Harvester Automation for Metadata Search Web Application

Part One: Methodologies by Terry Yuan, SIParCS Intern
Mentors: Nathan Hook, Saquib Aziz Khan, Christy Grant, Eric Nienhouse

Sama Manalai / Terry Yuan
SIParCS 2021: Project 7
Background

A Search Tool Built Under NCAR Made For NCAR
Background

GitHub Repository: XML Metadata

Harvester Service Collects
Apache Solr Indexes
Search Service Queries

Metadata Search Web Application
Motivation

How to Improve Harvester?

Fast, Accurate Metadata Accessibility
- For Search Service + User
- For Developers

Easier Future Work

Methodologies

Sama Manalai / Terry Yuan
SIParCS 2021: Project 7
Methodologies

- Agile Scrum
- SOLID Principles
- Pair Programming
- Layered Architecture
- Refactoring Code

Structure:
- Method
- Implementation
Methodologies: Agile Scrum

The Plan:

Weekly Sprints: Main Objective: Harvester

- Plan, Refine, Reflect
- Immediate Feedback from the Team
Methodologies: Agile Scrum

Sprint Planning

<table>
<thead>
<tr>
<th>Summary</th>
<th>Points</th>
<th>Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ignore Solr Core files during Commit</td>
<td>3</td>
<td>Update .gitignore in solr</td>
</tr>
<tr>
<td>Clean Up Extra XML Files in Harvester Project</td>
<td>N/A</td>
<td>Remove placeholder ISO XML files</td>
</tr>
<tr>
<td>Update Readme File</td>
<td>0.5</td>
<td>Reread and amend README files</td>
</tr>
<tr>
<td>Setup External Instance Properties Files for Development</td>
<td>3</td>
<td>External application.properties file</td>
</tr>
</tbody>
</table>

Example: Week 1 Sprint
Methodologies: Agile Scrum

Refine Backlog

Know the Problems
- Past
- Present
- Future

<table>
<thead>
<tr>
<th>Backlog</th>
<th>25 issues</th>
<th>Create sprint</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSETSP-129</td>
<td>Treat Git Repository as Authoritative Source</td>
<td>Automated Infrastructure</td>
<td>5</td>
</tr>
<tr>
<td>DSETSP-81</td>
<td>Search Provides Suggestions</td>
<td>Search Application</td>
<td></td>
</tr>
<tr>
<td>DSETSP-112</td>
<td>Unable to Delete a Record from Solr</td>
<td>Harvest Metadata</td>
<td></td>
</tr>
<tr>
<td>DSETSP-57</td>
<td>Search Index Isn't Updated Fast Enough After a Metadata Update</td>
<td>Automated Infrastructure</td>
<td></td>
</tr>
<tr>
<td>DSETSP-55</td>
<td>Queue Storage of Dataset Metadata into Solr</td>
<td>Automated Infrastructure</td>
<td></td>
</tr>
<tr>
<td>DSETSP-120</td>
<td>Provide a Site Map for All Search Records</td>
<td>Search Application</td>
<td></td>
</tr>
<tr>
<td>DSETSP-121</td>
<td>Validate ISO Records Contain Minimum Metadata</td>
<td>Search Application</td>
<td></td>
</tr>
<tr>
<td>DSETSP-105</td>
<td>Search on Lab or Organization</td>
<td>Search Application</td>
<td></td>
</tr>
<tr>
<td>DSETSP-122</td>
<td>Harvest Errors are Emailed to WAF Contact</td>
<td>Search Application</td>
<td></td>
</tr>
<tr>
<td>DSETSP-53</td>
<td>Unable to Search or Facet on Resource Type</td>
<td>Search Application</td>
<td></td>
</tr>
<tr>
<td>DSETSP-110</td>
<td>Non-GCMD Keywords are in our Solr Keywords Field</td>
<td>Search Application</td>
<td></td>
</tr>
<tr>
<td>DSETSP-98</td>
<td>Harvester Adds Github Xml Url Location to Solr Record</td>
<td>Harvest Metadata</td>
<td></td>
</tr>
</tbody>
</table>
Methodologies: Agile Scrum

Reflect Retrospective

<table>
<thead>
<tr>
<th>Happy Face</th>
<th>Frowny Face</th>
</tr>
</thead>
</table>
| Stories completed well before the end of sprint +1 +1  
Moving harvester forward/good direction +1  
Setting up asynchronous and/or cron jobs via spring Annotation is surprisingly quick and easy+1 +1  
** Running ahead w/ backlog review +1 +1 | Lots to learn when using frameworks. +1  
Really mispointed scheduling story. |

<table>
<thead>
<tr>
<th>Ideas</th>
<th>Flowers</th>
</tr>
</thead>
</table>
| Thinking about 2 separate presentations  
** More asynchronous stories +1  
*** Working on stories for search application +1  
*** Jamboard meeting to kickoff presentation/poster ideas. +1 | Sama and Terry working together. +1 +1 +1  
Sama and Terry looking at stories in the backlog. |
Methodologies: Agile Scrum

Focus on the Main Objective: Harvester Automation

Short Sprints = Progress!

