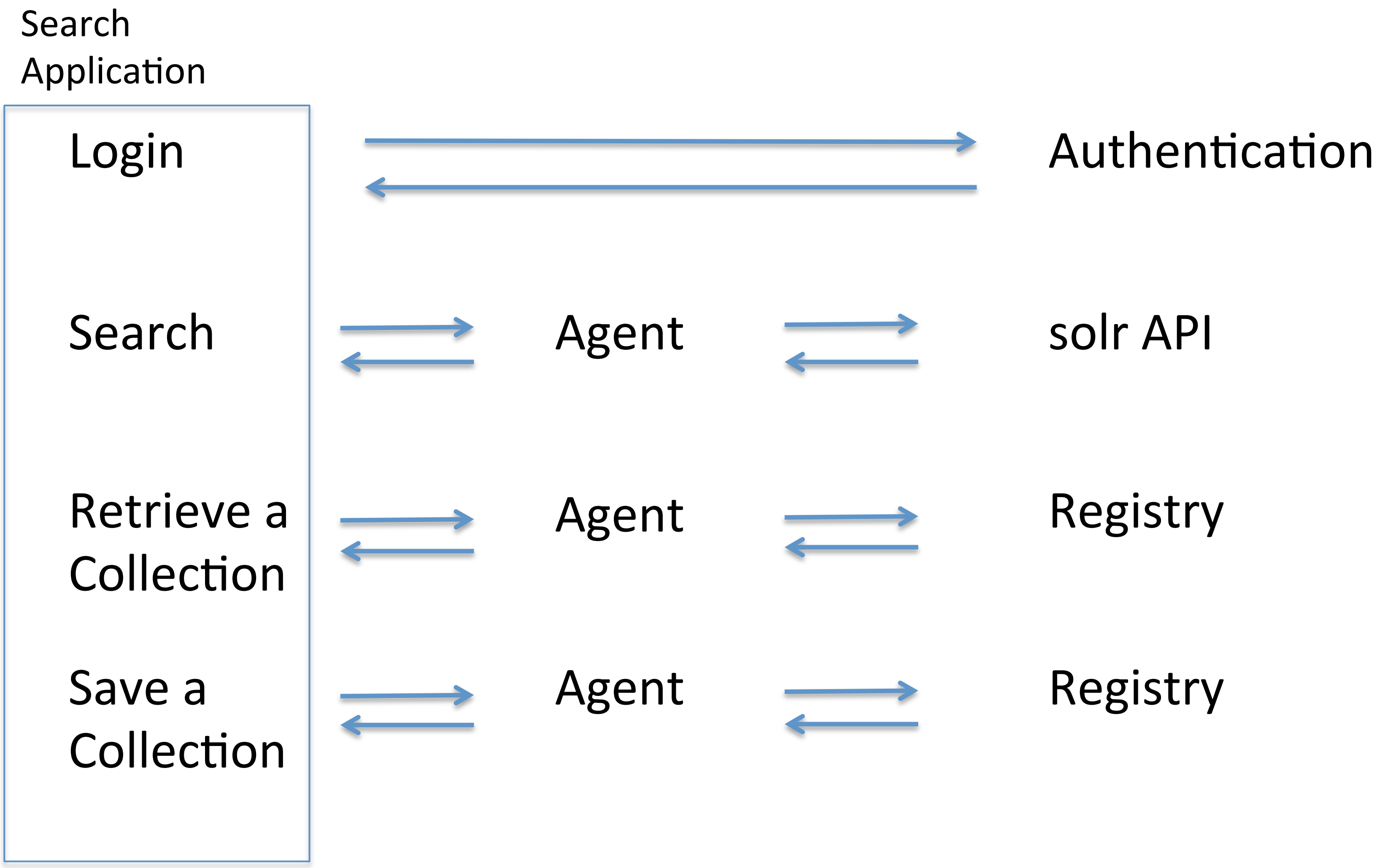
Infrastructure Components

- Authentication
- Agent
- Registry
- APIs
  - Data access
  - Solr access
- Secured computational resources

