Two Projects, One Challenge:
Common research data issues in MIREX and HTRC

Presented by

J. Stephen Downie

University of Illinois at Urbana-Champaign

Acknowledgements

- Most of today’s slides are directly drawn (aka copied) from the slides presented at the HTRC UnCamp events in Bloomington, Indiana and Champaign, Illinois.
- Today’s talk summarizes four days of excellent presentations and demonstrations!
- We thank the HTRC team and the UnCamp presenters for the use of their very informative slides.

Agenda

- Introduce MIREX
- Introducing the HathiTrust
- Non-Consumptive Research
  - MIREX and NEMA
  - HTRC Software Architecture
- Workset Creation for Scholarly Analysis
- Next Steps

In the Beginning…

- Work began on MIREX in Bloomington, Indiana in 2001
  - The "Indiana Manifesto"
- 2001-2003
  - Fact-finding meetings, planning meetings, and workshops funded by Mellon and NSF
  - Large-scale funding from NSF and Mellon 2003
- *Audio Description Contest* at ISMIR 2004
- MIREX first run at ISMIR 2005 in London

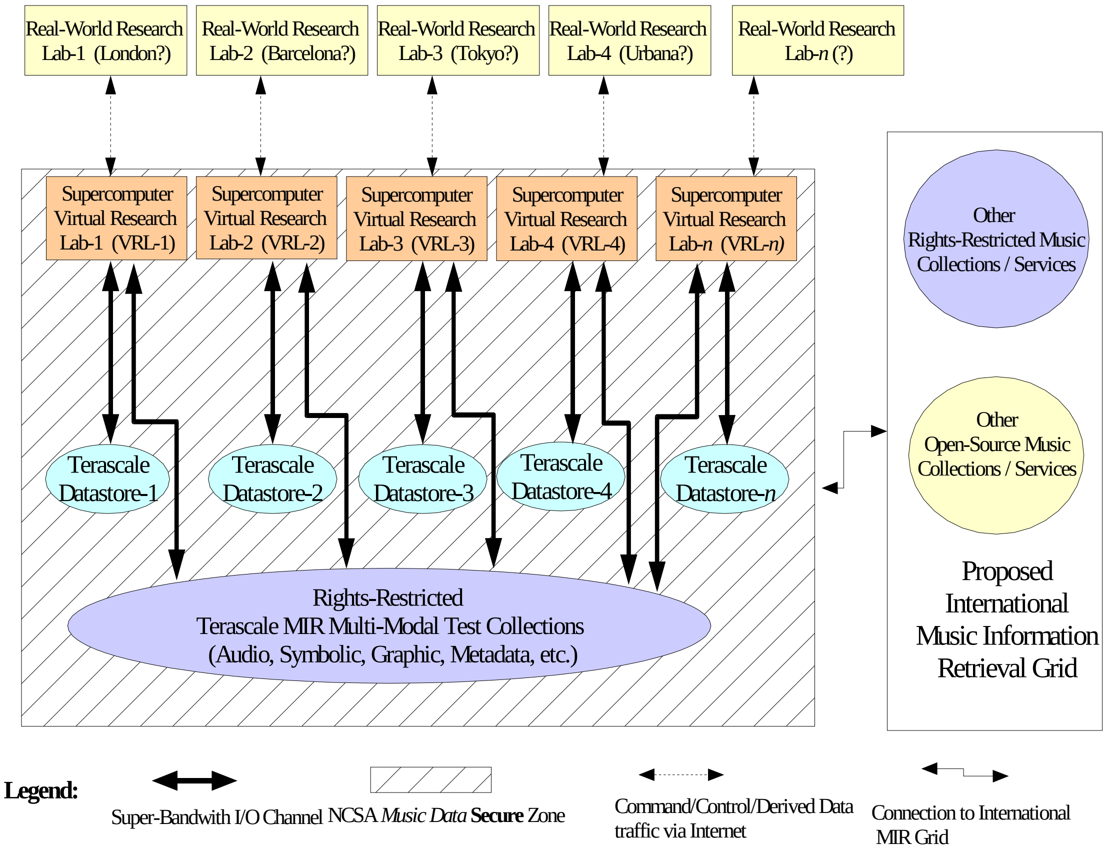
MIREX Model

- Based upon the TREC approach:
  - Standardized queries/tasks
  - Standardized collections
  - Standardized evaluations of results
- Not like TREC with regard to distributing data collections to participants
  - Music copyright issues, ground-truth issues, overfitting issues

IMIRSEL: First Principles

1. Security for the music materials
2. Accessibility for international, domestic and internal researchers
3. Sufficient computing and storage infrastructure for the computationally- and data-intensive MIR/MDL techniques examined

IMIRSEL Model



Real-World Research Lab-1 (London?)

Real-World Research Lab-2 (Barcelona?)

Real-World Research Lab-3 (Tokyo?)

Real-World Research Lab-4 (Urbana?)

Real-World Research Lab-*n* (?)

Supercomputer Virtual Research Lab-1 (VRL-1)

Supercomputer Virtual Research Lab-2 (VRL-2)

Supercomputer Virtual Research Lab-3 (VRL-3)

Supercomputer Virtual Research Lab-4 (VRL-4)

Supercomputer Virtual Research Lab-*n* (VRL-*n)*

Terascale Datastore-1

Terascale Datastore-2

Terascale Datastore-3

Terascale Datastore-4

Terascale Datastore-*n*

Rights-Restricted
Terascale MIR Multi-Modal Test Collections
(Audio, Symbolic, Graphic, Metadata, etc.)

Other
Rights-Restricted Music
Collections / Services

Other
Open-Source Music
Collections / Services

Proposed
International
Music Information
Retrieval Grid

Legend:

Super-Bandwith I/O Channel

▨ NCSA *Music Data* **Secure** Zone

Command/Control/Derived Data traffic via Internet

Connection to International MIR Grid

MIREX Overview

- Began as MIREX in 2005
- Tasks defined by community debate
- Data sets collected and/or donated
- Participants submit code to IMIRSEL
- Code rarely works first try ☺
- Huge labour consumption getting programmes to work
- Meet at ISMIR to discuss results

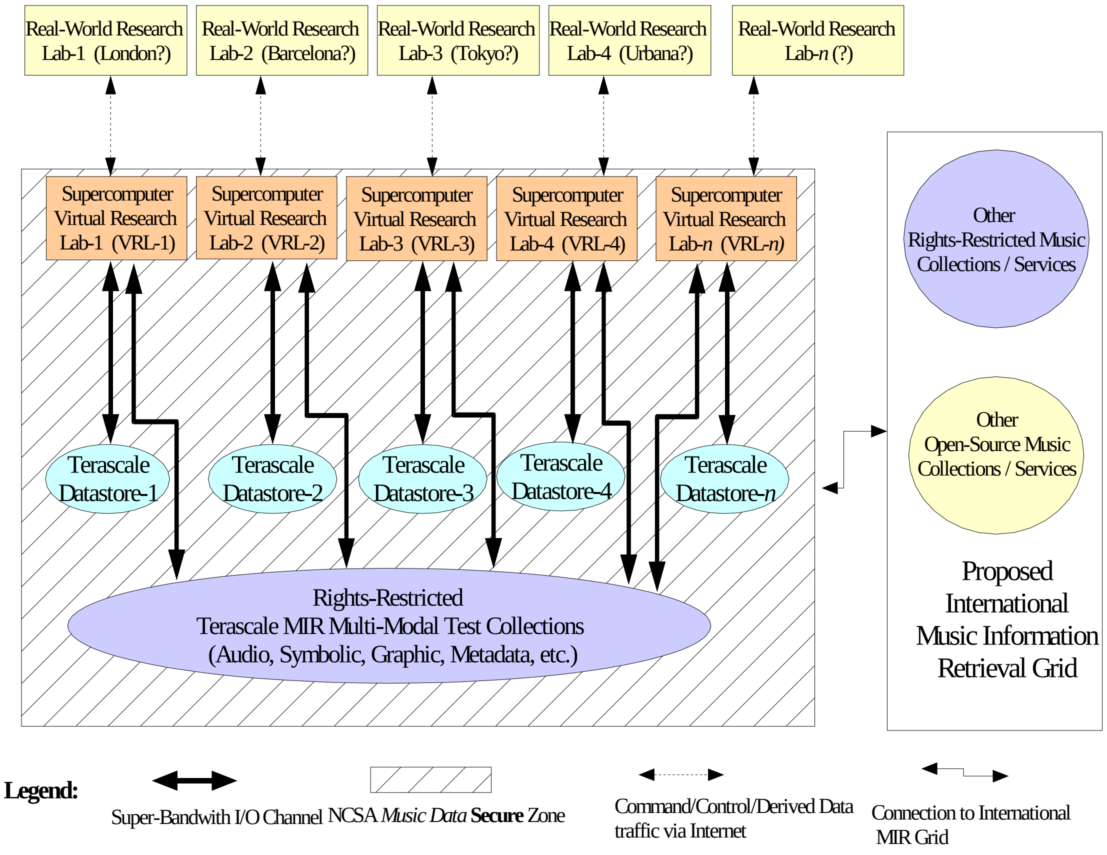
MIREX Model

- Based upon the TREC approach:
  - Standardized queries/tasks
  - Standardized collections
  - Standardized evaluations of results
- Not like TREC with regard to distributing data collections to participants
  - Music copyright issues, ground-truth issues, overfitting issues

IMIRSEL: First Principles

1. Security for the music materials
2. Accessibility for international, domestic and internal researchers
3. Sufficient computing and storage infrastructure for the computationally- and data-intensive MIR/MDL techniques examined

IMIRSEL Model



Real-World Research Lab-1 (London?)

Real-World Research Lab-2 (Barcelona?)

Real-World Research Lab-3 (Tokyo?)

Real-World Research Lab-4 (Urbana?)

Real-World Research Lab-*n* (?)

Supercomputer Virtual Research Lab-1 (VRL-1)

Supercomputer Virtual Research Lab-2 (VRL-2)

Supercomputer Virtual Research Lab-3 (VRL-3)

Supercomputer Virtual Research Lab-4 (VRL-4)

Supercomputer Virtual Research Lab-*n* (VRL-*n)*

Terascale Datastore-1

Terascale Datastore-2

Terascale Datastore-3

Terascale Datastore-4

Terascale Datastore-*n*

Rights-Restricted
Terascale MIR Multi-Modal Test Collections
(Audio, Symbolic, Graphic, Metadata, etc.)

Other Rights-Restricted Music Collections / Services

Other Open-Source Music Collections / Services

Proposed International Music Information Retrieval Grid

Legend:

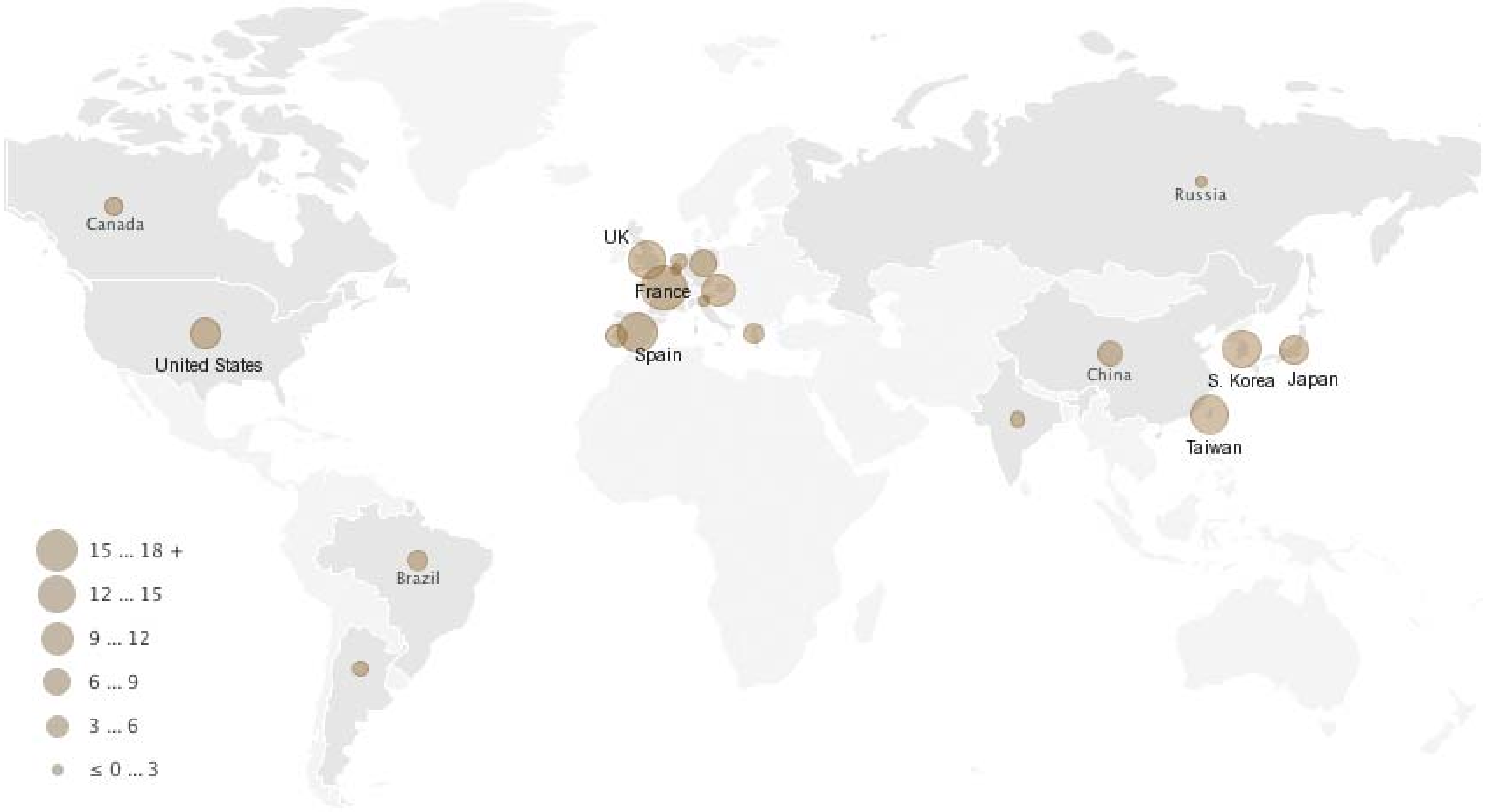
Super-Bandwith I/O Channel

NCSA *Music Data* **Secure** Zone

Command/Control/Derived Data traffic via Internet

Connection to International MIR Grid

MIREX Participation



MIREX 2012: 109 participants from 20 countries

MIREX Participation

	2005	2006	2007	2008	2009	2010	2011	2012	2013
Datasets	10	13	12	18	26	31	32	35	37
Individuals	82	50	73	84	138	152	156	109	116
Runs	86	92	122	169	289	337	312	302	328

