MUSE

Smart Connector Technology for Harvesting

MuseGlobal, Inc.
One Embarcadero
Suite 500
San Francisco, CA 94111
415 896-6873
www.museglobal.com

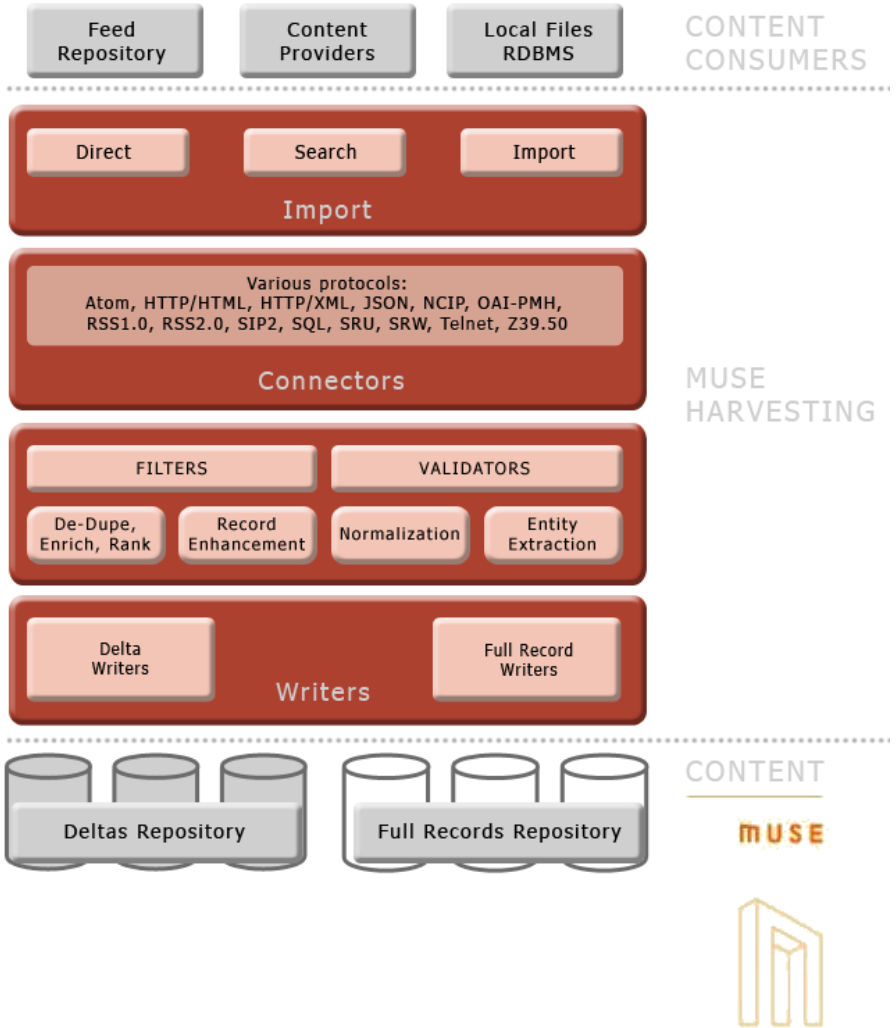
MuseGlobal S.A
Calea Bucuresti
Bl. 27B, Sc. 1, Ap. 10
Craiova, România
40 251-413496
www.museglobal.ro

EduLib, S.R.L.
Calea Bucuresti
Bl. 27B, Sc. 1, Ap. 2
Craiova, România
40 351-420970
www.edulib.com

Version: 1.2
Date: 15th November 2016
Author: EduLib, S.R.L.

Harvesting in MuseKnowledge

- **Harvesting in MuseKnowledge is based on different types of controlled extraction:**
  - Search-based
  - Date-based
  - Feed-based
- **Aggregated from any number of sources**
- **Gathered by feed or query**
- **Timed or 'on demand' operation**
- **Consistent delivery record format**
- **Enhanced, 'virtual records'**
- **Delivered by file or feed**
- **Deliver both full results or deltas**



Features

- **Strong Base, Strong Results.** Running on top of The MuseKnowledge™ Platform, its import capabilities, supported protocols, filters, validators and writers are capable of filling any repository with the right data.
- **Flexible Configuration.** MuseKnowledge™ Administrator Consoles, MuseKnowledge™ Control Center are used to administer MuseKnowledge™ Harvesting Applications which are not much different than a MuseKnowledge™ Federated Search Application.
- **Input and Output.** Up to 10K records per run per input, delivery via files, feeds, deltas or full records repository.
- **Retrieve Records from all Sources by:**
  - Searching the Deep Web, databases and Applications;
  - Polling Feeds (RSS, Atom, etc.);
  - Focused Crawling of web pages.
- **Detailed Extraction of Content:**
  - Page, Record and Source specific data extraction;
  - Multiple records per page or just one, or multiple pages per record;
  - Extract down to finer than field level granularity.
- **Contiguous Extraction.** Based on date interval, number of records, or combined. The extraction order of slices is guarantied for 2 successive extractions.

Features

- **Normalization and Consistency**
  - Consistent field level semantic normalization;
  - Specialized parsers for unstructured elements.
- **Regular Timed Processing**
  - Aligned on minutes granularity;
  - Irregular frequency.
- **Completeness**
  - Comparison of records retrieved and number reported;
  - Immediate (or next run), multiple re-tries (network congestion, etc.);
  - MD5 checksums for transmission.
- **The Right Content.** Use of filters and validators, depending on the semantics of the data extracted. New filters and validators can also be developed and used, depending on semantics of the data extracted as well as other criteria that may apply to a system or another.
- **Media Downloader.** Download all content (documents such as PDF, DOC, XLS, etc., media files such as movies, audio files, images, etc.) that accompanies the records and store it locally and/or upload on FTP.
- **Email Notifications.** Receive email notifications with execution reports in all cases: success/error/failure.

