Metadata Quality Enhancement for Large Digital Collections: Web Browser Automation with Selenium IDE

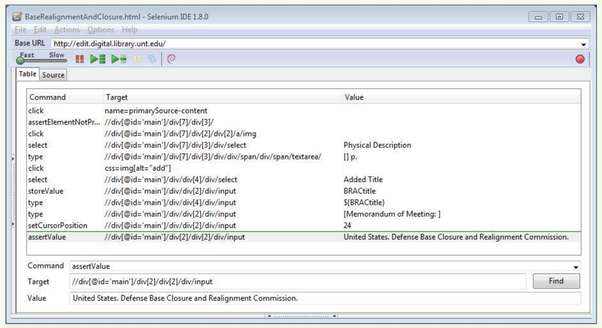
Andrew James Weidner and Daniel Gelaw Alemneh, Ph.D.
University of North Texas Libraries: Digital Projects Unit

Introduction

Creating and maintaining accurate metadata for digital objects is one of the best ways to connect with digital library users and maintain those connections over the long term. Good metadata empowers users to discover exactly what they searched for and to locate relevant resources that they did not expect to find. Metadata quality characteristics for digital libraries depend on many factors: the types of resources the repository offers and the users' needs, which vary across the spectrum of user communities.

The metadata quality issue is particularly acute if there are multiple institutions participating in collaborative digital projects that employ diverse naming schemes for their documents and files. Furthermore, harvesting large sets of documents from open repositories presents a number of challenges for creating accurate descriptive metadata. For example, metadata schema do not always map well, creating disconnects when published in the local repository. In the aforementioned cases, substantial rework is often required to create descriptive data that meets local repository standards.

Automated Metadata Editing: Selenium IDE for Content Management Systems



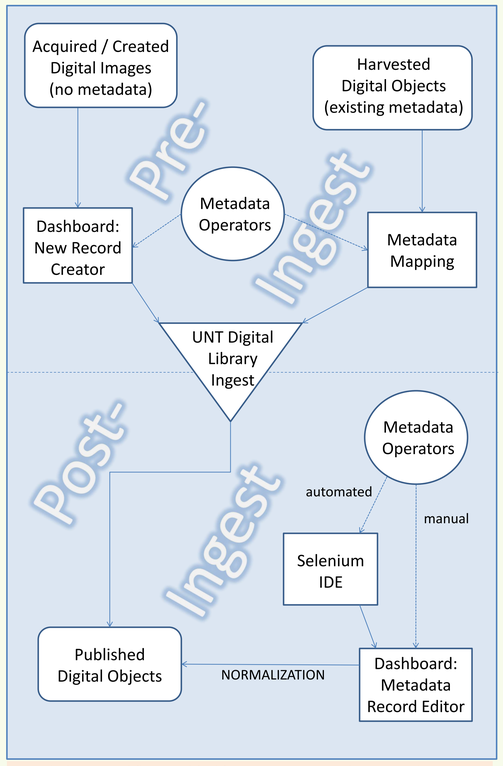
The Selenium IDE Interface Loaded with a Metadata Editing Automation Script.

Developed by the Web design community to simplify the process of testing Web applications, Selenium IDE is an open source Firefox browser add-on which provides an integrated development environment for creating, debugging, and running custom Web browser automation scripts encoded in HTML. Users record new scripts by clicking and typing directly in a browser window and modify existing scripts in either the Selenium IDE interface or a text editor.

Selenium IDE automates a variety of repetitive tasks including: clicking buttons and links, selecting from drop-down menus, and typing text. Automating all or part of the metadata editing process allows human operators to focus on descriptive content, thereby improving accuracy, efficiency and overall enjoyment of the task. Selenium IDE is ideal for situations in which standardized data must be changed across a large set of records and reduces the time required to accomplish batch normalization.

