the digitalarchive group

Making History Accessible

Digital Curation Plan

2018 - 2019

Prepared by the Digital Archive Group for:

Indiana University

INDIANA HISTORICAL SOCIETY

DR. MAYA ANGELOU FOUNDATION

Lori Moody

December 7, 2018

Version 1.2

TABLE OF CONTENTS

INTRODUCTION

The Digital Archive Group (TDAG)'s mission is to preserve historical digital data and artifacts for museums, libraries and organizations. We seek to support these institutions in creating high-resolution digital files and powerful digital asset management systems to view their historic content.

This document is the digital curation plan (DCP) for the new projects we will be working on for the 2018 – 2019 fiscal year. The DCP describes the data management life cycle for all data and artifacts that will be collected, processed and/or generated by the projects. The DCP describes the following:

- The handling of the data during and after the project's completion
- The type of data that will be collected, processed, or gathered
- What methodology and standards will be applied
- If and/or how the data will be made openly accessible
- How the data will be stored

Digital Collections Overview

TDAG has received three new digital collections to curate for a new digital asset management system. These collections include Indiana University's research on the South Aegean Volcanic Arc (SAVA), which was produced from a long-retired researcher awarded the Gold Medal Award for distinguished Archaeological Achievement from the Archaeological Institute of America. The next collection is in partnership with the Indiana Historical society, which hopes to digitally preserve artifacts of the Owens Family, a prominent family who settled in the New Harmony community in southern Indiana from the 1800s. The last collection is from the Maya Angelou Foundation, which seeks to preserve the digital footprint of famous writer, Dr. Maya Angelou. We will be archiving Angelou's social media accounts and providing a curated collection of her own words for the foundation. The DCP steps are outlined in the following pages. The procedures described in the DCP shall be followed by all member of the Digital Archive group to ensure that data is transferred and used in a secure setting.

Metadata for Curation

The metadata standards are outlined in Figure 1. Each collection will include technical, intellectual, and object metadata. Metadata completeness will be checked using controlled vocabularies. These guidelines are to increase accessibility of client's files in a digital repository. Dublin Core Metadata standards are to be used.

Dublin Core Metadata Fields:

Title	Creator	Subject
Description	Publisher	Contributor
Date	Type	Format
Identifier	Source	Language
Relation	Coverage	Rights

*Staff should make sure each type of metadata is incorporated before delivering the final product.

Figure 1

Metadata Workflow
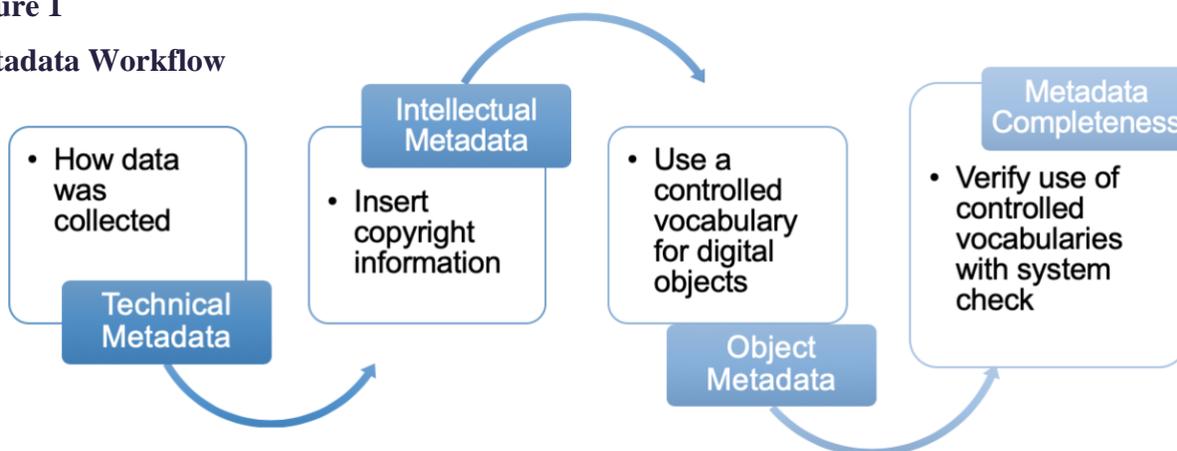
Discovery & Accessibility

- **Standards of metadata creation**

A metadata schema will be created, and a key maintained of the available metadata in the digital asset management system.

- **Approach toward search keywords**

In addition to metadata embedded in the file, we will add keywords, and specific filters based on the metadata elements. We will allow these keywords to be refined across the various types of data in the digital asset management system (e.g. search dataset by clay elements, social media account, year, etc.) A list of common vocabulary and keywords will also be available to the client.