Harvesting Implementation Phases

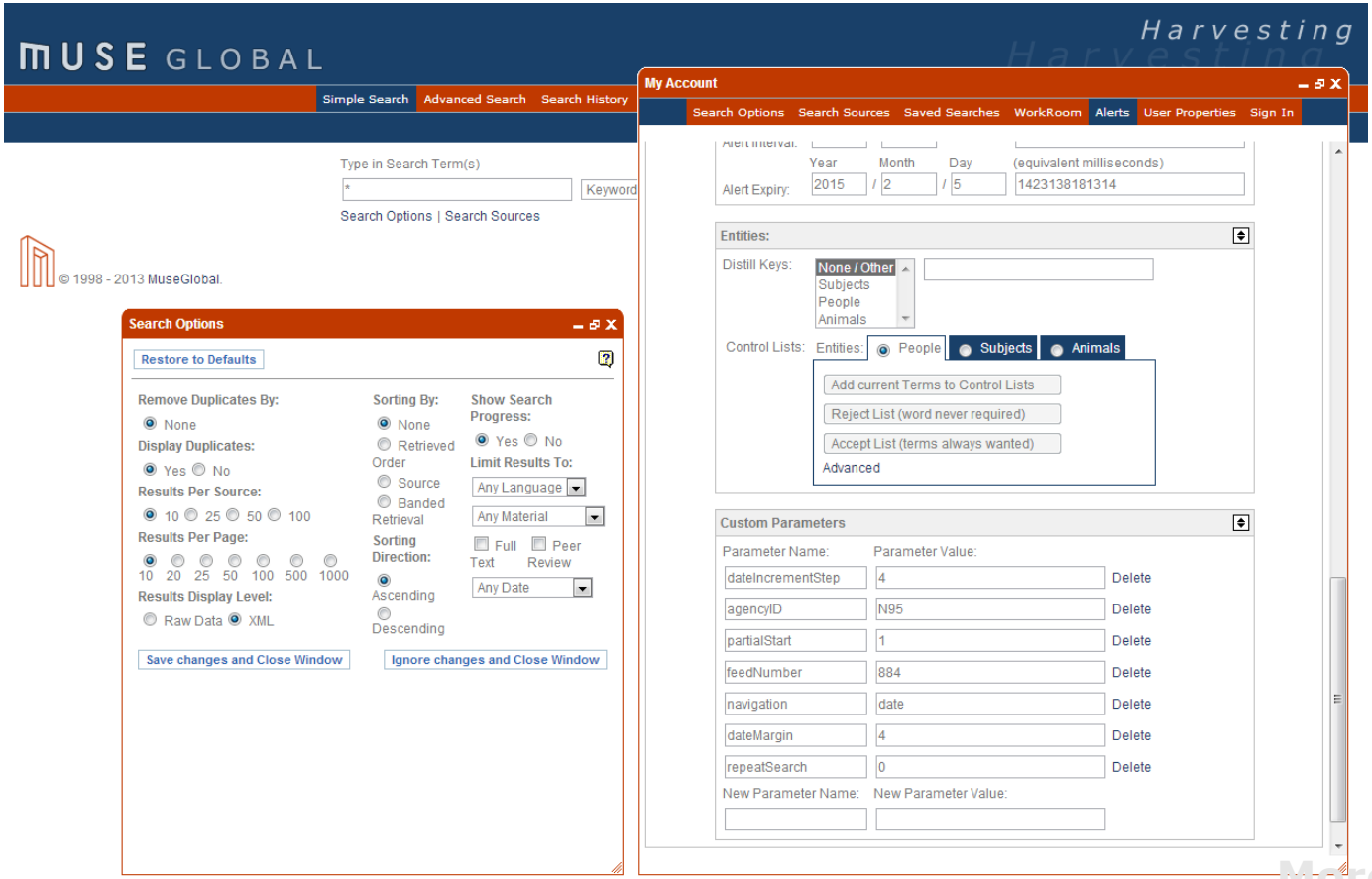
- **Planning time and gathering the pre-requisites**
  - Contact vendors for access to APIs;
  - Obtain access details;
- **Implementation time**
  - Create the necessary MuseKnowledge™ Source Packages for the sources to be harvested;
  - Implement the requested business logic e.g. create the necessary ICE Scripts, Ant Scripts, MuseKnowledge™ Control Center tasks, etc.;
- **Tests**
  - Corrections can be made during tests;
- **Acceptance**
  - The customer must accept the harvesting system before going further;
- **Historical phase.** This is the extraction process of data up to the current date.
- **Monitoring.** The extraction process is monitored and necessary corrections are applied.
- **Incremental phase.** This is the regular extraction process which harvests the deltas.
- **Monitoring.** Monitoring procedures are followed to make sure the harvesting process is smooth.

Key Components

- **MuseKnowledge™ Harvesting Application -** MuseKnowledge™ Application specially created and preconfigured for harvesting process; Multiple records display options are available for better management of extracted fields: XML, RAW, Atom.
- **MuseKnowledge™ Record Tracking System -** Track the statuses of the processed records; The MuseKnowledge™ RTS Console connects to the MuseKnowledge™ RTS DB and provides a basic HTML interface for the processed records.
- **MuseKnowledge™ Control Center -** Execute tasks at any given moment; The system functionality is given by a set of dynamically loaded tasks, which are driven by events generated by the core system or by the other tasks running into system; Every task generates its own events to inform about a certain status or exceptions, which can arise during execution.
- **MuseKnowledge™ Console for Application Administration (MCAA) -** Manage the MuseKnowledge™ Source Packages from inside MuseKnowledge™ Harvesting Applications; Configure many areas of the Muse System such as the administration users and their rights; Problem reporting to MuseGlobal Support.
- **MuseKnowledge™ Alerts -** Searches saved that are re-run automatically by Muse (more exactly MuseKnowledge™ Control Center will run the appropriate Alerts script).
- **ICE Scripts –** Scripts that are used to define the business logic of the harvesting process.

MuseKnowledge™ Harvesting Application

- Individual instance that runs inside MuseKnowledge™ at a customer site or at a group of sites to provide harvesting of resources. As the harvesting process is complex, there are MuseKnowledge™ Application templates specially created and preconfigured for harvesting processes.