MIREX 2013

- 116 researchers
- More than 29 countries
- 37 datasets
- 24 tasks
- 328 completed runs

MIREX 2013 TASKS	
Audio Artist Identification	Audio Onset Detection
Audio Beat Tracking	Audio Tag Classification
Audio Chord Detection	Audio Tempo Extraction
Audio Classical Composer ID	Multiple F0 Estimation
Audio Cover Song Identification	Multiple F0 Note Detection
Audio Drum Detection	Query-by-Singing/Humming
Audio Genre Classification	Query-by-Tapping
Audio Key Finding	Score Following
Audio Melody Extraction	Structural Segmentation
Audio Mood Classification	Symbolic Genre Classification
Audio Music Similarity	Symbolic Key Finding
Discovery of Repeated Themes & Sections	Symbolic Melodic Similarity

MIREX Participation

	2005	2006	2007	2008	2009	2010	2011	2012	2013
Datasets	10	13	12	18	26	31	32	35	37
Individuals	82	50	73	84	138	152	156	109	116
Runs	86	92	122	169	289	337	312	302	328

Total Runs: 2037!

Introducing the HathiTrust

Partnership

Allegheny College
Arizona State University
Baylor University
Boston College
Boston University
Brown University
California Digital Library
Colby College
Columbia University
Cornell University
Dartmouth College
Duke University
Emory University
Florida State University
Getty Research Institute
Harvard University Library
Indiana University
Johns Hopkins University
Lafayette College
Library of Congress
Massachusetts Institute of Technology
McGill University`
Michigan State University
New York Public Library
New York University
North Carolina Central University
North Carolina State University
Northwestern University
The Ohio State University
The Pennsylvania State University
Princeton University
Purdue University
Stanford University
Temple University
Texas A&M University
Tufts University
Universidad Complutense de Madrid
University of Alberta
University of British Columbia
University of Arizona
University of Calgary
University of California
- Berkeley
- Davis
- Irvine
- Los Angeles
- Merced
- Riverside
- San Diego
- San Francisco
- Santa Barbara
- Santa Cruz

The University of Chicago
University of Connecticut
University of Delaware
University of Florida
University of Houston
University of Illinois
University of Illinois at Chicago
The University of Iowa
University of Maryland
University of Massachusetts
University of Miami
University of Michigan
University of Minnesota
University of Missouri
University of Nebraska-Lincoln
The University of North Carolina at Chapel Hill
University of Notre Dame
University of Oklahoma
University of Pennsylvania
University of Pittsburgh
University of Queensland
University of Tennessee,Knoxvile
University of Utah
University of Virginia
University of Washington
University of Wisconsin-Madison
Utah State University
Wake Forest University
Washington University
Yale University Library

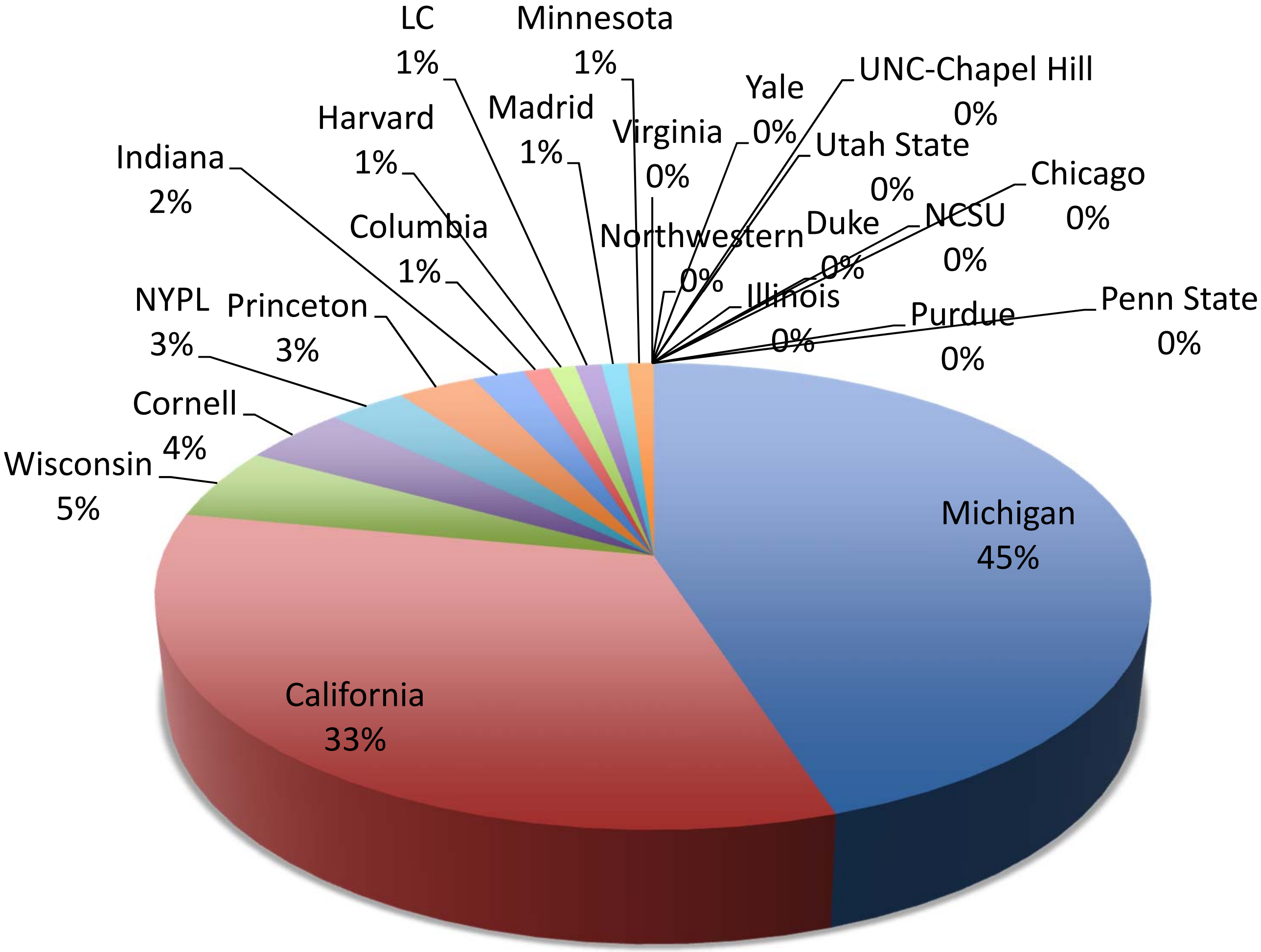
Mission

To contribute to the common good by collecting, organizing, preserving, communicating, and sharing the record of human knowledge

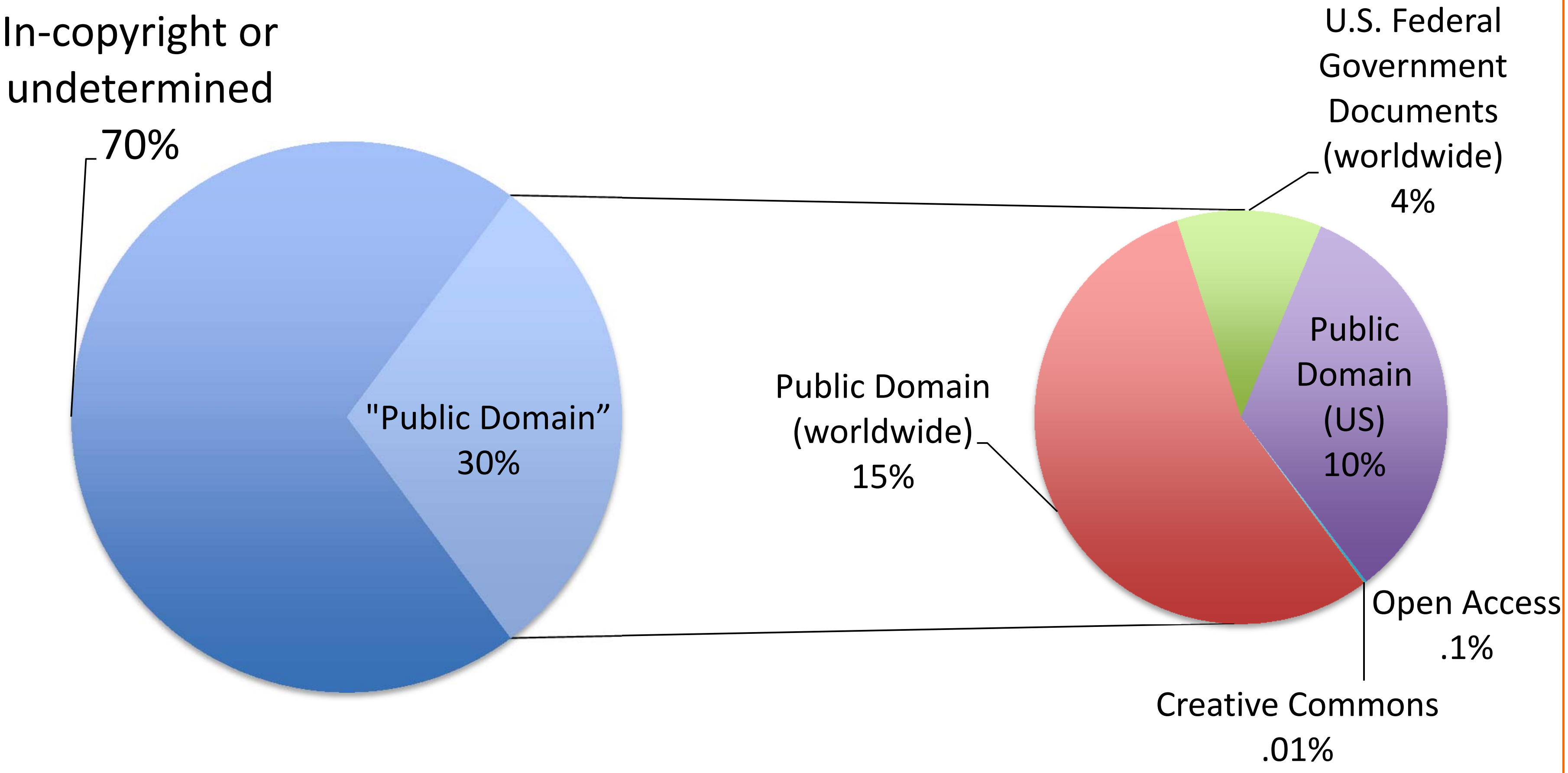
HathiTrust “Wow” Numbers

- 10,924,244 total volumes
- 5,719,252 book titles
- 285,776 serial titles
- 3,823,485,400 pages
- 490 terabytes
- 129 miles
- 8,876 tons
- 3,565,657 volumes(~33% of total) in the public domain

Content Sources

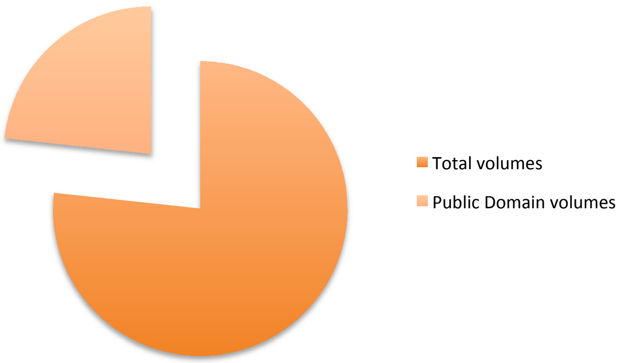


Content Distribution



Google PD Research Collection

- Public Domain Materials of the HatihTrust
  - 2,592,097 Volumes
  - Gigabytes
    - 2.3 TB in raw OCR'd text
    - 3.7 TB of managed OCR'd text
    - 1.85 TB solr Index
  - Monthly Updates
    - And irregular data 'take down' requests



Non-Consumptive Research Model

Non-Consumptive Research Paradigm

- *No action or set of actions on part of users, either acting alone or in cooperation with other users over duration of one or multiple sessions can result in sufficient information gathered from collection of copyrighted works to reassemble pages from collection.*
- Definition disallows collusion between users, or accumulation of material over time. Differentiates human researcher from proxy which is not a user. Users are human beings.