- Fast Deadline
- Quick Results
- Small Steps
Objective: Make Software More Maintainable

- One object has one purpose
  - Single Responsibility Principle

- Reduce knowledge = Be modular
  - Interface Segregation Principle

Not every principle was followed, and that's okay
One object has one purpose
Example: Evolution of a Class's Responsibilities

Because if you use "AND", you know it's too much
Methodologies: SOLID Principles

Reduce knowledge = Be modular
Example: Manual Update from GitHub

What does it take to do one thing?

UpdateMediator ➔ GitUpdates ➔ GitRepository ➔ GitFileMediator ➔ HarvesterService

We are just about to start reharvesting...

Update | Clean and Reharvest | Cancel

Sama Manalai / Terry Yuan
SIParCS 2021: Project 7
Methodologies: Layered Architecture

Example: Automatic Update from GitHub (+ Restaurant Parallel)

Presentation (Customer)    Application/Domain (Waiter)    Persistence (Kitchen)

Update Request (START)  →  Repository Update Check  →  Repository Database

Search Results (END)
Methodologies: Layered Architecture

Example: Same but Complicated (+ Restaurant Parallel)

Presentation (Customer)

Application /Domain (Waiter)

Persistence (Kitchen)

GitRepositoryController

UpdateMediator

GitUpdates

GitFileMediator

HarvesterService

GitRepository

Sama Manalai / Terry Yuan
SIParCS 2021: Project 7
Methodologies: Layered Architecture

Objective: Implementing SOLID "theory"

- One Thing, One Job
- Communicate One Layer At A Time
- I Know My Code
Methodologies: Pair Programming

Double the Brain Cells, Double the Vision

Literally just teamwork...

Variable Communication
Methodologies: Pair Programming

- Interactive Conversations
- Double Engagement
- Teamwork
- Awkward Silences
- Network Issues When Virtual
- Not Rated E For Everyone
Methodologies: Refactoring Code

Organize

Shift

Rewrite

Balance Between:

Readability

Optimization
Methodologies: Refactoring Code

- Boost Confidence, Settle the Bare Necessities
  - Unit Tests
- Remove repeats, Simplify code, Prioritize Readability
  - Iterative Cycles

Sama Manalai / Terry Yuan
SIParCS 2021: Project 7
What is your project?

Recommendations:

- Modular?
- Maintainable?
- Team Size > 1?
Thank You For Listening

Thank You To...

SIParCS Mentors: Nathan Hook, Saquib Aziz Khan, Eric Nienhouse, Christy Grant

SIParCS Program Leads AJ Lauer, Virginia Do, Jerry Cyccone, Max Cordes Galbraith

...and everyone else for making this program happen
Shift to Technologies

Harvester Automation for Metadata Search Web Application

NEXT UP!

Part Two: Technologies
by Sama Manalai, SIParCS Intern
References

- https://www.atlassian.com/agile/scrum
- https://medium.com/mindorks/solid-principles-explained-with-examples-79d1ce114ace
- https://hackernoon.com/solid-principles-simple-and-easy-explanation-f57d86c47a7f
- https://blog.ndepend.com/layers-architecture-solid-approach/
- https://dzone.com/articles/layers-architecture-is-good
- http://fewagainstmany.com/blog/introduction-to-layers-architecture-part-one
- https://www.agilealliance.org/glossary/pairing/
- https://lvivity.com/what-is-code-refactoring
Digital Media

- https://www.hiclipart.com/free-transparent-background-png-clipart-dphnr - GitHub logo
- https://webiconspng.com/icon/5129 - Binoculars icon
- https://www.flaticon.com/free-icon/thumbs-up-hand-symbol_25423 - Thumbs up icon
- https://pluspng.com/png-82868.html - Test/Checklist icon
- https://www.onlinewebfonts.com/ - For all other icons