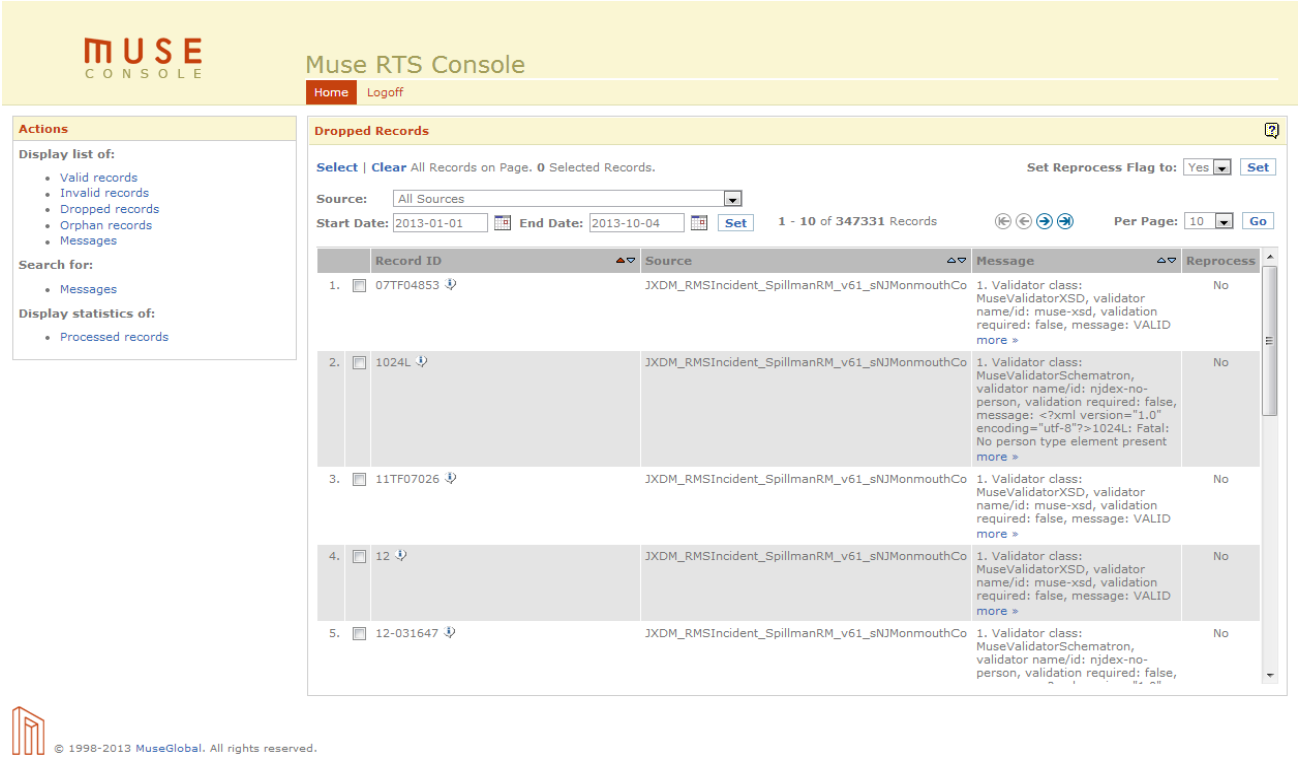


MuseKnowledge™ Harvesting Application

- The application comes with an extended Alerts editing interface for configuring each aspect of a MuseKnowledge™ Alert:
  - Basic parameters such as name, query, query type and comments associated with the Alert;
  - Searched Sources;
  - Expression Sources. This is needed for constructing a hierarchy of MuseKnowledge™ Alerts.
  - Search Options such as: Enrich Results By, Remove Duplicates By, Dedupe Mode, Dedupe Mix Mode, Sorting By, Sorting Mode, Sorting Direction, Results Per Source, Results Per Page.
  - Result Set details;
  - Alert Options such as the Alert Interval and Expiry;
  - Entities. Change the Content Mining control lists for various entities such as: People, Subjects and Animals. One can set here both accept and reject control lists for each of these three entities;
  - Custom Parameters. One can define custom parameters needed to run the Alert script through MuseKnowledge™ Control Center;
  - Alerts Parents. This is part of the Alerts inheritance mechanism for defining parents for the current Alert.

MuseKnowledge™ Record Tracking System

- Track the statuses of the records processed by the system from the moment these records enter the system either through a search connector or imported from 3rd party databases to the moment these records are exported by the system in one of the various outputs and the recipient acknowledges the correct input.