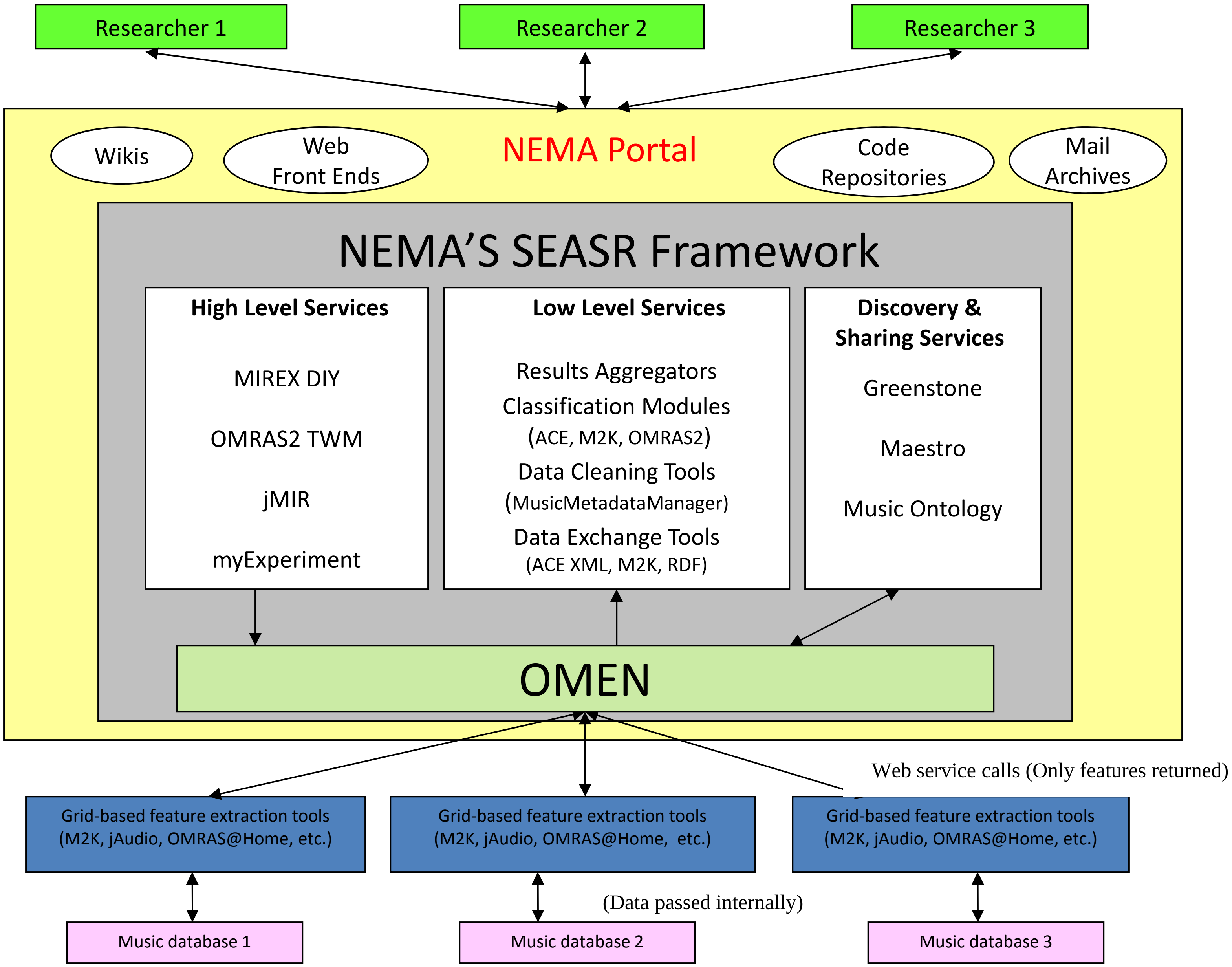
Non-Consumptive Research Paradigm

Bring the

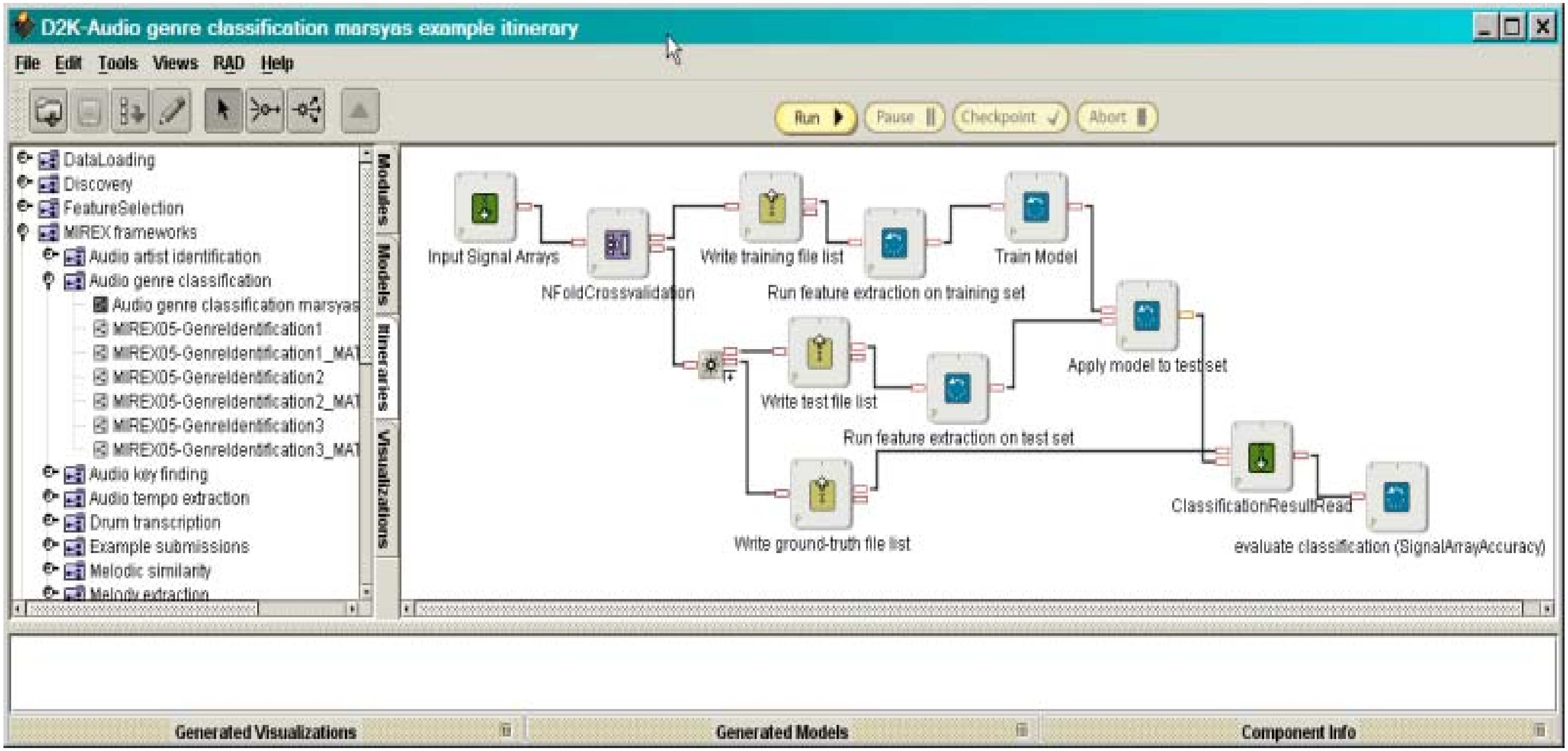
COMPUTATION

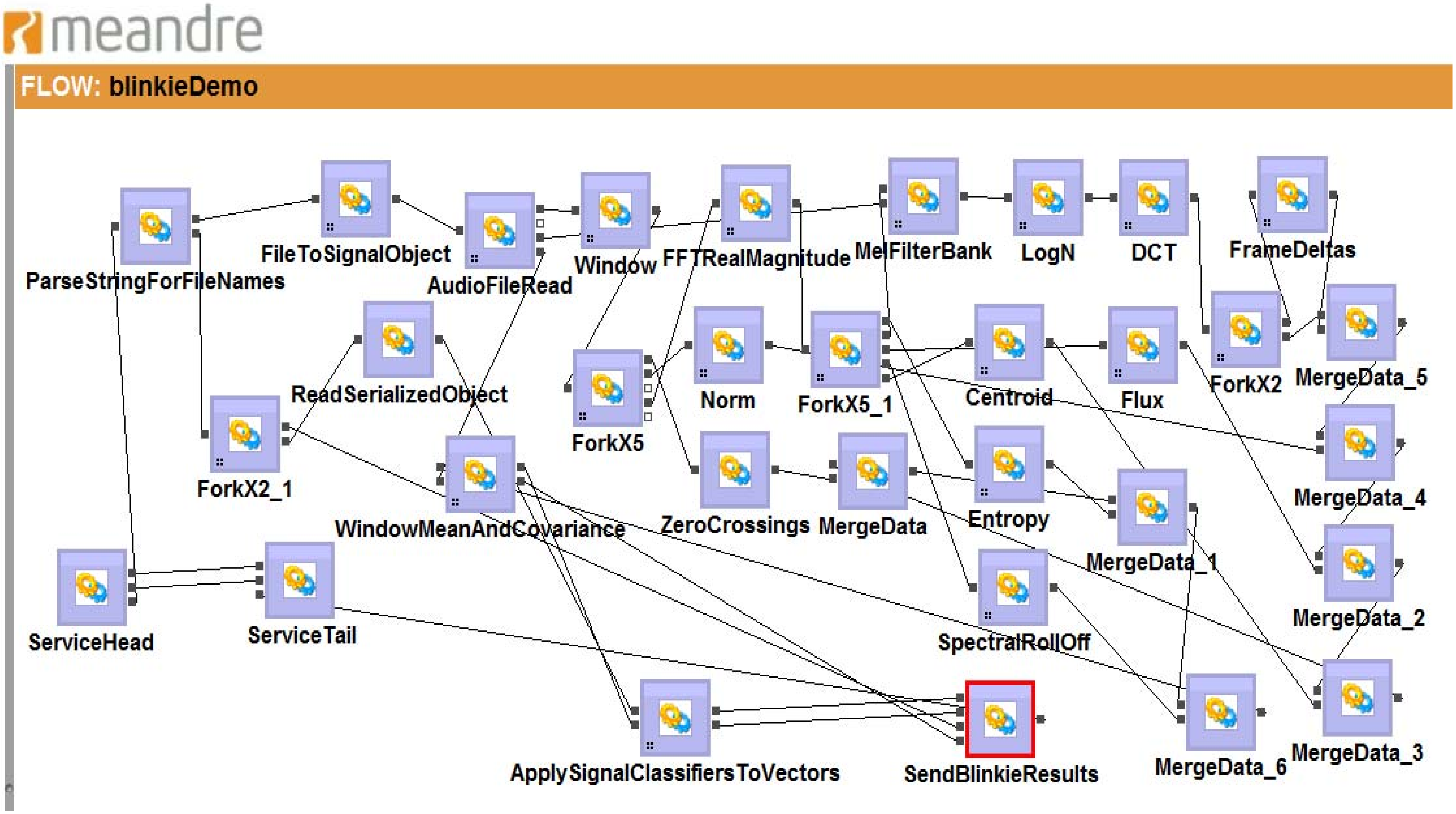
to the

DATA!


Researcher 1
Researcher 2
Researcher 3
Wikis
Web Front Ends
NEMA Portal
Code Repositories
Mail Archives
NEMA'S SEASR Framework
High Level Services
MIREX DIY
OMRAS2 TWM
jMIR
myExperiment
Low Level Services
Results Aggregators
Classification Modules (ACE, M2K, OMRAS2)
Data Cleaning Tools (MusicMetadataManager)
Data Exchange Tools (ACE XML, M2K, RDF)
Discovery & Sharing Services
Greenstone
Maestro
Music Ontology
OMEN
Web service calls (Only features returned)
Grid-based feature extraction tools (M2K, jAudio, OMRAS@Home, etc.)
Grid-based feature extraction tools (M2K, jAudio, OMRAS@Home, etc.)
Grid-based feature extraction tools (M2K, jAudio, OMRAS@Home, etc.)
(Data passed internally)
Music database 1
Music database 2
Music database 3

