Metadata on My Mind:
Automating Troubleshooting to Increase Discovery of Collections

Juliana Nykolaiszyn
& Gautham Ponnaganti

Introduction

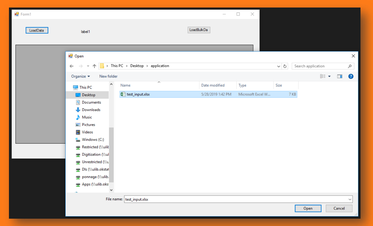
Established in 2014, SHAREOK (www.shareok.org) is the joint institutional repository shared by the Oklahoma State University Libraries, the University of Oklahoma Libraries, and the University of Central Oklahoma Library. It serves as the home for the intellectual output of these communities and includes digital dissertations, faculty publications, digital special collections, open access publications, open educational resources and more. Each institution is responsible for ingesting content into their communities. This poster focuses on metadata remediation efforts with respect to Oklahoma State University's content housed in SHAREOK.

SHAREOK at OSU

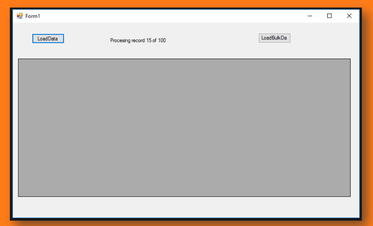
Oklahoma State University's largest SHAREOK collection includes electronic theses and dissertations (ETDs), upwards of 24,000+ submissions. Today, born digital theses and dissertations are loaded upon receipt from ProQuest/UMI. The collection is also comprised of scanned theses and dissertations from the early years of the University through approximately 2004. This scanning project was undertaken at a time when in-house metadata expertise was limited, and employed the help of student workers to complete the project. In addition to ETDs, OSU also ingests a variety of related scholarship, including faculty papers, Agricultural Experiment Station and Cooperative Extension Service materials, a variety of conference proceedings, and undergraduate research, for example.

Library

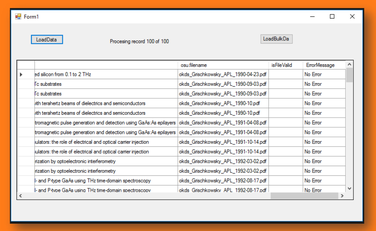
By creating a **desktop application**, we can quickly **validate**, **identify**, and **remediate record** inconsistencies within the IR.



In preparing to run the application, the 'LoadData' button will open a dialog box where the CSV file containing metadata can be selected.



This is the application under execution, processing one entry of metadata at a time. The status text shows the number of handles processed by the application from the CSV file.



The application lists potential problems in the 'ErrorMessage' column. When an entry has an issue, it describes the validation failure.

Take a picture to view the code on GitHub

Metadata Remediation

Upon reviewing OSU metadata ingested in SHAREOK, a number of issues started to arise including spelling errors and inconsistent use of name authorities. A larger problem within the IR also involved a mix of the following:

- Wrong attachments
- No attachments
- Duplicate records

Plan of Attack

We brainstormed ways to automate identification of errors involving record duplication and attachment issues within the IR. Enlisting the help of an OSU Library developer, he proposed creating a validation tool to help streamline the process.

Validation Tool Creation

The tool was created using .NET framework as a Windows Desktop Application. The tool uses internal libraries to read single or multiple CSV files at once.

How it Works

The tool takes one or more CSV files as an input. The CSV files contain, along with other information, DSpace handles and the name of the attachment associated with it. The application then makes an HTTP request to those handles and looks for the following:

- Is the page available?
- Does the page have an attachment?
- Is the attachment name the same as the one in the input CSV?

The application automatically creates an output file where it describes the status of each handle noting any validation failures.

Results