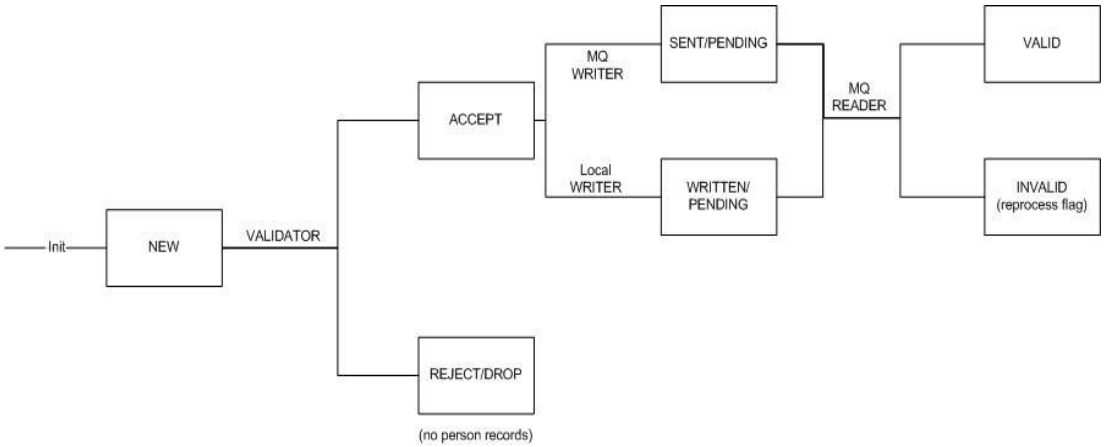


MUSE

More

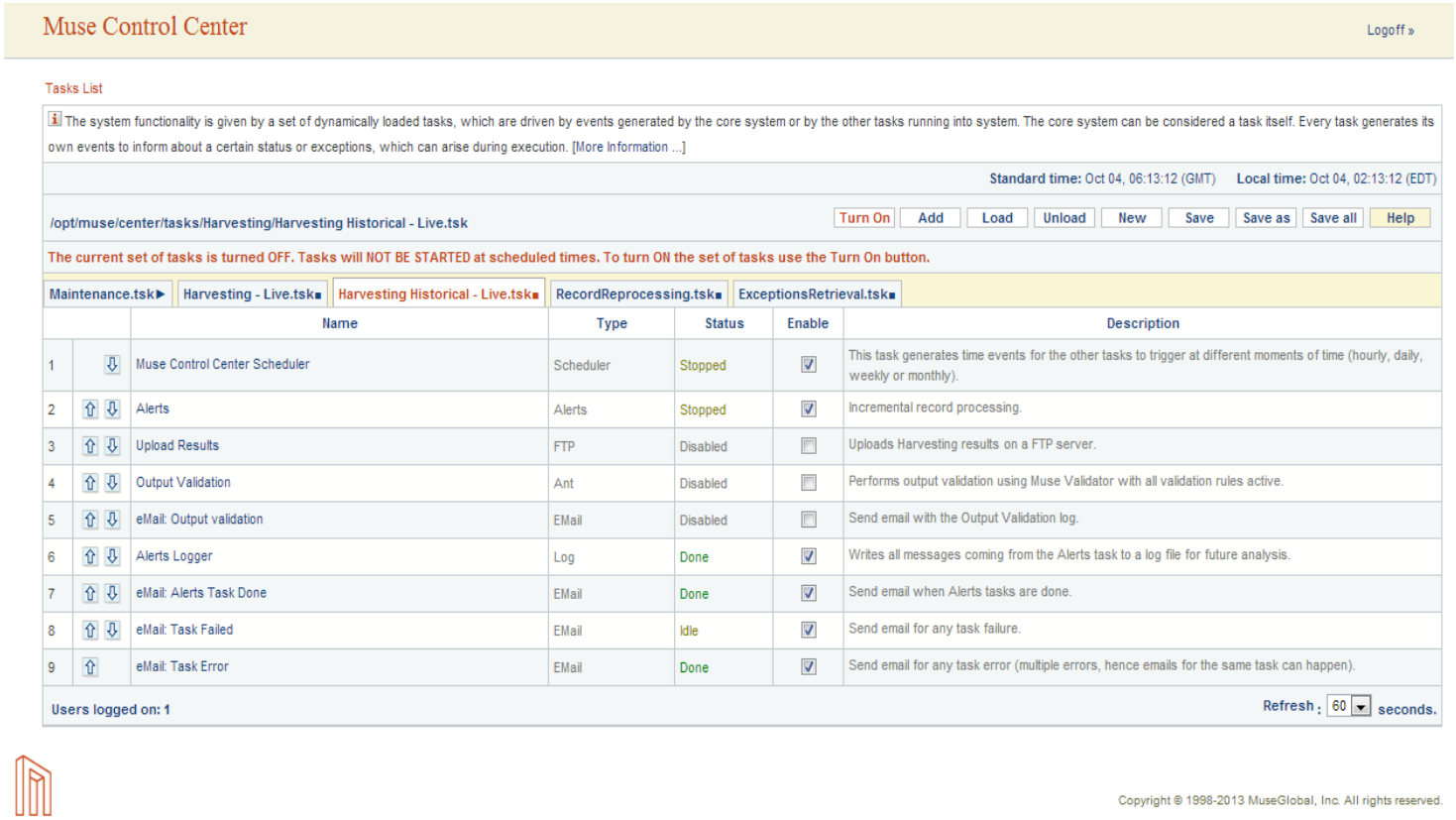
MuseKnowledge™ Record Tracking System

- Record Tracking System statuses:
  - record has entered the system;
  - record has been validated by the system: valid/invalid;
  - record has been exported by the system;
  - accepted or rejected by target system (ex. via Queue Manager such as WebSphere MQ).



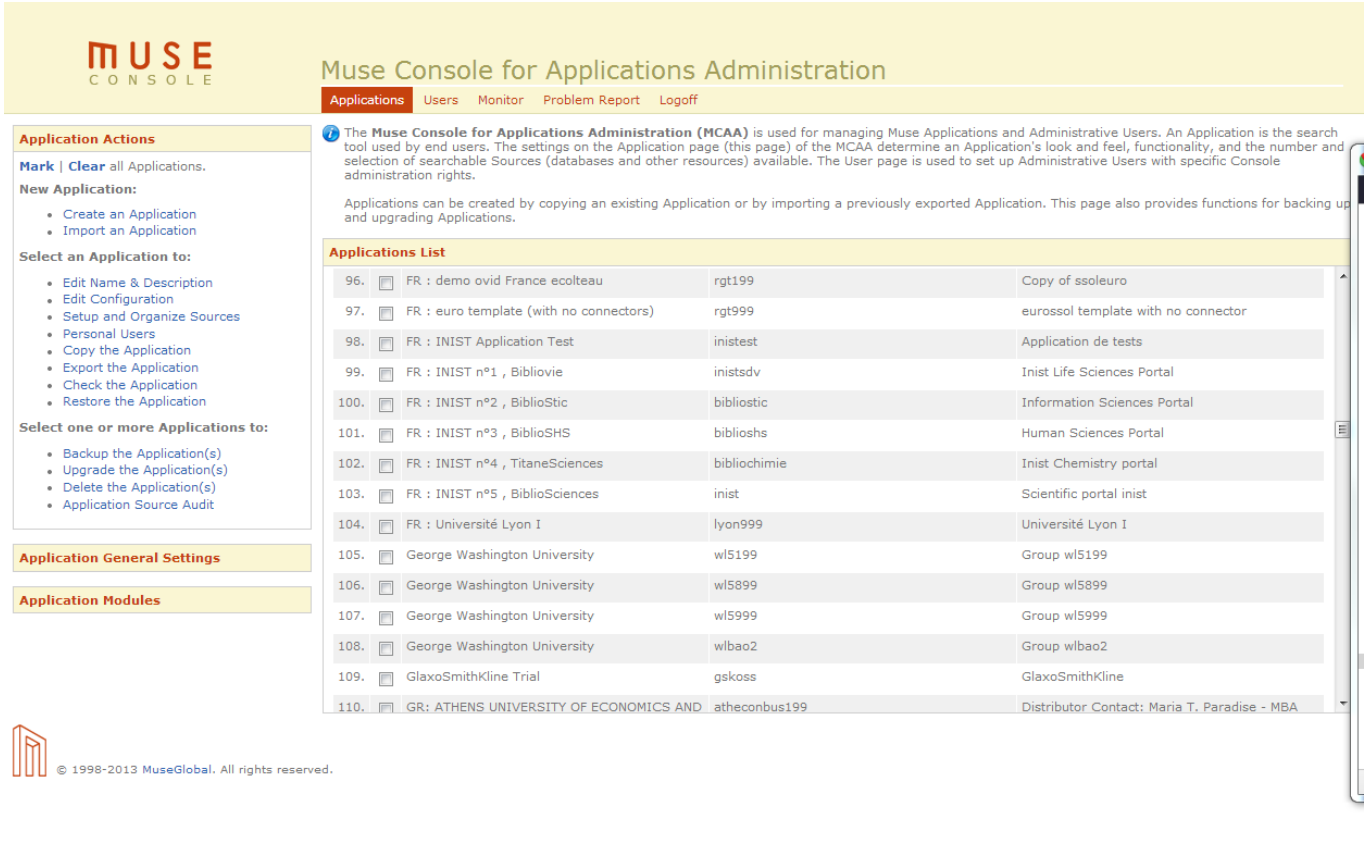
MuseKnowledge™ Control Center

- Used to execute tasks at any given moment. The tasks lists can be used for various purposes, starting from regularly updating the MuseKnowledge™ Source Packages to sending emails to users or even generate usage statistics of the system they run on.
- Runs saved queries (MuseKnowledge™ Alerts) on various sources at any moment, without any intervention.



MuseKnowledge™ Console for Application Administration (MCAA)

- Administration tools used to add, delete and manage access rights to MuseKnowledge™ Applications. By means of administrator consoles there is access to any settings of the system.



Usage Scenarios

Extraction, Transform and Load tools

- **Step 1.** Extract from a variety of distinct protocols, distinct data types, distinct sources, structured or non-structured :
  - CMS, search engines, repositories, database systems;
  - Magazines, news, journal, library archives, books, articles, images, web formats, videos, blogs, real objects;
  - Traditional and online publishers (subscription/premium content);
  - Online content aggregators.
  - Examples of standards: Atom, HTTP/HTML, HTTP/XML, JSON, NCIP, OAI-PMH, RSS 1.0, RSS 2.0, SIP2, SQL, SRU, SRW, Telnet, Z39.50.
- **Step 2.** Transform into a single standard protocol and format: XML, SQL, JSON, etc.
- **Step 3.** Filter, Validate: Syntactic, Semantic.
- **Step 4.** Feed into the target system's bus: queue managers, databases, content repository, etc.
- **Step 5.** Receive transaction feedback, analyze it, store it, etc.

Target: Justice and Law enforcement, Enterprise, Media.

MUSE

More

Usage Scenarios

Muse Meters - Want to know what's hot? And what's not? And what's hotter today than it was yesterday?