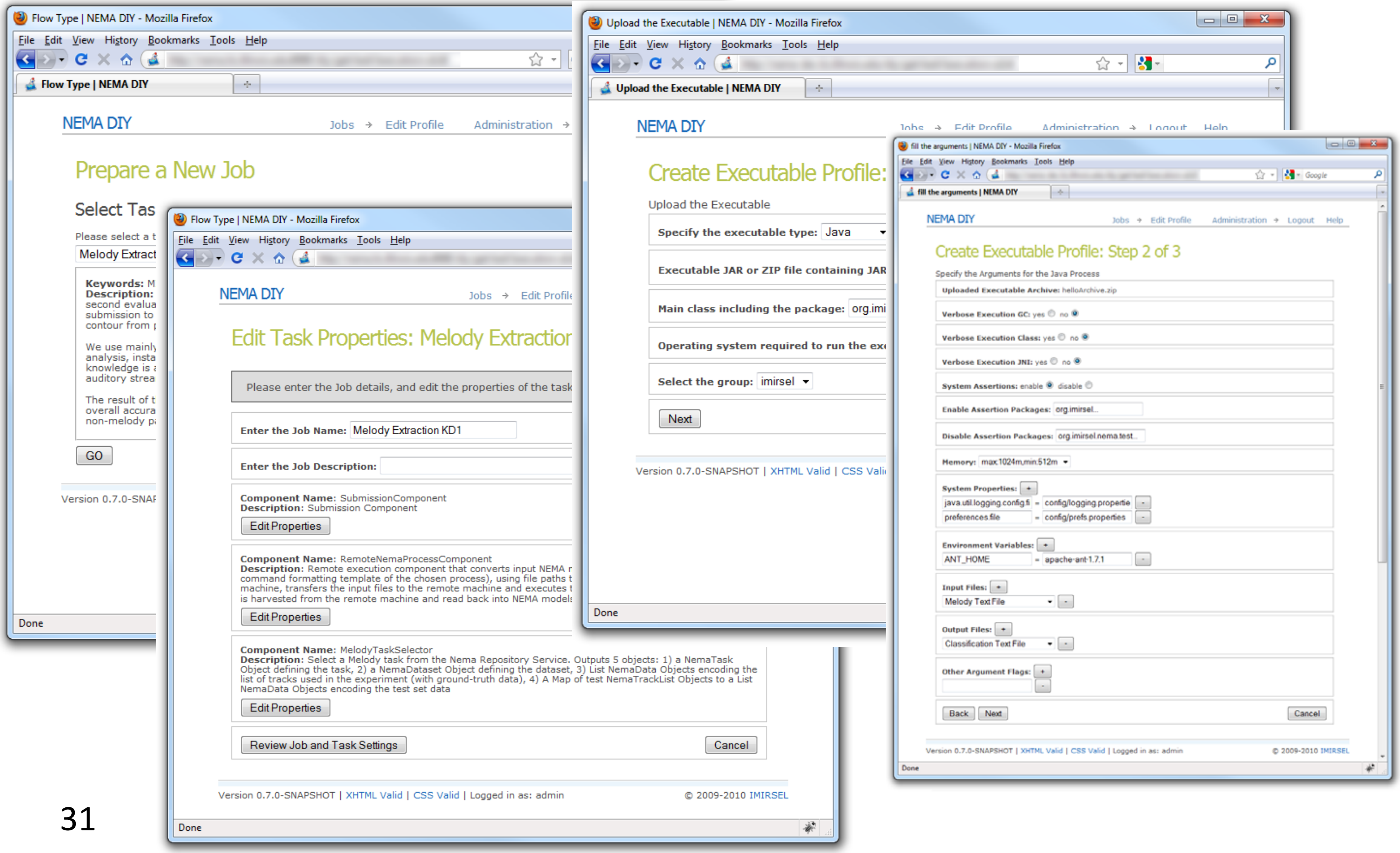
An External Classification Algorithm


D2K-Audio genre classification marsyas example itinerary
File
Edit
Tools
Views
RAD
Help
Run
Pause
Checkpoint
Abort
DataLoading
Discovery
FeatureSelection
MIREX frameworks
Audio artist identification
Audio genre classification
Audio genre classification marsyas
MIREX05-GenreIdentification1
MIREX05-GenreIdentification1_MAT
MIREX05-GenreIdentification2
MIREX05-GenreIdentification2_MAT
MIREX05-GenreIdentification3
MIREX05-GenreIdentification3_MAT
Audio key finding
Audio tempo extraction
Drum transcription
Example submissions
Melodic similarity
Melody extraction
Modules
Models
Itineraries
Visualizations
Input Signal Arrays
NFoldCrossvalidation
Write training file list
Run feature extraction on training set
Train Model
Apply model to test set
Write test file list
Run feature extraction on test set
Write ground-truth file list
ClassificationResultRead
evaluate classification (SignalArrayAccuracy)
Generated Visualizations
Generated Models
Component Info


meandre
FLOW: blinkieDemo
ParseStringForFileNames
FileToSignalObject
AudioFileRead
Window
FFTRealMagnitude
MelFilterBank
LogN
DCT
FrameDeltas
ReadSerializedObject
ForkX5
Norm
ForkX5_1
Centroid
Flux
ForkX2
MergeData_5
ForkX2_1
WindowMeanAndCovariance
ZeroCrossings
MergeData
Entropy
MergeData_1
MergeData_4
MergeData_2
ServiceHead
ServiceTail
SpectralRollOff
ApplySignalClassifiersToVectors
SendBlinkieResults
MergeData_6
MergeData_3

Networked Environment for Music Analysis

NEMA DIY Interface



Flow Type | NEMA DIY - Mozilla Firefox

NEMA DIY   Jobs → Edit Profile   Administration →

Prepare a New Job

Select Tas

Please select a t

Melody Extract

Keywords: M
Description:
second evalua
submission to
contour from

We use mainly
analysis, insta
knowledge is
auditory strea

The result of t
overall accura
non-melody p

GO

Version 0.7.0-SNAF

Done

Flow Type | NEMA DIY - Mozilla Firefox

NEMA DIY   Jobs → Edit Profile

Edit Task Properties: Melody Extraction

Please enter the Job details, and edit the properties of the task

Enter the Job Name: Melody Extraction KD1

Enter the Job Description:

Component Name: SubmissionComponent
Description: Submission Component
Edit Properties

Component Name: RemoteNemaProcessComponent
Description: Remote execution component that converts input NEMA r
command formatting template of the chosen process), using file paths t
machine, transfers the input files to the remote machine and executes t
is harvested from the remote machine and read back into NEMA models
Edit Properties

Component Name: MelodyTaskSelector
Description: Select a Melody task from the Nema Repository Service. Outputs 5 objects: 1) a NemaTask Object defining the task, 2) a NemaDataset Object defining the dataset, 3) List NemaData Objects encoding the list of tracks used in the experiment (with ground-truth data), 4) A Map of test NemaTrackList Objects to a List NemaData Objects encoding the test set data
Edit Properties

Review Job and Task Settings   Cancel

Version 0.7.0-SNAPSHOT | XHTML Valid | CSS Valid | Logged in as: admin   © 2009-2010 IMIRSEL

Done

Upload the Executable | NEMA DIY - Mozilla Firefox

NEMA DIY   Jobs → Edit Profile   Administration → Logout   Help

Create Executable Profile:

Upload the Executable

Specify the executable type: Java

Executable JAR or ZIP file containing JAR

Main class including the package: org.imi

Operating system required to run the exe

Select the group: imirsel

Next

Version 0.7.0-SNAPSHOT | XHTML Valid | CSS Vali

Done

fill the arguments | NEMA DIY - Mozilla Firefox

NEMA DIY   Jobs → Edit Profile   Administration → Logout   Help

Create Executable Profile: Step 2 of 3

Specify the Arguments for the Java Process

Uploaded Executable Archive: helloArchive.zip

Verbose Execution GC: yes ○ no ●

Verbose Execution Class: yes ○ no ●

Verbose Execution JNI: yes ○ no ●

System Assertions: enable ● disable ○

Enable Assertion Packages: org.imirsel...

Disable Assertion Packages: org.imirsel.nema.test...

Memory: max:1024m,min:512m

System Properties: +
java.util.logging.config.fi = config/logging.propertie -
preferences.file = config/prefs.properties -

Environment Variables: +
ANT_HOME = apache-ant-1.7.1 -

Input Files: +
Melody Text File -

Output Files: +
Classification Text File -

Other Argument Flags: +
-

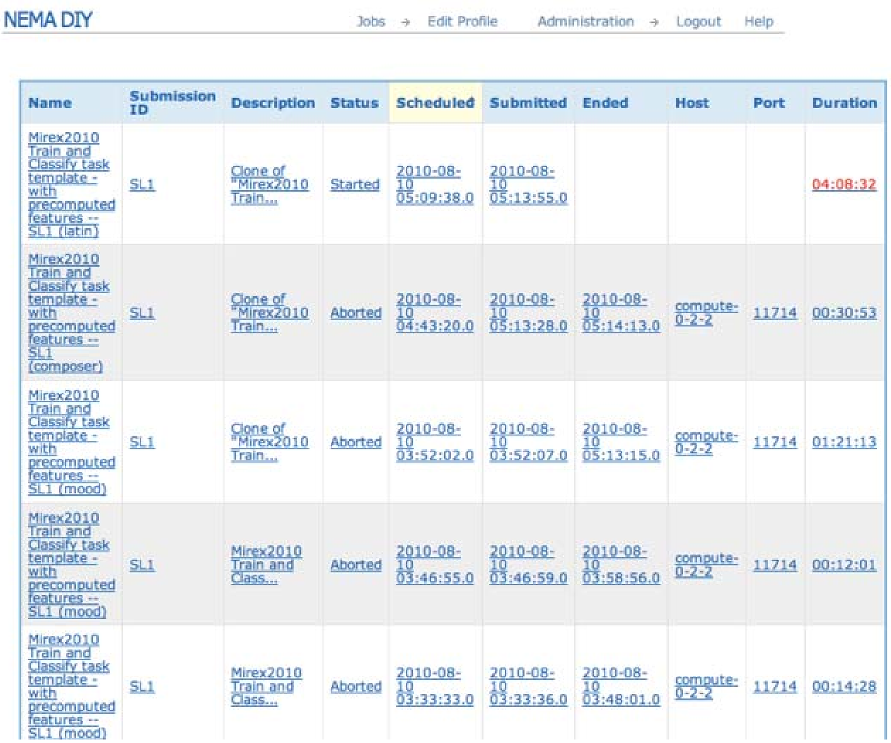
Back   Next   Cancel

Version 0.7.0-SNAPSHOT | XHTML Valid | CSS Valid | Logged in as: admin   © 2009-2010 IMIRSEL

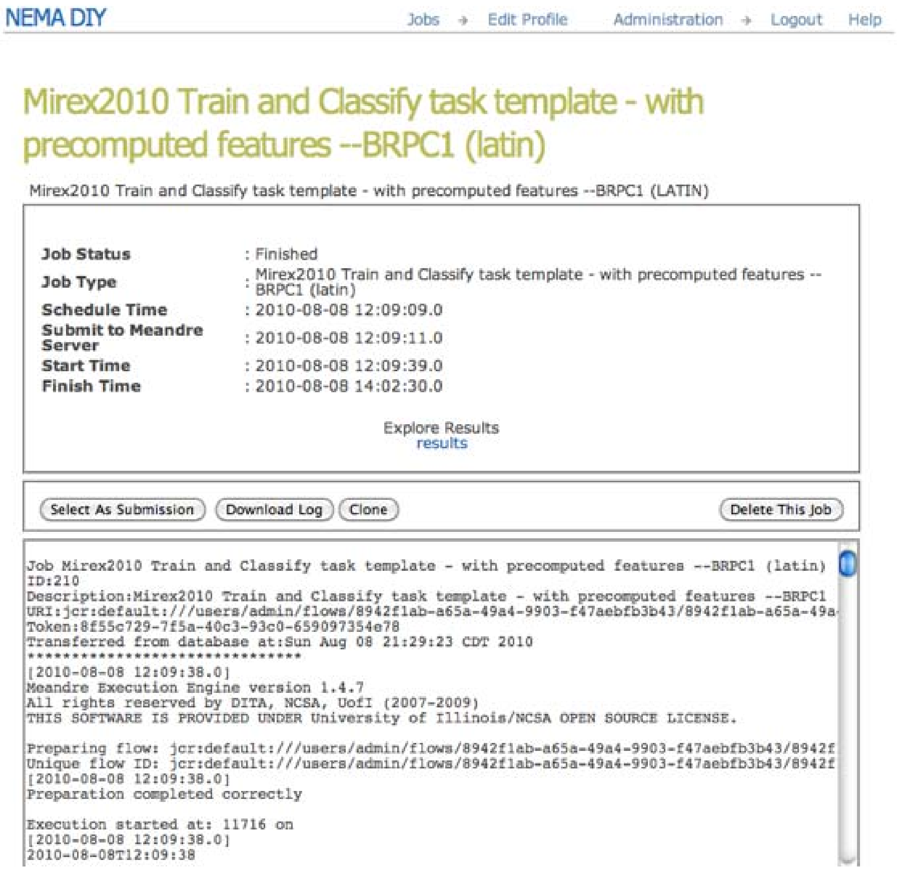
Done

Networked Environment for Music Analysis

NEMA DIY Interface



Name	Submission ID	Description	Status	Scheduled	Submitted	Ended	Host	Port	Duration
Mirex2010 Train and Classify task template - with precomputed features -- SL1 (latin)	SL1	Clone of "Mirex2010 Train...	Started	2010-08-10 05:09:38.0	2010-08-10 05:13:55.0				04:08:32
Mirex2010 Train and Classify task template - with precomputed features -- SL1 (composer)	SL1	Clone of "Mirex2010 Train...	Aborted	2010-08-10 04:43:20.0	2010-08-10 05:13:28.0	2010-08-10 05:14:13.0	compute-0-2-2	11714	00:30:53
Mirex2010 Train and Classify task template - with precomputed features -- SL1 (mood)	SL1	Clone of "Mirex2010 Train...	Aborted	2010-08-10 03:52:02.0	2010-08-10 03:52:07.0	2010-08-10 05:13:15.0	compute-0-2-2	11714	01:21:13
Mirex2010 Train and Classify task template - with precomputed features -- SL1 (mood)	SL1	Mirex2010 Train and Class...	Aborted	2010-08-10 03:46:55.0	2010-08-10 03:46:59.0	2010-08-10 03:58:56.0	compute-0-2-2	11714	00:12:01
Mirex2010 Train and Classify task template - with precomputed features -- SL1 (mood)	SL1	Mirex2010 Train and Class...	Aborted	2010-08-10 03:33:33.0	2010-08-10 03:33:36.0	2010-08-10 03:48:01.0	compute-0-2-2	11714	00:14:28



Mirex2010 Train and Classify task template - with precomputed features --BRPC1 (latin)

Mirex2010 Train and Classify task template - with precomputed features --BRPC1 (LATIN)

Job Status	: Finished
Job Type	: Mirex2010 Train and Classify task template - with precomputed features -- BRPC1 (latin)
Schedule Time	: 2010-08-08 12:09:09.0
Submit to Meandre Server	: 2010-08-08 12:09:11.0
Start Time	: 2010-08-08 12:09:39.0
Finish Time	: 2010-08-08 14:02:30.0

