Introducing Tabula

The University of Minnesota Libraries Digital Preservation System

Carol Kussmann
University of Minnesota Libraries
United States of America
kussmann@umn.edu

Abstract – The University of Minnesota Libraries journey with preserving digital materials has been a long one. After completing an RFP for a preservation system, and then testing that system for multiple years, we decided it was not the system for us. Over 2021 and 2022, we took our requirements along with the lessons we learned from testing, and began to design our own preservation system. Our main goal with this new system is to preserve the unique materials of the Libraries and to be able to provide access to staff that need copies of preservation files for publication or research requests. This poster highlights the development process of Tabula, our digital preservation system.

Keywords – digital preservation, digital preservation system, libraries, implementation

Conference Topics – From Theory to Practice

I. Introduction

Digital preservation at the University of Minnesota Libraries is managed by the Digital Preservation & Repository Technologies Department (DPRT). DPRT works with a variety of stakeholders across the Libraries and beyond to manage and preserve over 350 TB of materials in all formats.

When we shifted to developing our own digital preservation system, we began by reviewing our requirements and ensuring that we understood the goals and purpose of the preservation system. We focused on five main areas of development utilizing an iterative process: metadata requirements, ingest processes, the hardware/software environment, reporting functionality, and preservation activities.

II. Development Areas

A. *Metadata Requirements*

Existing descriptive metadata schema from multiple sources were studied and crosswalks were developed. The goal was to create a minimal set of descriptive metadata that would assist with the preservation of the materials. With this approach, only two out of 16 various descriptive metadata fields are required, making the system an accessible and effective tool for materials across Library repositories and departments. Administrative and technical metadata requirements were also developed with the goal of long-term preservation activities in mind.

B. *Ingest Process*

When testing the previous repository, we found that other organizations were performing ingest processes prior to system ingest because it was 'easier' than having the system do it. We wanted to make sure our system did the work for us. Tabula's current ingest process walks through 18 steps that work to add content to the database by assigning ids, associating metadata to the content, verifying that files exist for ingest, performing an anti-virus check, creating/verifying checksums, extracting technical metadata, copying files to multiple storage locations, creating derivatives if needed, and producing a report of the ingest process. Our user interface allows us to build and ingest a SIP as well as check the status of individual SIP steps.

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

https://hdl.handle.net/2142/121068

C. Hardware And Software Environment

Encompassing both a web-based graphical user interface and command line menu driven tool, Tabula utilizes the RESTful API Design Methodology [RADM]. At Tabula's core a series of microservices written in the Python scripting language interact with a MySQL database employing a string of RESTFUL APIs built on Spring Boot, a Java-based Framework. Tabula's web interface is built upon Python Flask, a lightweight web application framework, also known as a Web Server Graphical Interface [WSGI].

The modular design of Tabula at the highest level has an external Web Proxy Server, a primary Application Server, secondary Application Servers, an internal proxy server, the RESTful API server, and the MySQL Database Server. The application Servers run Red Hat Enterprise Linux 8 and have 128 GB of RAM and 8 CPUs. The working application space is 10 TB and our "permanent storage" has twin 1 PB of storage and a Tape Library with three LTO9 drives and 100 slots.

D. Database Tables

A series of database tables work together to build the foundation of Tabula. The Element table records information about the 'things' that we want to preserve. Elements are classified as either an Asset or a File. An asset represents one or more files. Both assets and files have their own unique ID numbers.

The Affiliation database documents the names and contact information of organizations, institutions, and repositories for which the elements are related. For example, at the point of ingest we associate materials with the organization from which they came from as well as the access repository in which the materials can be found.

A Metadata table documents descriptive and technical metadata elements available for use within the system. Title, author, description, date of creation, publisher, and geospatial information are some of the descriptive metadata fields. This database also records the technical metadata elements that are captured during the ingest process using DROID.

All actions taken on the Elements are tracked within the Event database. Creating and verifying checksums, performing a virus scan, creating a copy, and moving a file from one location to another, are some of the events that are tracked. We record the type of event, what tool was used, who initiated the event, and the outcome of the event.

Organizing the system with these database tables offers flexibility to develop new tables as needed for additional functionality or needs. We expect to add new tables to assist with our preservation activities.

E. Reporting

We are in the process of developing both internal and external reporting functionality. We want to be able to understand what happens to objects and when, so we document these internal events on SIPs and objects. We also want to be able to answer questions such as: How much content is being preserved? What file formats do we have? When did we receive the materials? We intend to utilize Tableau (a reporting software) to query our databases to produce not only standard reports for our own use but to also create on demand reports in response to questions from our stakeholders about the contents of the repository.

F. Preservation Activities

We are currently building a workbench area where 'problem' objects will be sent to be addressed. The workbench area will be used to address issues discovered upon ingest as well as issues discovered at a later point. It will also be used if a stakeholder needs to review an object. A preservation planning area, an area that will be used to monitor and address preservation concerns, will be developed in the future. We expect to use this area to address objects associated with at risk formats. The preservation planning area will have access to tools to assist with file format identification, migration, and more.

III. Ongoing Work

Developing Tabula continues to be an iterative process. To date, our main focus has been on the environment and ingest process, so we can at minimum preserve our materials. We continue to improve and add functionality to our reporting activities, preservation actions, and the user interface. The end goal is to have a responsive design with UMN branding that will allow users to ingest a set of objects, perform and complete preservation work on one or more of those objects, and allow for non-preservation staff to search, find, and download objects when needed.