- **Step 1.** Extract from a variety of sources, structured or non-structured: RSS feeds, blogs, newsletters, newspapers.
- **Step 2.** Transform into a single standard protocol and format: XML, JSON, etc.
- **Step 3.** Perform content mining to identify relevant terms for the day/hour and present the evolution.
- **Step 4.** Construct the lists of people, teams, products, topics, etc. and feed into the Meter.

Muse Multi-Meters – A variation of Muse Meters.

- **Step 1.** One or more 'Multi-Meter Alerts' are used to download the available records from a set of sources.
- **Step 2.** A second, user-set alert will inherit the properties from one or more 'multi-meter alerts' which will be calling parent alerts from now on. The only difference between this user alert and a normal one is that the user alert will run the search - with the user set query - over the result set created by the 'multi-meter alert'.
- **Step 3.** Perform content mining and construct the list of items to feed into the Multi-Meter.

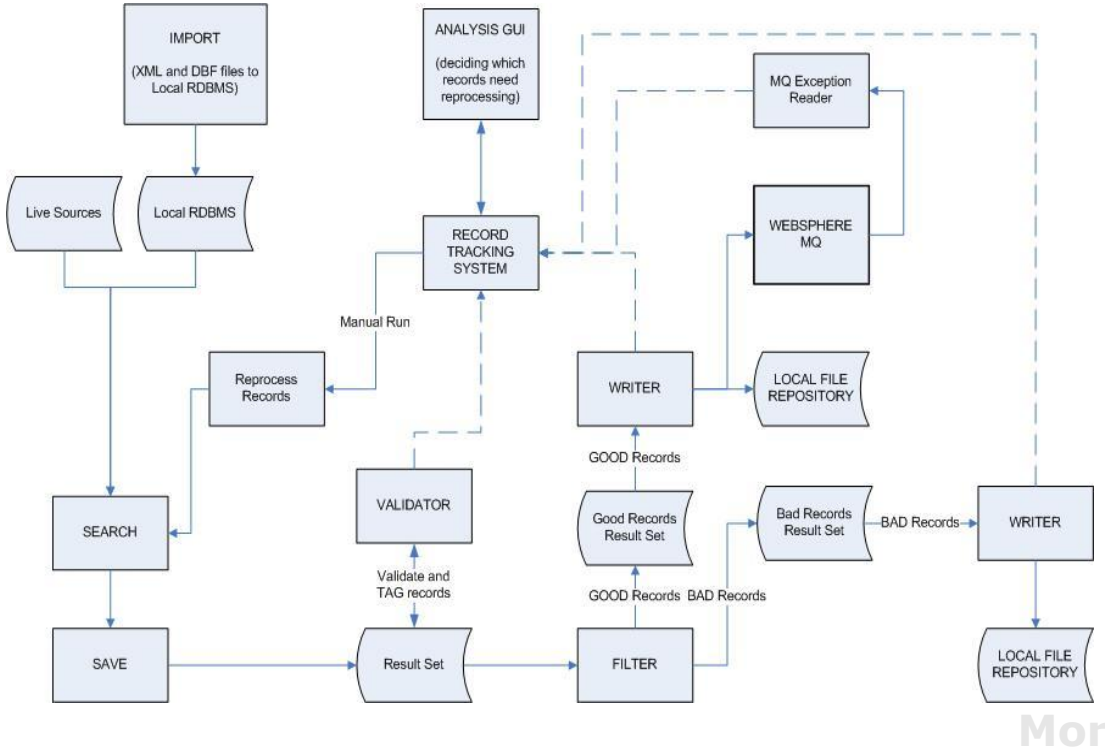
Target: Media, Marketing, News, Sports, Politics.

Usage Scenarios - Harvesting Police Data

Harvesting Police Data

Automatically harvest name and incident information from the police departments' Records Management Systems (RMSs) and uploads the information to a target union database (state, national, etc.) - IBM WebSphere MQ (WMQ) system.

Workflow



Usage Scenarios - Harvesting Police Data

Incremental and Historical Import

Incremental import workflow: extraction, validation, local storing, RTS storing, WMQ exporting, exceptions retrieval, record reprocessing for WMQ invalid records.

Historical import - particular case of the incremental with a few exceptions:

- It runs at a higher frequency to import all the records until the current day;
- Records extracted using a historical task are not sent to WMQ, but sent rather using other methods: disk, tape, FTP;
- Store in RTS for tracking purposes.

Validating the Result Set

- The validation is handled by the Muse Validator module;
- The validation can be done with multiple validation schemas;
- Each record from the Result Set is appended the result of the validation;
- Currently there are only 2 implementations: XSD and Schematron validation;

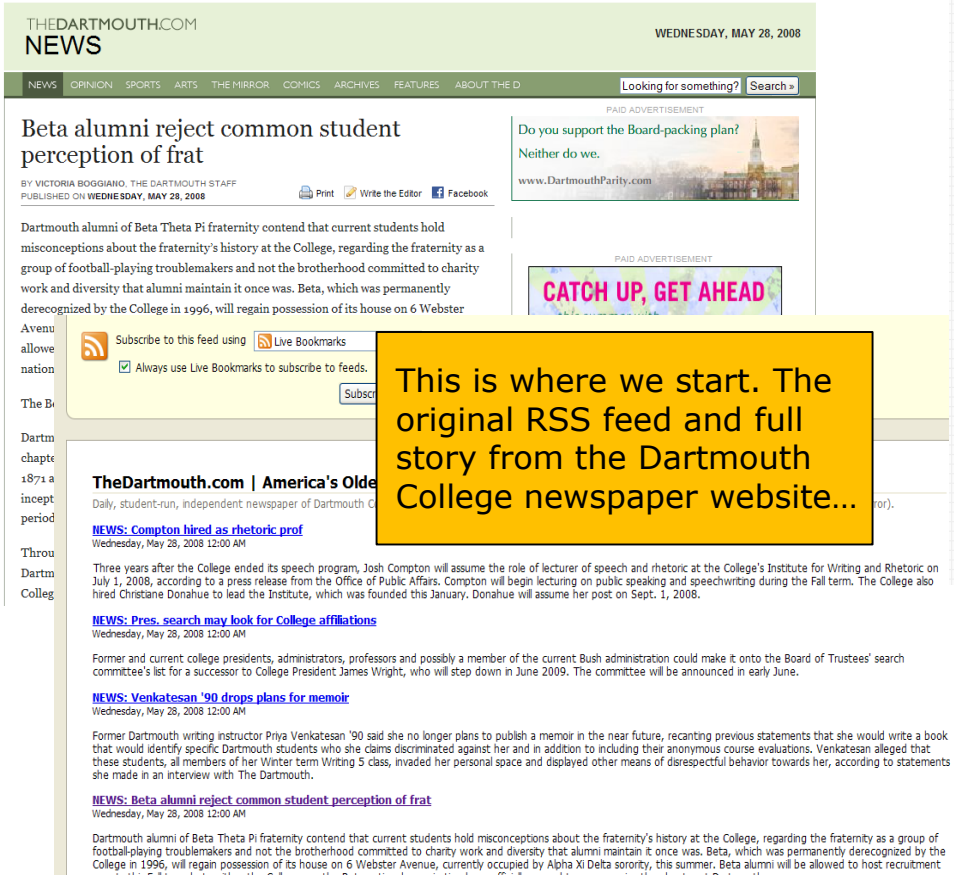
Usage Scenarios - Harvesting Police Data