Metadata Quality Tools in UNT's Digital Library Workflow

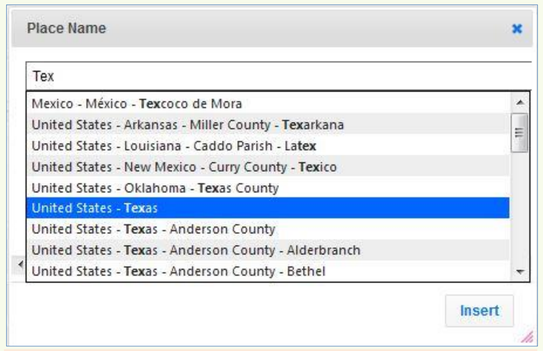
The University of North Texas (UNT) digital libraries group utilizes various tools and mechanisms to ensure metadata consistency and precision across all digital resources. Careful mapping prior to ingest facilitates accurate conversions among various metadata element sets. Crosswalks also facilitate exporting records to other systems throughout the metadata management lifecycle.



Digital Library Workflow Modified for Descriptive Metadata Normalization.

The UNT Digital Library features a Web-based editing interface, called the Dashboard, for metadata operators to use during metadata creation and normalization. Pre-populated controlled vocabulary terms enable operators to easily select standard values via drop-down menus. Auto-suggest for text input fields draws values from controlled vocabularies while simultaneously allowing users to input new values when necessary.

To support these activities, the UNT Libraries recently implemented Selenium IDE as a tool for expediting the editing process for large sets of metadata records in cases when post-ingest metadata normalization enhances recall and precision for its digital objects.



Dashboard Pop-up Window with Auto-Suggested Controlled Vocabulary Values.

Command	Function	Usage
assertValue	Tests the value of an element.	Verifies that a Web page element, such as a drop-down menu or text input field, contains a specific value. If the element contains the specified value, the script continues. If not, the script fails. Use regular expressions to match dynamic text content.
click	Performs a mouse click.	Clicks links, checks boxes, selects radio buttons, etc.
clickAndWait	Performs a mouse click and waits for a new page to load.	Use this command when a clicked link loads a new page. The script waits for the new page to load before executing the next command.
close	Closes the tab or browser window.	The UNT digital libraries group uses this command to enable fully automated batch editing with a suite of identical scripts. After opening a set of records in multiple tabs, the *close* command at the end of each script allows the next script in the suite to work on the next tab in succession until there are either no more scripts in the suite or no more tabs to edit.
keyPress	Performs a specified key press.	Selectively deletes or adds characters in a text input field. Unlike the *type* command, *keyPress* allows users to modify existing text rather than replace it entirely.
select	Selects a value for an element.	Selects a choice from a drop-down menu.
setCursor-Position	Places the cursor at a specified point in a text string.	Use this command with *keyPress* to modify existing text at the appropriate point in a text string.
storeValue	Stores a text string in a variable.	Use this command to copy existing text from an input field. The stored text, contained in a user-defined variable, can then be pasted in another field with the *type* command.
type	Enters text in an input field.	Populates an input field with the specified text, including stored variables. This command overwrites any existing text.

Basic Selenese: Nine Useful Selenium IDE Commands.

Summary

Any institution with a Web-based content management interface can potentially benefit from Selenium IDE's automation capabilities.

Large digital collections present many challenges when producing descriptive metadata. Naming schemes and element definitions can vary widely, requiring substantial rework to meet local repository standards. Successful metadata enhancement strategies involve mechanisms for both pre- and post-ingest metadata normalization such as the built-in tools in UNT's Dashboard editing interface. Web browser automation with Selenium IDE improves operator efficiency and accuracy during the time- and labor-intensive post-ingest data entry process. Selenium IDE enables rapid editing for large sets of published metadata records that might otherwise languish in sub-optimal form because of time and labor constraints. Selenium IDE is a highly recommended addition to the metadata enhancement toolkit for any institution that employs a Web-based content management interface.