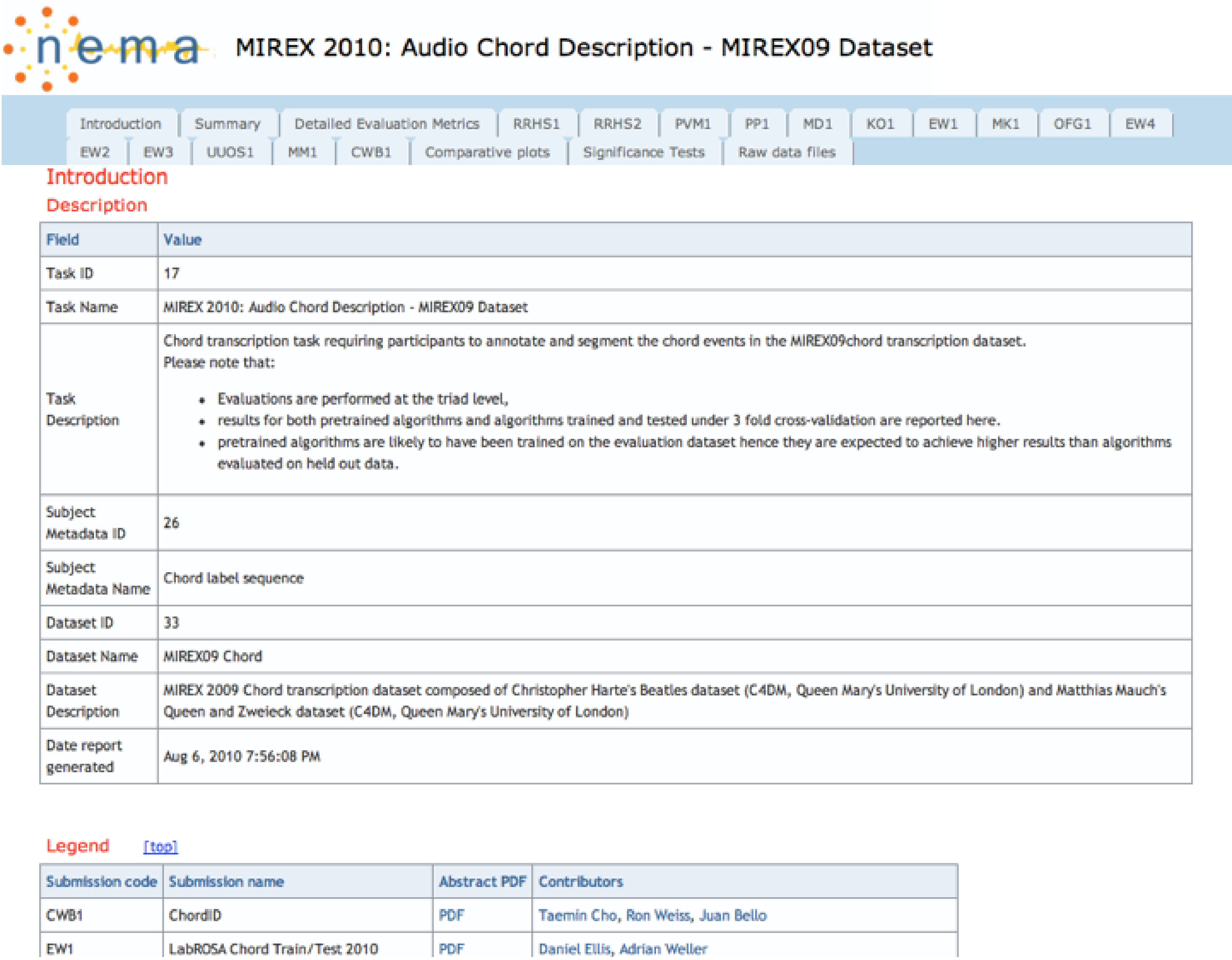
Explore Results
results

```
Job Mirex2010 Train and Classify task template - with precomputed features --BRPC1 (latin)
ID:210
Description:Mirex2010 Train and Classify task template - with precomputed features --BRPC1
URI:jcr:default:///users/admin/flows/8942f1ab-a65a-49a4-9903-f47aebfb3b43/8942f1ab-a65a-49a-
Token:8f55c729-7f5a-40c3-93c0-659097354e78
Transferred from database at:Sun Aug 08 21:29:23 CDT 2010
*******************************
[2010-08-08 12:09:38.0]
Meandre Execution Engine version 1.4.7
All rights reserved by DITA, NCSA, UofI (2007-2009)
THIS SOFTWARE IS PROVIDED UNDER University of Illinois/NCSA OPEN SOURCE LICENSE.

Preparing flow: jcr:default:///users/admin/flows/8942f1ab-a65a-49a4-9903-f47aebfb3b43/8942f
Unique flow ID: jcr:default:///users/admin/flows/8942f1ab-a65a-49a4-9903-f47aebfb3b43/8942f
[2010-08-08 12:09:38.0]
Preparation completed correctly

Execution started at: 11716 on
[2010-08-08 12:09:38.0]
2010-08-08T12:09:38
```

Evaluation Reports



MIREX 2010: Audio Chord Description - MIREX09 Dataset

Introduction

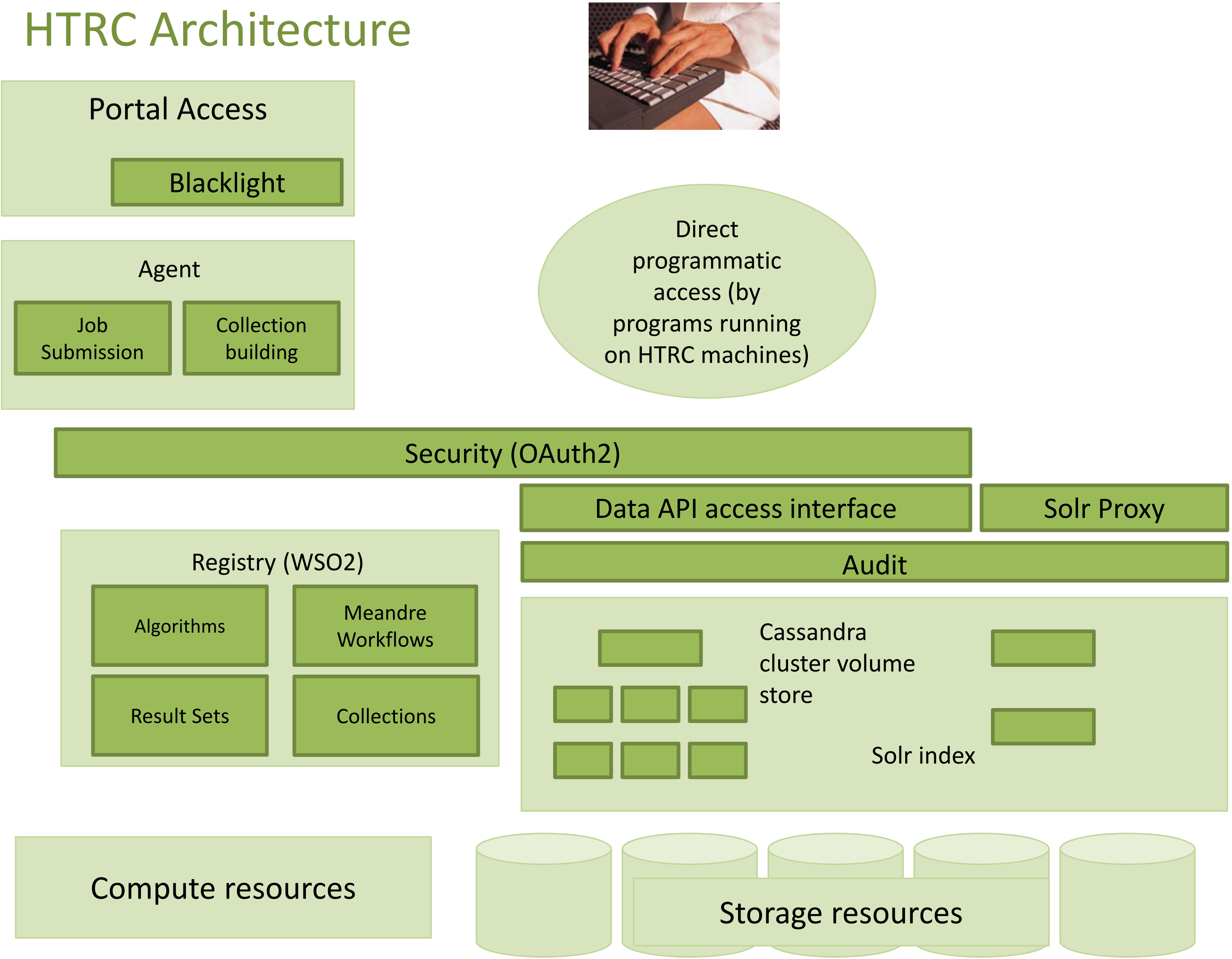
Description

Field	Value
Task ID	17
Task Name	MIREX 2010: Audio Chord Description - MIREX09 Dataset
Task Description	Chord transcription task requiring participants to annotate and segment the chord events in the MIREX09chord transcription dataset. Please note that: • Evaluations are performed at the triad level, • results for both pretrained algorithms and algorithms trained and tested under 3 fold cross-validation are reported here. • pretrained algorithms are likely to have been trained on the evaluation dataset hence they are expected to achieve higher results than algorithms evaluated on held out data.
Subject Metadata ID	26
Subject Metadata Name	Chord label sequence
Dataset ID	33
Dataset Name	MIREX09 Chord
Dataset Description	MIREX 2009 Chord transcription dataset composed of Christopher Harte's Beatles dataset (C4DM, Queen Mary's University of London) and Matthias Mauch's Queen and Zweieck dataset (C4DM, Queen Mary's University of London)
Date report generated	Aug 6, 2010 7:56:08 PM

Legend [top]

Submission code	Submission name	Abstract PDF	Contributors
CWB1	ChordID	PDF	Taemin Cho, Ron Weiss, Juan Bello
EW1	LabROSA Chord Train/Test 2010	PDF	Daniel Ellis, Adrian Weller

HTRC Architecture



Portal Access

Blacklight

Agent

Job Submission

Collection building

Direct programmatic access (by programs running on HTRC machines)

Security (OAuth2)

Data API access interface

Solr Proxy

Registry (WSO2)

Algorithms

Meandre Workflows

Result Sets

Collections

Audit

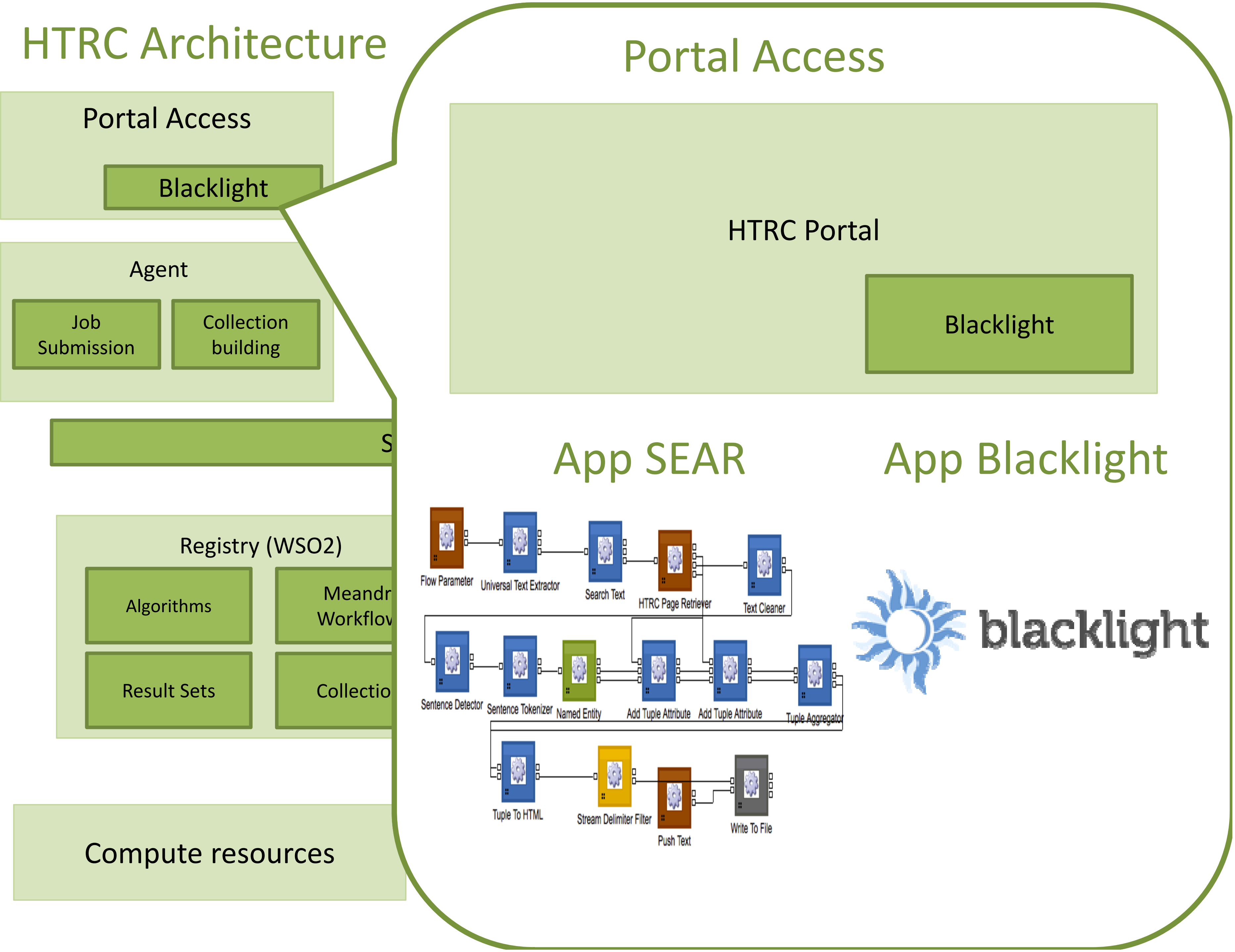
Cassandra cluster volume store

Solr index

Compute resources

Storage resources

HTRC Architecture



Portal Access

Blacklight

Agent

Job Submission

Collection building

S

Registry (WSO2)