Splitting and Writing the Result Set

- **Split module:**
  - filters the records received from the records stream and pass them to different modules;
  - adds in the RTS entries corresponding to the validation results;
  - splits the Result Set into 4 result sets:
    - a Result Set which contains all the VALID records;
    - a Result Set containing all the DROPPED records (records for which the validation output was ERRORS for Schematron validation);
    - a Result Set containing all the M_INVALID records (records for which the validation output was ERRORS for XSD validation);
    - a Result Set containing all other records. Normally this is empty.
- **Writer module:**
  - writes the records after split into Result Sets;
  - has several implementations:
    - for the valid records uses the local writer implementation and the WMQ writer implementation, with RTS;
    - for the dropped records uses only the local writer, no RTS;
    - for the XSD invalid records uses only the local writer, with RTS;

Usage Scenarios - Harvesting Stories

Harvests stories from 275 college newspapers every hour. Removes duplicates, normalizes the records, and delivers the new (deltas) and updated stories via an FTP feed.


THEDARTMOUTH.COM
NEWS

WEDNESDAY, MAY 28, 2008

NEWS OPINION SPORTS ARTS THE MIRROR COMICS ARCHIVES FEATURES ABOUT THE D

Looking for something? Search »

Beta alumni reject common student perception of frat

BY VICTORIA BOGGIANO, THE DARTMOUTH STAFF
PUBLISHED ON WEDNESDAY, MAY 28, 2008

Print Write the Editor Facebook

Dartmouth alumni of Beta Theta Pi fraternity contend that current students hold misconceptions about the fraternity's history at the College, regarding the fraternity as a group of football-playing troublemakers and not the brotherhood committed to charity work and diversity that alumni maintain it once was. Beta, which was permanently derecognized by the College in 1996, will regain possession of its house on 6 Webster

PAID ADVERTISEMENT

Do you support the Board-packing plan?
Neither do we.
www.DartmouthParity.com

PAID ADVERTISEMENT

CATCH UP, GET AHEAD

Subscribe to this feed using Live Bookmarks

Always use Live Bookmarks to subscribe to feeds.

Subsc

TheDartmouth.com | America's Olde

Daily, student-run, independent newspaper of Dartmouth C

NEWS: Compton hired as rhetoric prof
Wednesday, May 28, 2008 12:00 AM

Three years after the College ended its speech program, Josh Compton will assume the role of lecturer of speech and rhetoric at the College's Institute for Writing and Rhetoric on July 1, 2008, according to a press release from the Office of Public Affairs. Compton will begin lecturing on public speaking and speechwriting during the Fall term. The College also hired Christiane Donahue to lead the Institute, which was founded this January. Donahue will assume her post on Sept. 1, 2008.

NEWS: Pres. search may look for College affiliations
Wednesday, May 28, 2008 12:00 AM

Former and current college presidents, administrators, professors and possibly a member of the current Bush administration could make it onto the Board of Trustees' search committee's list for a successor to College President James Wright, who will step down in June 2009. The committee will be announced in early June.

NEWS: Venkatesan '90 drops plans for memoir
Wednesday, May 28, 2008 12:00 AM

Former Dartmouth writing instructor Priya Venkatesan '90 said she no longer plans to publish a memoir in the near future, recanting previous statements that she would write a book that would identify specific Dartmouth students who she claims discriminated against her and in addition to including their anonymous course evaluations. Venkatesan alleged that these students, all members of her Winter term Writing 5 class, invaded her personal space and displayed other means of disrespectful behavior towards her, according to statements she made in an interview with The Dartmouth.

NEWS: Beta alumni reject common student perception of frat
Wednesday, May 28, 2008 12:00 AM

Dartmouth alumni of Beta Theta Pi fraternity contend that current students hold misconceptions about the fraternity's history at the College, regarding the fraternity as a group of football-playing troublemakers and not the brotherhood committed to charity work and diversity that alumni maintain it once was. Beta, which was permanently derecognized by the College in 1996, will regain possession of its house on 6 Webster Avenue, currently occupied by Alpha Xi Delta sorority, this summer. Beta alumni will be allowed to host recruitment

This is where we start. The original RSS feed and full story from the Dartmouth College newspaper website…

URL: http://thedartmouth.com/2008/05/28/news/beta/
IDENTIFIER: http://thedartmouth.com/2008/05/28/news/beta/
ABSTRACT: Dartmouth alumni of Beta Theta Pi fraternity contend that current students hold misconceptions about the fraternity's history at the College, regarding the fraternity as a group of football-playing troublemakers and not the brotherhood committed to charity work and diversity that alumni maintain it once was. Beta, which was permanently derecognized by the College in 1996, will regain possession of its house on 6 Webster Avenue, currently occupied by Alpha Xi Delta sorority, this summer. Beta alumni will be allowed to host recruitment events this Fall term but neither the College nor the Beta national organization have officially agreed to re-recognize the chapter at Dartmouth.
DATE: Wednesday, May 28, 2008
BODY: Dartmouth alumni of Beta Theta Pi fraternity contend that current students hold misconceptions about the fraternity's history at the College, regarding the fraternity as a group of football-playing troublemakers and not the brotherhood committed to charity work and diversity that alumni maintain it once was. Beta, which was permanently derecognized by the College in 1996, will regain possession of its house on 6 Webster Avenue, currently occupied by Alpha Xi Delta sorority, this summer. Beta alumni will be allowed to host recruitment events this Fall term but neither the College nor the Beta national organization have officially agreed to re-recognize the chapter at Dartmouth. The Beginning: Beta comes to Dartmouth Dartmouth's chapter of Beta began as Sigma Delta Pi fraternity, which established a chapter at Dartmouth in 1858. The organization was renamed the Vitruvian Society in 1871 and eventually became the Alpha Omega Chapter of Beta Theta Pi in 1889. From its inception, Beta maintained its affiliation with the national organization, except for a brief period from 1961 to 1962. Throughout the 1930s and 1940s, Beta's membership included athletes, members of The Dartmouth and the Jack-O-Lantern, performers in the marching band and members of the College's glee club, according to Beta's chapter book from 1941. The Glory Days: Beta turns 100 During the winter of 1953, Dartmouth's chapter of Beta and eight other Beta chapters in New England passed a resolution that condemned racial discrimination within the national fraternity and demanded that each chapter lift its membership restraints, according to the Manchester Union Leader from December of that year. Further disagreement over racial discrimination in 1961 led Dartmouth's Beta chapter to disassociate from the national organization for two years, Dmitri Gerakaris '69, a member of Dartmouth Beta Board of Trustees, said in an interview. Several other national fraternities at the College went local at this time for similar reasons, according to Deb Carney, director of coed, fraternity and sorority administration. "During the late '50s and early '60s, trustees made a statement that all of our fraternities must be open and inclusive to all," Carney said. "And some of our national frats had discrimination clauses — those fraternities either changed or went local." Dartmouth's Beta chapter, geographically isolated from the fraternity's national headquarters, had minimal ties to the national organization after reaffiliating in 1962, according to Gerakaris, who was president of both Beta and the InterFraternity Council during his senior year at the College. The End is Near: Beta in the 1980s and 1990s Following coeducation in 1972, women would often frequent Beta for its parties and comfortable atmosphere, Beta alumnus and board … grew, women felt