Collection Builder Demo

RESEARCH CENTER

Collection Builder Process Flow



Search Application

Login

Authentication

Search

Agent

solr API

Retrieve a Collection

Agent

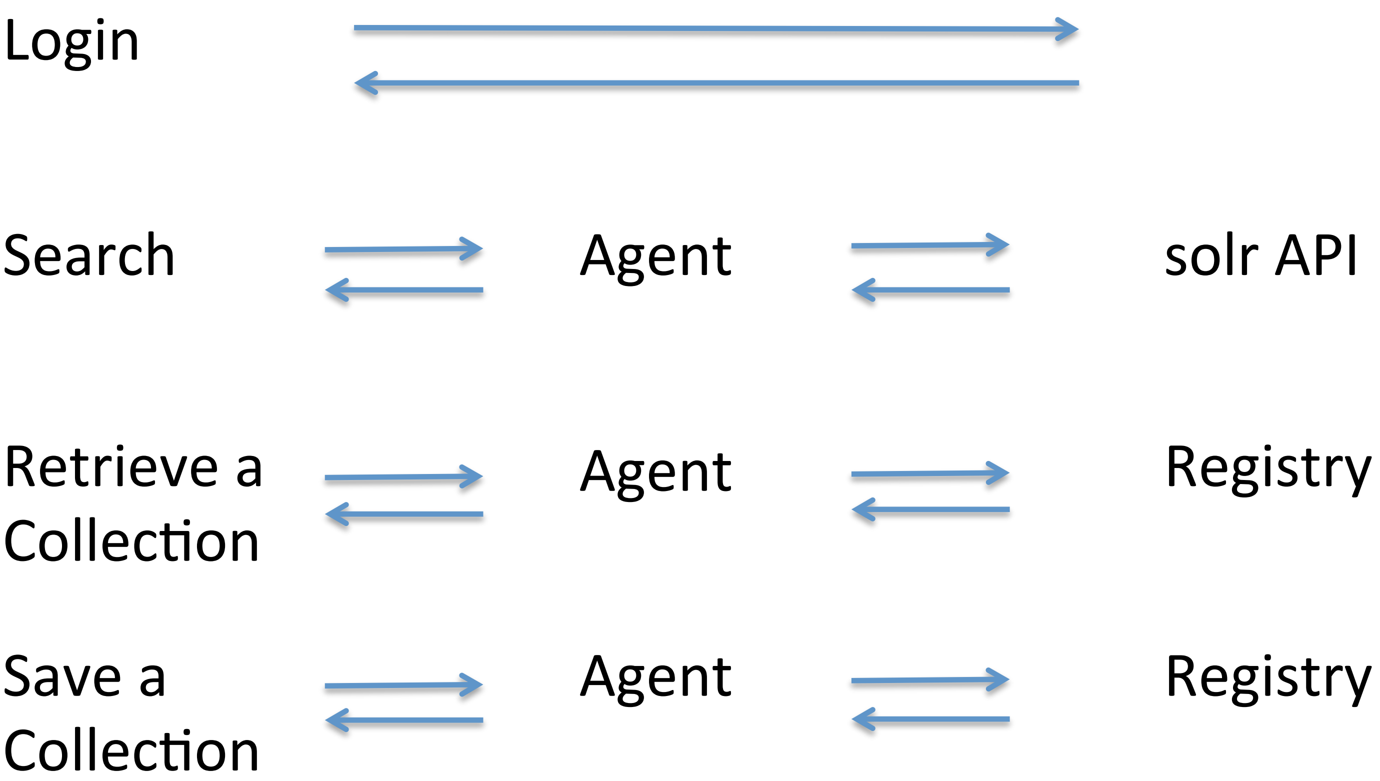
Registry

Save a Collection

Agent

Registry

Collection Builder Process Flow



Login

Search

Agent

solr API

Retrieve a Collection

Agent

Registry

Save a Collection

Agent

Registry

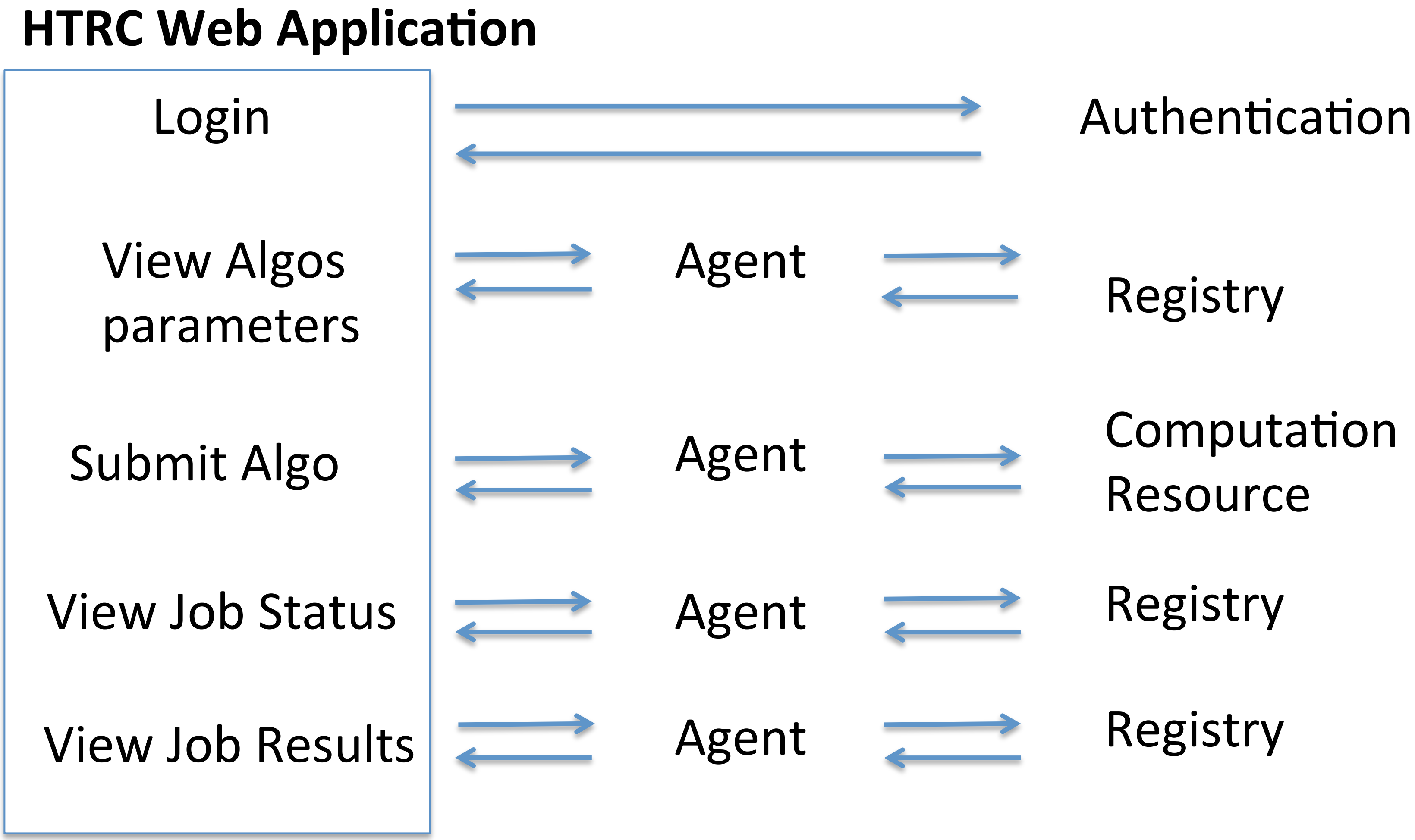
Build a Collection

- http://htrc.mine.nu/blacklight
- Use Firefox, Safari, Chrome browsers
- Login using the ID provided
- Create a collection to be analyzed
  - Search
  - Create a collection
  - Retrieve the collection
  - Add or remove a volume
  - Save revised collection

HTRC Web Interface Demo

RESEARCH CENTER

Algorithm Process Flow



HTRC Web Application

Login

View Algos parameters

Submit Algo

View Job Status

View Job Results

Authentication

Agent

Registry

Agent

Computation Resource

Agent

Registry

Agent

Registry

Run an Algorithm

- http://smoketree.cs.indiana.edu:8999/HTRC-UI-Portal2/LogoutAction
- Sign on with the IDs provided
- View your collections
- Run an algorithm
  - Word count
  - Simple tag cloud
- Results
  - Job management
  - View Results

Feedback

- What worked for you?
- What did not work?
- Where there interactions that were awkward?
- What additional functions would you like to see?
- What types of algorithms do you use in your research?
- What environments do these algorithms require?
- What is the approximate size of the collections you use in your work?
- Were you able to find useful materials for your research? Would subsetting the collection apriori help you find materials?