Algorithms

Meandr Workflov

Result Sets

Collectio

Compute resources

Portal Access

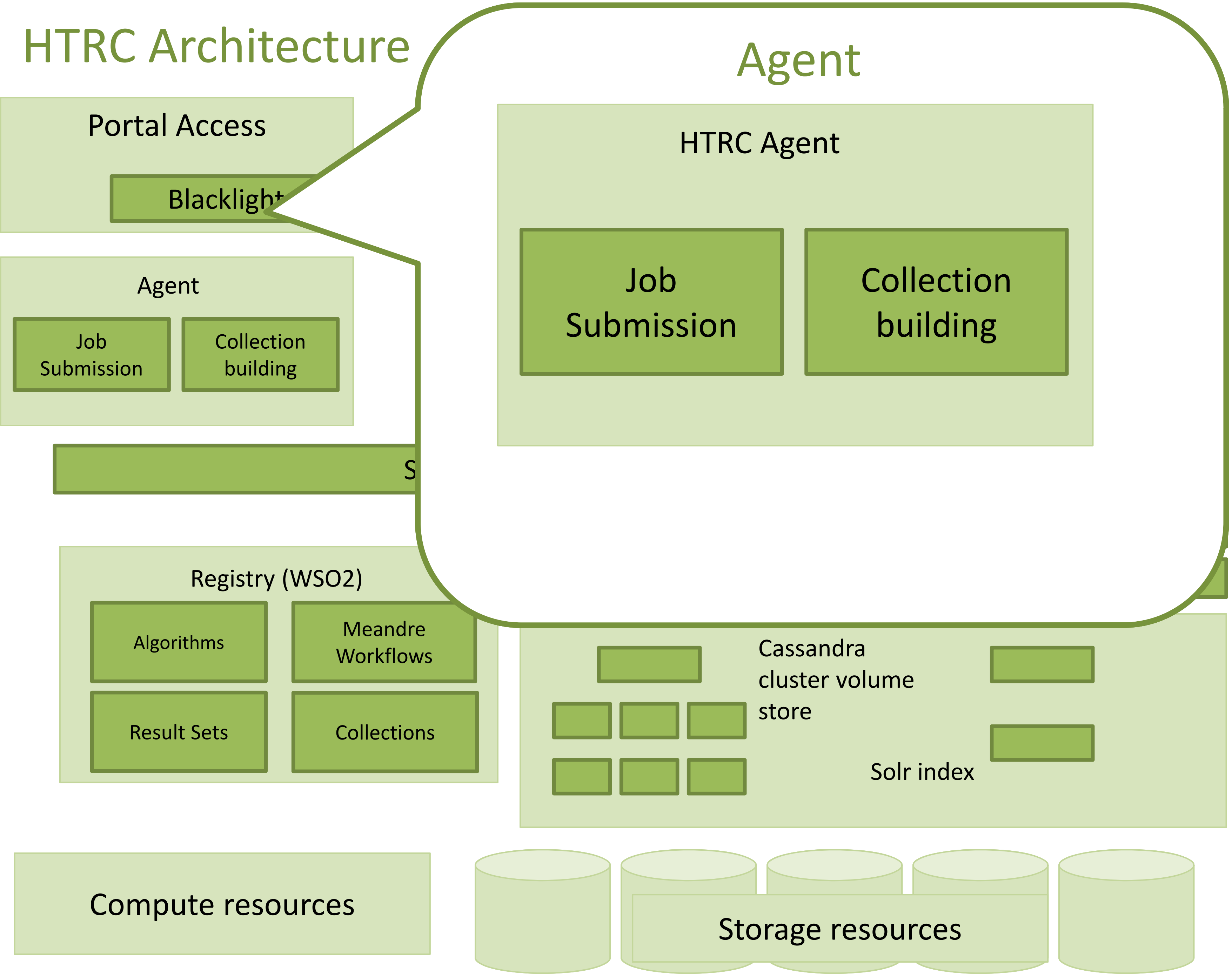
HTRC Portal

Blacklight

App SEAR

App Blacklight

HTRC Architecture



Agent

HTRC Agent

Job Submission

Collection building

Portal Access

Blacklight

Agent

Job Submission

Collection building

S

Registry (WSO2)

Algorithms

Meandre Workflows

Result Sets

Collections

Cassandra cluster volume store

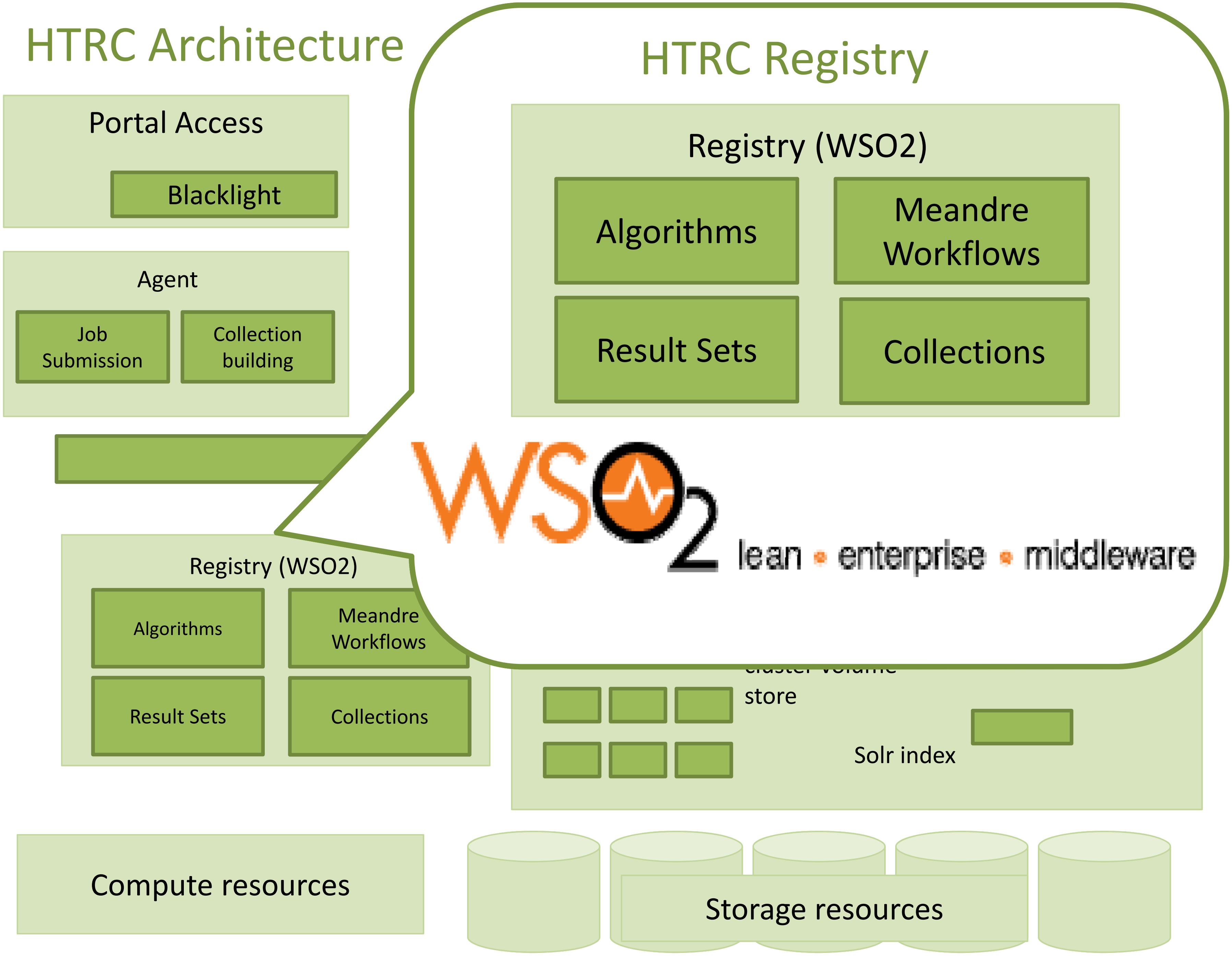
Solr index

Compute resources

Storage resources

HTRC Architecture

HTRC Registry



Registry (WSO2)

Algorithms

Meandre Workflows

Result Sets

Collections

WSO2 lean • enterprise • middleware

Portal Access

Blacklight

Agent

Job Submission

Collection building

Registry (WSO2)