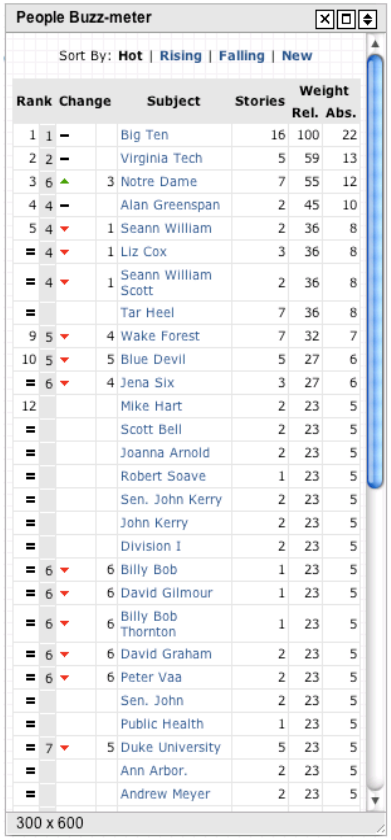
… and this is where we end up. A normalized, structured record, ready for export.

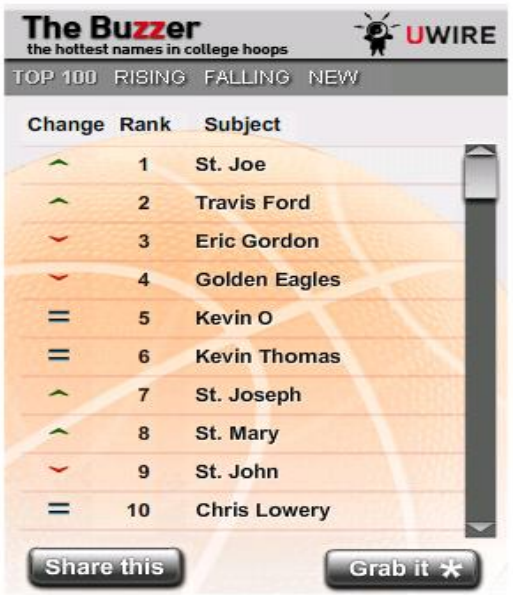
MUSE

Usage Scenarios - BuzzGauge

Buzz-meter: A set of Muse functions which gather content from a wide range of Sources, analyze it for Entities, generate secondary content about it, and distribute that "buzz" for widgets, newsletters, mashups, dashboards, etc.

- Display lists of people, teams, products, topics
- Show movement, rank, importance, Stories, etc.
- Skin the widget to you look & feel
- Flash® and JavaScript widgets supported
- Direct secure connection to Muse server
- Various lists and data elements
- Live links from list items to results list and original records

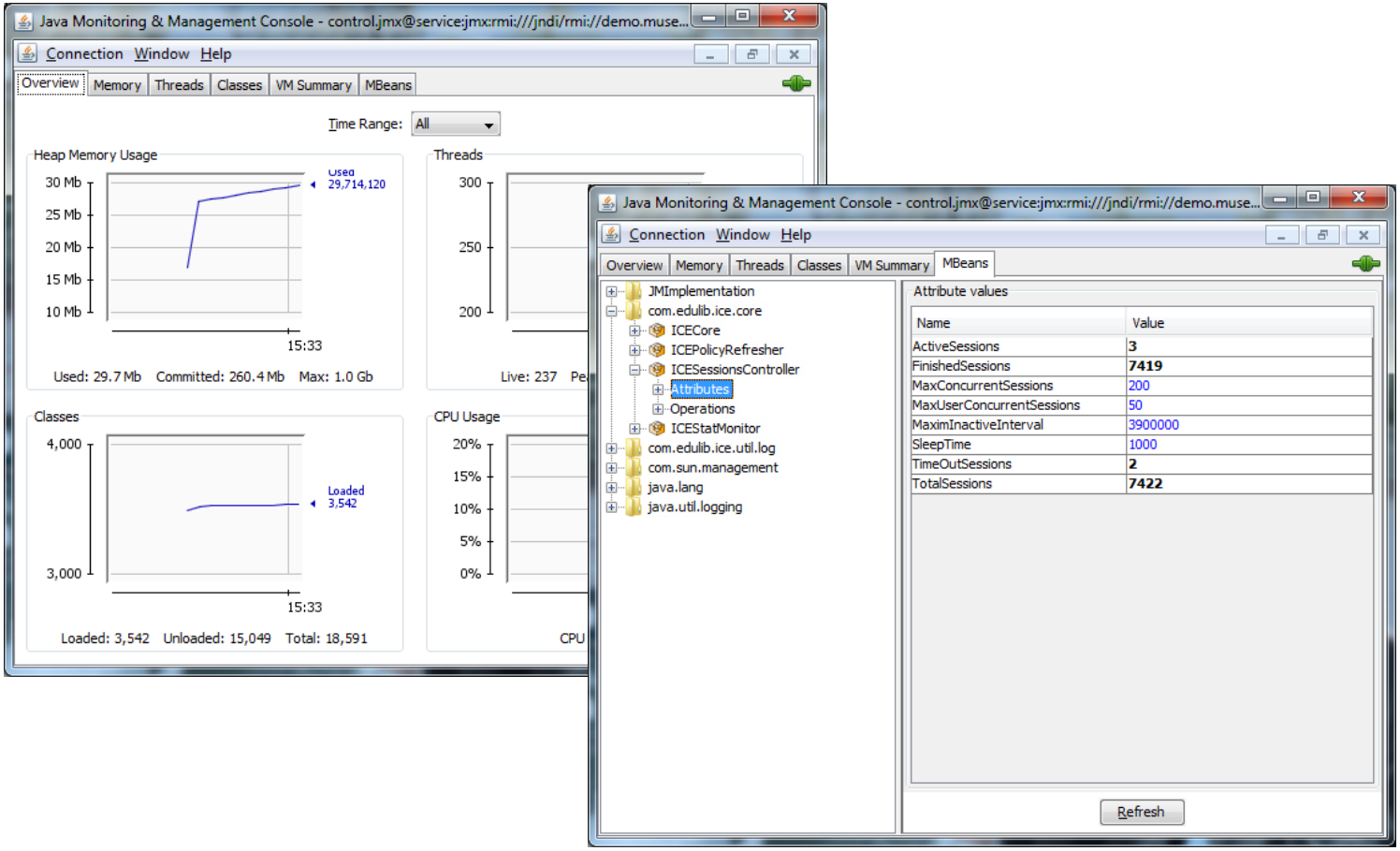




MUSE

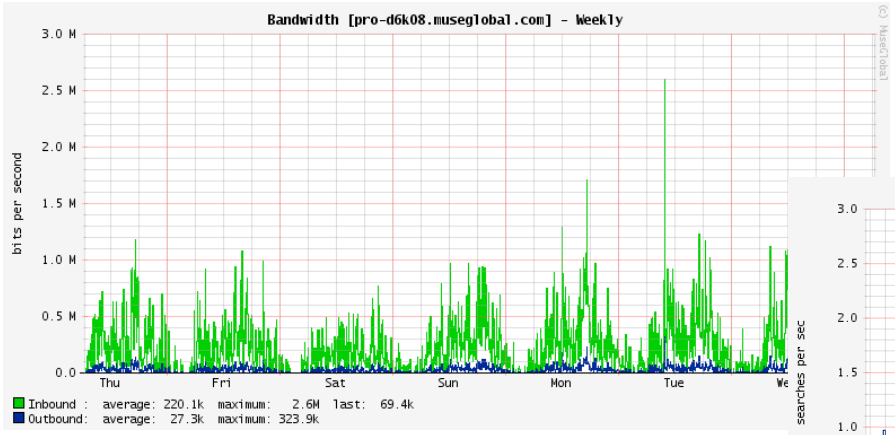
Monitoring the Harvesting

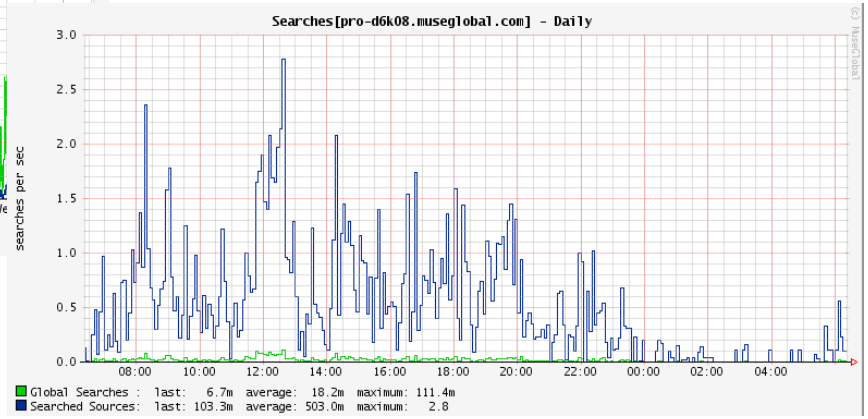
Advanced real time monitoring of Muse servers through JMX

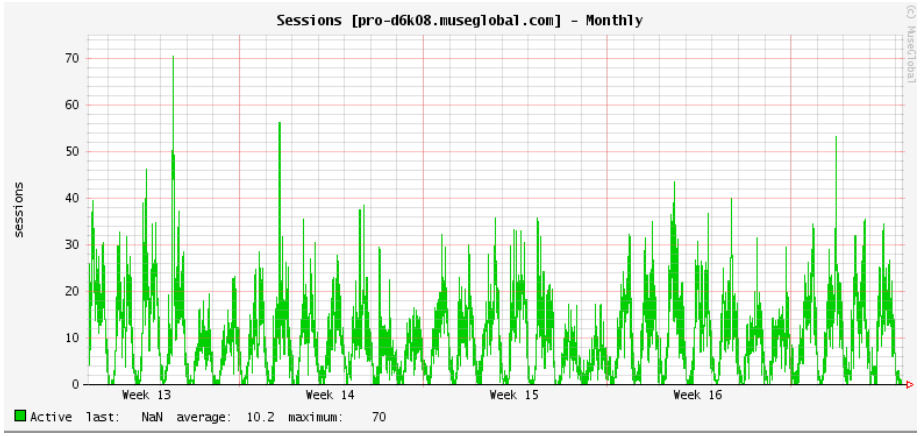


Monitoring the Harvesting

Historical JMX graphs with RRD Grapher







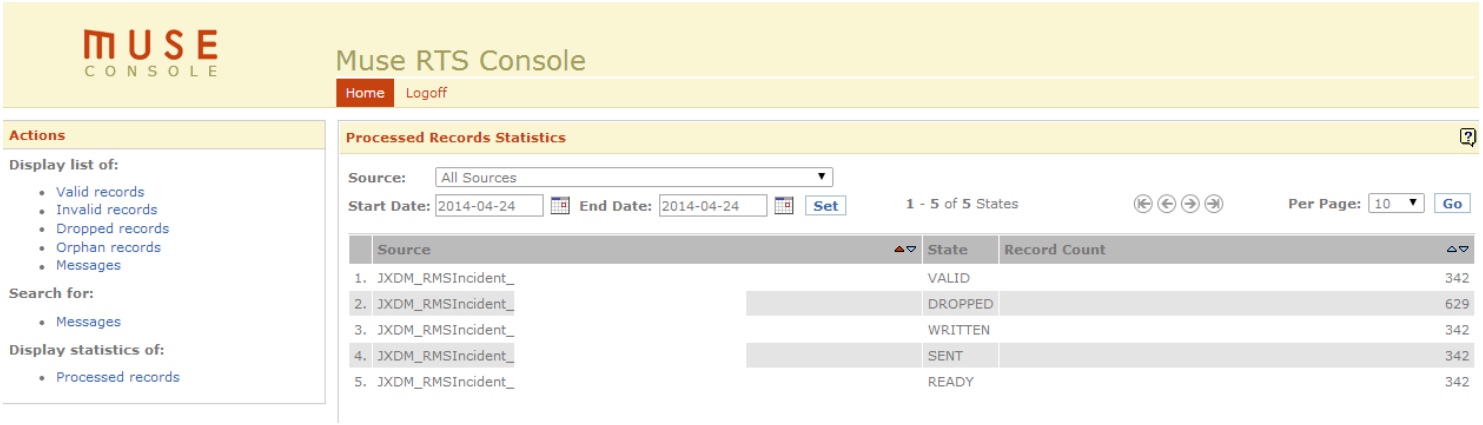
Monitoring the Harvesting

Get email notifications for success/error/failure executions

- MuseKnowledge™ Control Center email tasks are used for this purpose;
- Email tasks can be configured to be sent on completion, error or failure events;
- Log file(s) with execution reports can be attached to the emails as they are or archived;

Log files - Muse reports various types of statistical and debugging information in log files.

MuseKnowledge™ Record Tracking System statistics



Muse Documentation

- Muse Console for Application Administration.pdf
- Muse Control Center.pdf
- Muse Alerts.pdf
- Record Tracking System.pdf

MUSE

Smart Connector Technology for Harvesting