Algorithms

Meandre Workflows

Result Sets

Collections

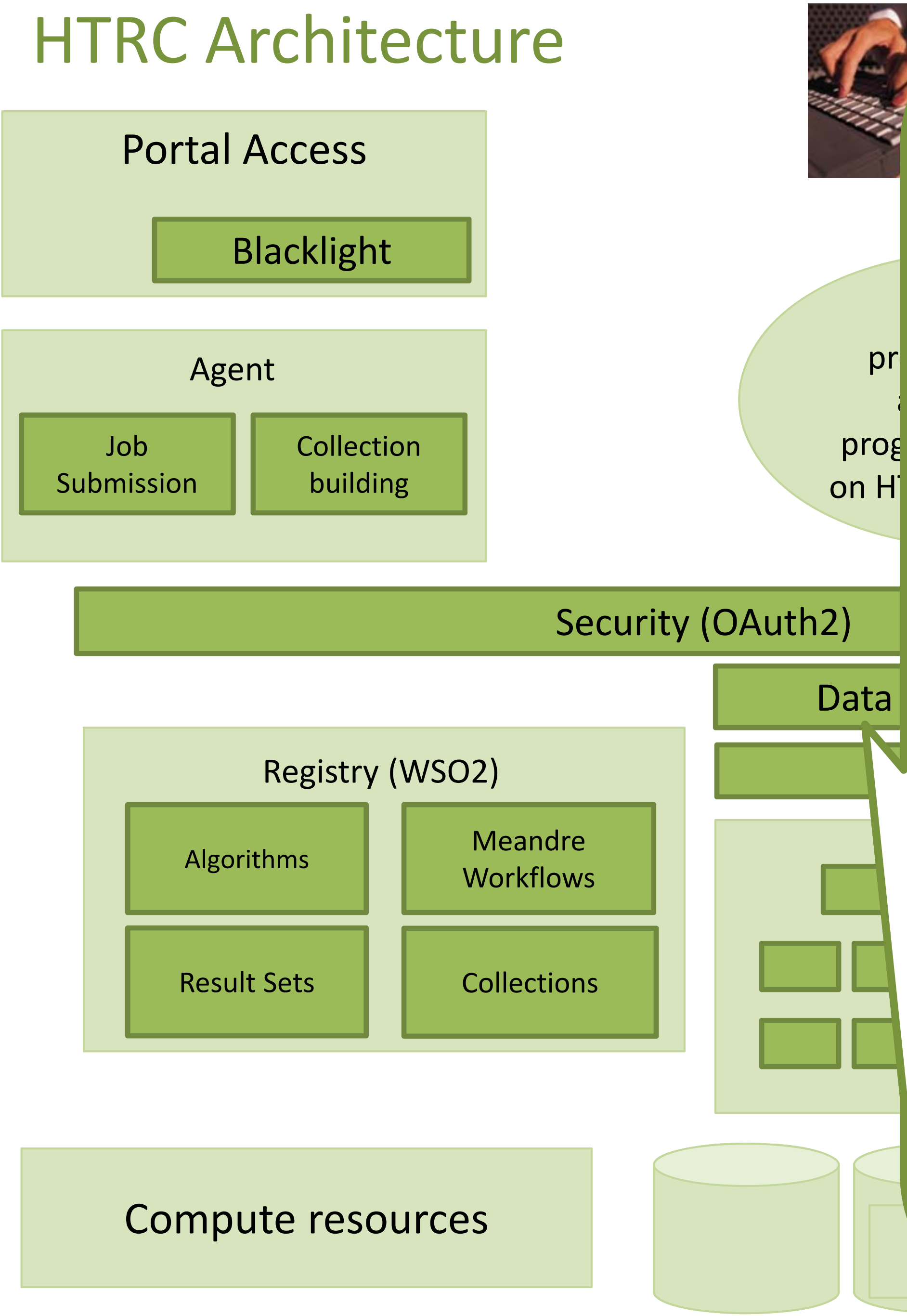
cluster volume store

Solr index

Compute resources

Storage resources

HTRC Architecture

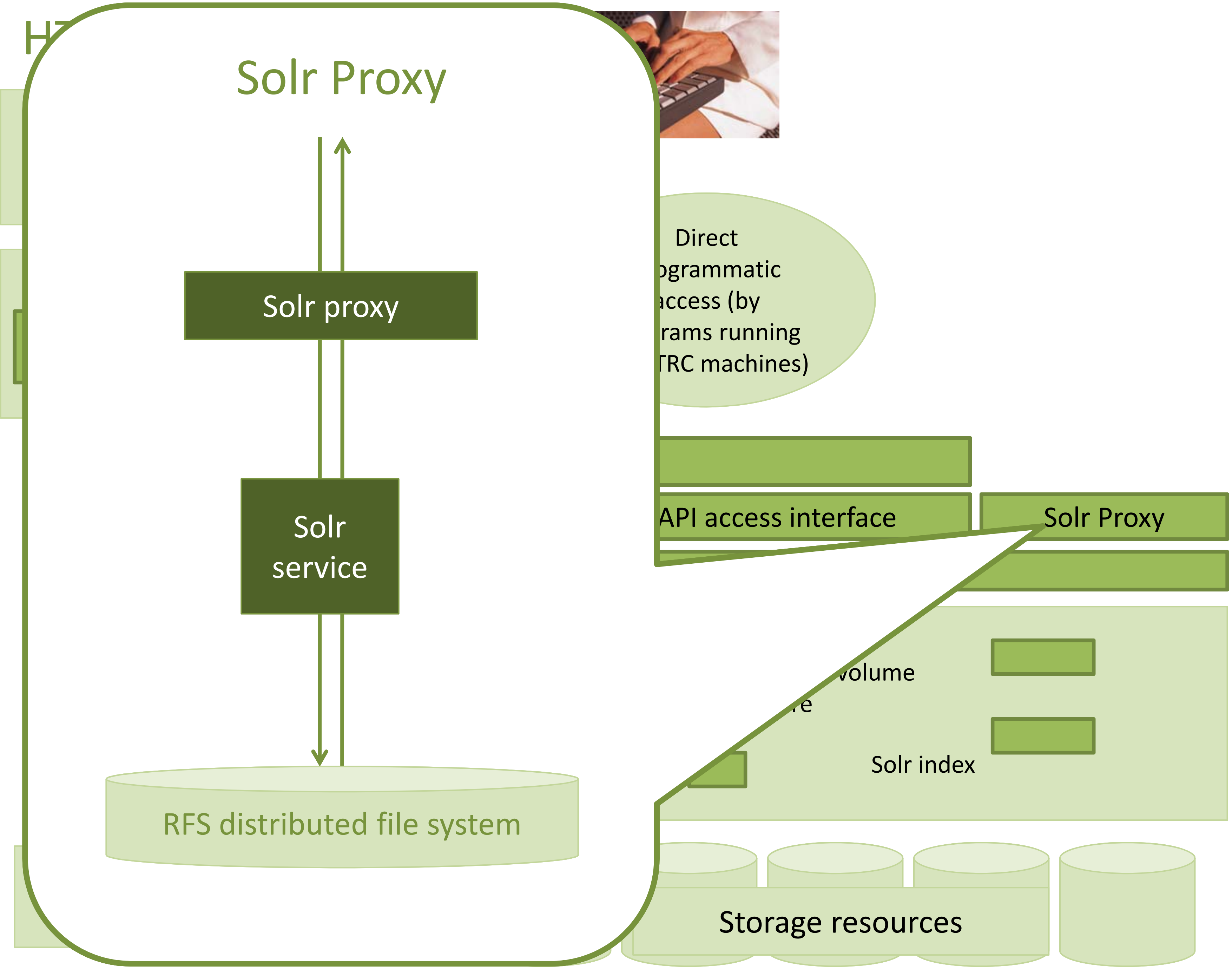


Secure Data API

- RESTful Web Service
  - Language agnostic
  - Clients don’t have to deal with Cassandra
- Simple OAuth2 authentication
- HTTP over SSL
- Audits client access
- Protected behind firewall, accessible only to authorized IPs

HTRC

Solr Proxy



Solr proxy

Solr service

RFS distributed file system

HT

Direct
ogrammatic
access (by
rams running
TRC machines)

API access interface

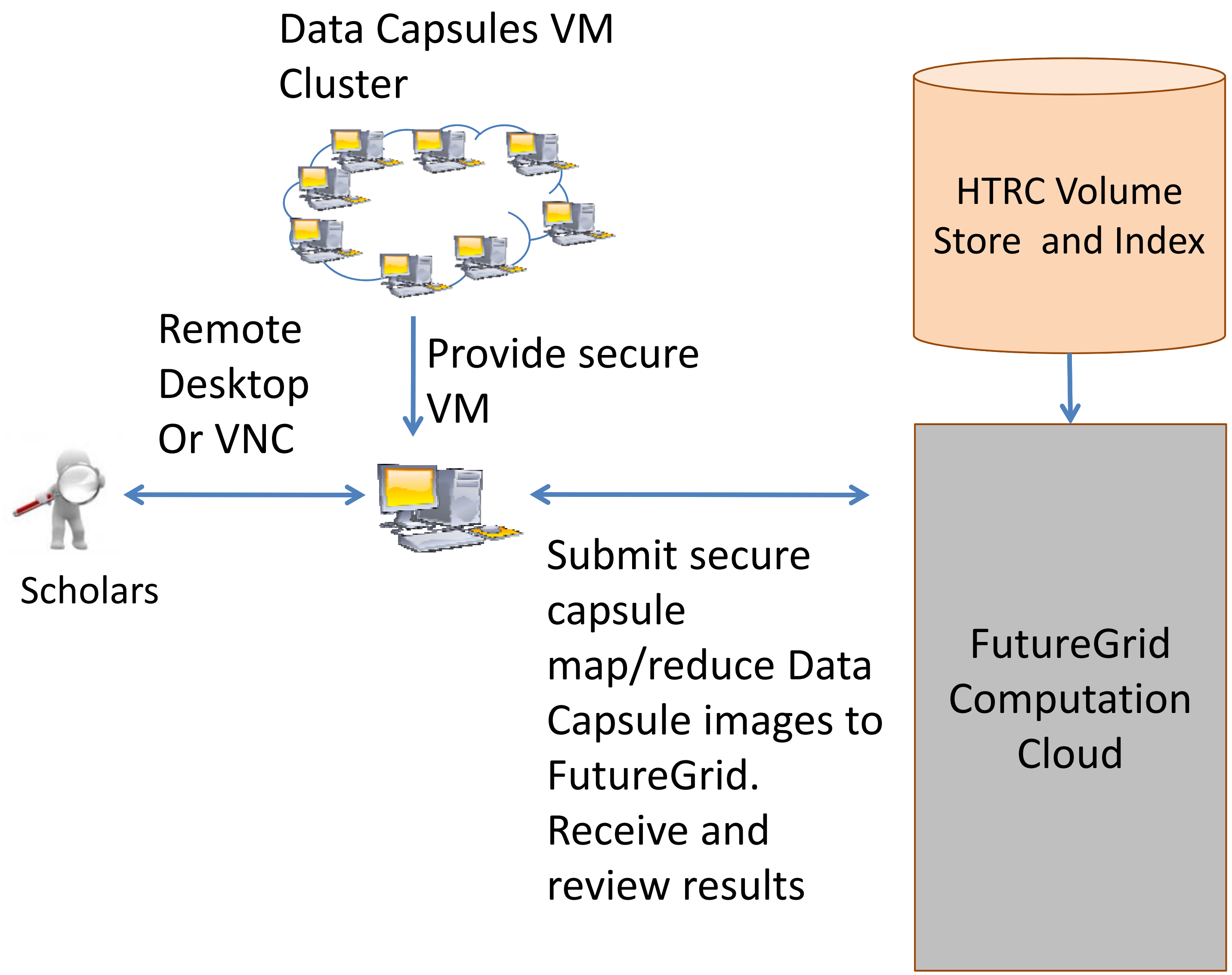
Solr Proxy

volume

e

Solr index

Storage resources



Data Capsules VM Cluster

Remote Desktop Or VNC

Provide secure VM

Scholars

Submit secure capsule map/reduce Data Capsule images to FutureGrid. Receive and review results

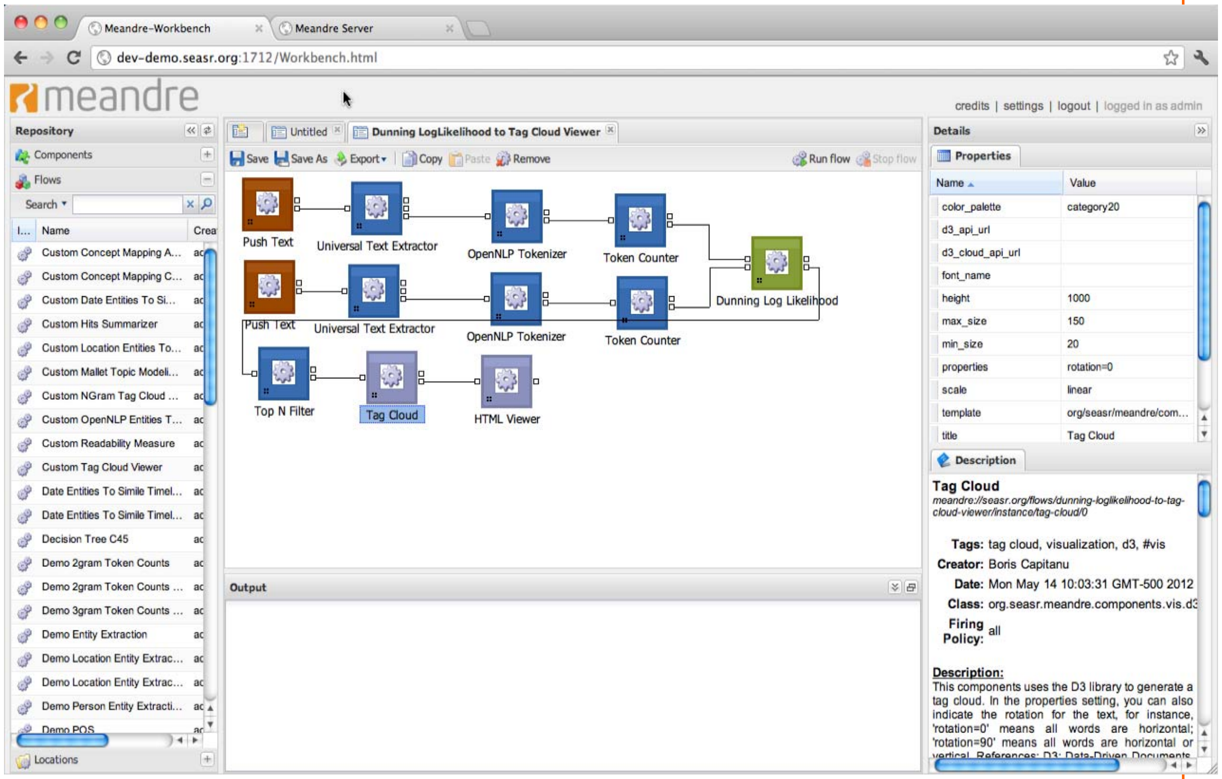
HTRC Volume Store and Index

FutureGrid Computation Cloud

Non-Consumptive Research-Secure Data Capsule

Meandre: Workbench Existing Flow

- Web-based UI
- Components and flows are retrieved from server
- Additional locations of components and flows can be added to server
- Create flow using a graphical drag and drop interface
- Change property values
- Execute the flow



I...	Name	Crea
	Custom Concept Mapping A...	ac
	Custom Concept Mapping C...	ac
	Custom Date Entities To Si...	ac
	Custom Hits Summarizer	ac
	Custom Location Entities To...	ac
	Custom Mallet Topic Modeli...	ac
	Custom NGram Tag Cloud ...	ac
	Custom OpenNLP Entities T...	ac
	Custom Readability Measure	ac
	Custom Tag Cloud Viewer	ac
	Date Entities To Simile Timel...	ac
	Date Entities To Simile Timel...	ac
	Decision Tree C45	ac
	Demo 2gram Token Counts	ac
	Demo 2gram Token Counts ...	ac
	Demo 3gram Token Counts ...	ac
	Demo Entity Extraction	ac
	Demo Location Entity Extrac...	ac
	Demo Location Entity Extrac...	ac
	Demo Person Entity Extracti...	ac
	Demo POS	ac

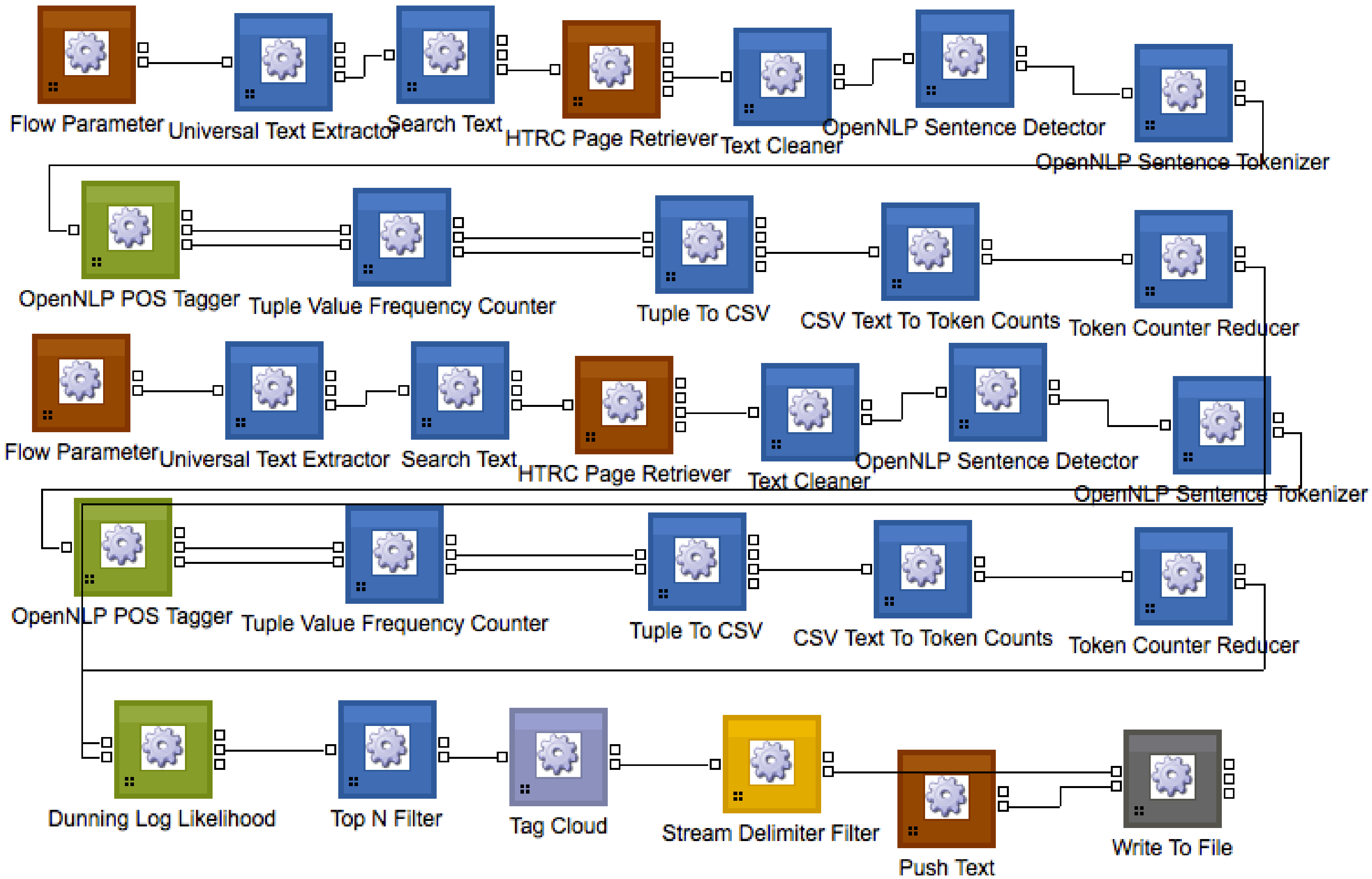
Name	Value
color_palette	category20
d3_api_url	
d3_cloud_api_url	
font_name	
height	1000
max_size	150
min_size	20
properties	rotation=0
scale	linear
template	org/seasr/meandre/com...
title	Tag Cloud

Tag Cloud
meandre://seasr.org/flows/dunning-loglikelihood-to-tag-cloud-viewer/instance/tag-cloud/0

Tags: tag cloud, visualization, d3, #vis
Creator: Boris Capitanu
Date: Mon May 14 10:03:31 GMT-500 2012
Class: org.seasr.meandre.components.vis.d3
Firing Policy: all

Description:
This components uses the D3 library to generate a tag cloud. In the properties setting, you can also indicate the rotation for the text, for instance, 'rotation=0' means all words are horizontal; 'rotation=90' means all words are horizontal or vertical. References: D3: Data-Driven Documents

Meandre Flow



Workset Creation for Scholarly Analysis (WCSA)

Workset Creation for Scholarly Analysis: Prototyping Project

- Collection analysis and prototype tools & services to facilitate work-set creation
  - J. Stephen Downie, Tim Cole, Beth Plale
  - Andrew W. Mellon Foundation
  - 1 July 2013 - 30 June 2015
- Proposal Narrative:
  - http://bit.ly/htrrcworksetgrant

WSCA Project Goals

- The ***Workset Creation for Scholarly Analysis: Prototyping Project*** (WCSA) seeks to address three sets of tightly intertwined research questions regarding:

1. ***Enriching*** the metadata describing the HathiTrust corpus through mining of the resources themselves and leveraging end-user annotations;

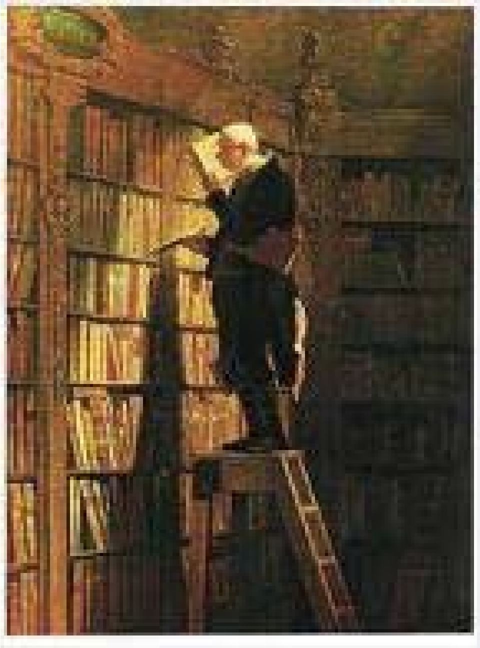
2. ***Augmenting*** string-based metadata with URIs to leverage external services and Linked Open Data to facilitate discovery and the process of organizing HathiTrust resources into collections and worksets; and,

3. ***Formalizing*** the notion of collections and worksets in the context of the HathiTrust Research Center.

Motivation & Models

Collections, corpora, worksets, ...:

- Aggregations of items brought together in some context:
  - Archival
  - Curatorial
  - Experimental
  - Referential
  - Thematic (for research)



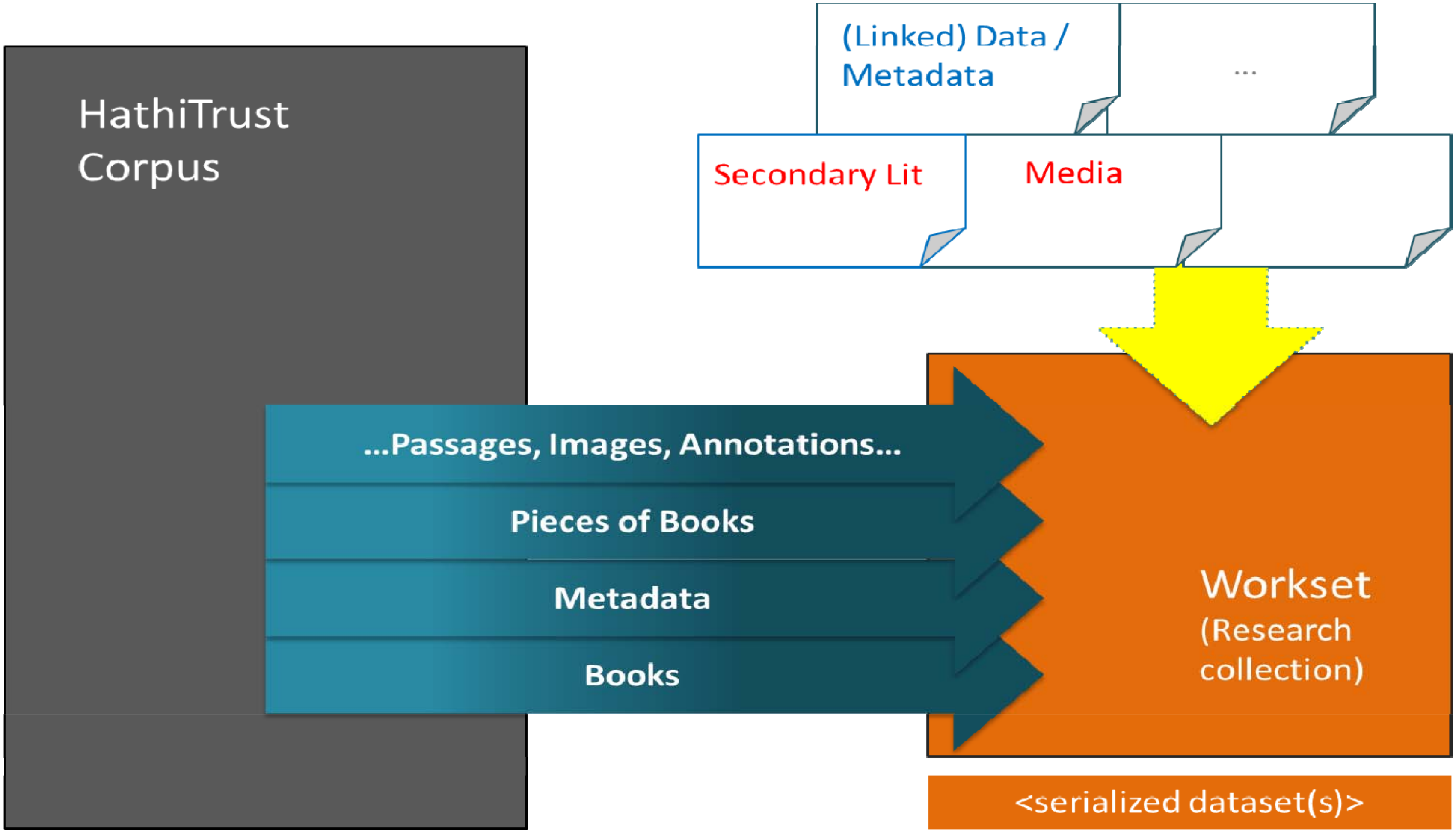
Carl Spitzweg. 1850
The Bookworm (*Der Bücherwurm*)

Grand Motivation

- *The ability to slice through a massive corpus constructed from many different library collections, and out of that to construct the precise workset required for a particular scholarly investigation, is an example of the "game changing" potential of the HathiTrust...*

What is a Workset?

1. A workset is an aggregation of materials brought together for the purpose of analysis.
2. Worksets are conceptual and must be expressible in a variety of ways
   - Need to allow creation outside of HathiTrust
   - Need to facilitate inclusion of resources beyond HathiTrust
   - Need to facilitate the inclusion of resources at many different levels of granularity beyond the book
3. Worksets encapsulate the specific materials that underwent analysis.
   - Need to capture provenance information
   - Possible recording of parameters
4. Worksets should be able to spawn descendants but otherwise immutable



HathiTrust Corpus

(Linked) Data / Metadata

...

Secondary Lit

Media

...Passages, Images, Annotations...

Pieces of Books

Metadata

Books

Workset (Research collection)

<serialized dataset(s)>

Scope

Dimensions of Workset Creation (Illustrative)

My work-set should contain (inspired by 2012 UnCamp):

- Volumes pertaining to Japan / in Japanese
- All volumes relevant to the study of Francis Bacon
- Music scores or notation extracted from HT volumes
- Images of Victorian England extracted from HT vols.
- Volumes in HT similar to TCP-ECCO novels
- 19th c. English-language novels by female authors
- Representative sample (by pub date & genre) of French language items in HT

MARC Metadata Shortcomings I

MARC Field	Percent of records in OCLC having instance of this field
245 Title Statement	> 99%
260 Publication Distribution, etc.	92%
500 General Note	41%
650 Topical Term / 653 Index Term – Uncontrolled	39% / 13%
050 LC Classification No / 082 Dewey Classification No	17% / 13%
655 Index Term -- Genre Form	12%

Table 2. Frequency of MARC fields in OCLC Records

MARC Metadata Shortcomings II

MARC Field	Percent of British Novel MARC records having instance of this field
650 Topical Term	6%
050 LC Classification No / 082 Dewey Classification No	27% / 4%
655 Index Term -- Genre Form	5%

Table 3. Frequency of MARC fields used in 2,386 descriptions of 19th century British novels digitized from UIUC collections

Why Worksets?

- The result of a first-level, rough filter
- Better scale for intensive analytics
- Provides essential scope for certain analytics
  - Word frequency scope over Bacon's essays
- Some tools (are trained to) work best on a narrow, homogeneous work-set
- Eliminate noise that would otherwise arise by asking questions across whole of HT

Research Questions (Illustrative only)

- Can we enrich the HathiTrust corpus metadata by distilling analytics over full text?
- Can we augment string-based metadata with URIs for recognized entities – e.g., names, subjects, publication location, etc. -- and by doing so can we leverage external services to facilitate discovery and clustering of resources?
- Can we leverage existing, well-defined external corpora to identify complementary subsets of HT volumes, and having done so can we demonstrate the ability to create and perform analytics over an integrated workset that includes resources external to HT?

Key Workset Questions

- Can we formalize the notion of collections and worksets in the HTRC context?
- What are the necessary elements of a “collection”? What are the necessary elements of a “workset”?
- How can we balance rigor with extensibility and flexibility?
- What roles do “data”, “metadata”, “annotations”, “tags”, “feature sets”, and so on, all play in the conception, creation, use and reuse of collections and worksets?

Two Project Streams

- Workset formal structures and semantics
  - Work in conjunction with Center for Informatics Research in Science and Scholarship at the Graduate School of Library and Information Science
- WCSA Prototyping Projects
  - Four projects funded by the grant but conducted by community teams

WCSA Timeline

- July 2013: Project Start
- Q1: User needs assessments / focus groups
- Q2: HT Corpus characterization
Request For Prototype Proposals
- Q3: RFP Finalist Workshop (Chicago) February 20
Prototype experiment funding awarded
- Q4-6: Prototype experiments done
Metadata workflow & work-set modeling
- Q7-8: Planning for prototype to production
Report out
- June 2015: Project ends

Prototype Grants

As part of project, HTRC will make 4 sub-awards

- $40K awarded to each of 4 non-HTRC teams
- HTRC will collaborate with each team
  - Access to representative test data / metadata set
  - Collaborate on work with HT / HTRC APIs, etc.

RFP & Sub-Award Schedule	
2013-11-15	RFP Available
2013-12-16	Letters of Intent Due (preferred)
2014-01-15	Final Proposals Due
2014-02-20	Finalist Meeting
~ 2014-03-15	Award Notification for projects running April-Dec, 2014

WCSA Summary

- Worksets are fundamental to the scholarly computational analysis enterprise
- We need a better understanding of their:
  - Constituent parts
  - Creation
  - Manipulation
  - Use and reuse
- Prototypes to lead to deeper tool development and metadata enhancement

Next Steps

Personal Goals for HTRC

- Engage in more collaborative projects
- Expand to have truly international partnerships
- Make sure to move beyond text
- Make sure to move beyond humanities!

Redux: Ongoing Challenges

How do we actually unlock the potential of 3 billion pages of human knowledge?

- Data quality issues
- Data structure challenges
- Metadata shortcomings
- Overcoming copyright barriers to research
  - Non-consumptive research
  - Computation to the data
- Moving beyond text
- Community building important and ongoing

Questions? Comments? Suggestions?

Special thanks to:

Jeremy York, Stacy Kowalczyk and Loretta Auvil