- **Use of persistent and unique identifiers**

Each digital artifact will be supplied with a persistent identifier (PID) and managed in a CSV spreadsheet by the TDAG group. Instructions on how to maintain and check each PID will be provided to the organization upon completion.

- **OCR of Documents**

Each document, paper or presentation will run through an OCR (Optical Character Recognition) process, which will allow any system to search these documents for specific titles, words, headlines, etc.

- **Naming Conventions**

For metadata, datasets and file names, we define naming conventions based on these mandatory parts:

- A prefix, indicating if it is a dataset, an image, presentation, Facebook post, etc.
- A root composed by:
  - The short and meaningful name of the dataset/metadata
  - The acronym/short name of the data provider/organization
  - A suffix indicating the date of the last upload into the digital repository in in YYYYMMDD format.
- **Clear versioning**

The version of the file may change once metadata and other elements are added to the file. The version of the file must be present in the naming convention and the use of the date as a suffix. This will indicate the last version of the file.

Directly accessible assets via Image Relay DAM system

Image Relay (https://www.imagerelay.com/) is the digital asset management (DAM) system that will store all assets for the collections. It provides access to data through a web interface and a platform to view, edit metadata, and share files across devices easily. It uses an open architecture and extensible API for applications and plugins, which will work for most storage solutions. Some constraints include the inability to create folders via a CSV spreadsheet. Folders for images must be created on Image Relay, but then the files are easy to upload using a CSV spreadsheet or FTP software.

Storage and Backup

The archive will be stored using a three-part storage system. Files will be stored on an internal hard drive in the office and another hard drive to be stored in a separate unit off-site. The third copy of the file will be stored in a cloud-based storage system for online viewing in the Image Relay DAM system. TDAG will store the archives in a Dark Archive after 10 years, specifically used for storage only at the request and monetary agreement of the organization. Amazon's AWS system will be used to store these files safely in the cloud until the client requests a copy.

Data Recording and Retention

The retention of digital assets for each organization is of utmost importance to TDAG. The curatorial staff will have the responsibility of recording, retaining, and storing all data of the projects. The records will include sufficient detail on metadata schemas, naming conventions, checksums, and persistent identifiers for each artifact. Since different types of data will be generated based on each specific collection, we will use different data standards to manage the data artifacts.

Each collection will be stored with TDAG for up to 5 – 7 years as a backup for each organization. These files will be available on the Image Relay digital asset management system as well as accessible from one of our external storage drives. The collection will then be stored in Amazon's AWS dark archive for an additional 10 years. Dark archiving service must be renewed with TDAG in order for more long-term preservation.

Staffing

- **Curatorial**
    - o The curatorial staff assists in creating and ingesting files into the DAM system with keywords, file names and metadata
- **Technical**
    - o The technical staff maintains the networks and computer systems as well as implementing APIs from the physical digital repository to the cloud-based DAM system.
- **Management**
    - o Management is in charge of disaster management, authentication of digital asses, data rights and customer/organization outreach.
- **Clerical**
    - o Must maintain records of all metadata and naming conventions used. The staff will also take care of filing and managing documents for external shareholders.

Technologies Required

TDAG will make use of free and/or open source software when possible. There will also be a need for the Adobe Creative Suite to modify images and text for ingestion. OCR software is required to optimize document findability. The digital repository will also need access to the JHOVE software to run fixity and MD5 checksums to check file validity. TDAG will need at the minimum two 10 TB storage hard drives to store working and archived files of each project. In addition, a reliable high-speed internet connection with a 200+ MB upload/download speed is required to store the data on Image Relay's cloud-based digital asset management system.

Disaster Recovery Plan

The main systems include desktop computers, digital repository, and public catalog collection. The type of archival materials ranges from the scientific to humanities. There is a small chance the organization could face a few natural disasters. This includes floods, blizzards, heat waves, lightning strikes, power failure, and even tornados. The disaster recovery plan takes into account these natural disasters, which can cause hardware issues, massive storage failure and power failures. These include saving all data on three storage solutions, including an offsite backup (such as a cloud solution such as Amazon AWS and other online digital repositories) and updating servers daily.

Other risks comprise of server failures, data hacks or cyberterrorism. A preventative measure includes installing anti-virus software on personal PCs, which are often transferring files to our digital repository. We want to mitigate any risk of these files carrying malicious software to the system. Firewalls should also be setup on the public catalog collection to prevent hacking or malicious software from taking down the server. Most systems can afford to be down for several hours, but we want to make sure our public-facing site is operable in two hours in case of a disaster. It's important to install redundant backup servers in and offsite to prevent total system inaccessibility.

Definitions:

- **UPS** – Uninterruptible power supply (power cords, sockets, etc.)
- Recovery Point Objective (**RPO**) – the acceptable latency of data that will not be recovered
- Recovery Time Objective (**RTO**) – the acceptable amount of time to restore the function
- Maximum Tolerable Period of Disruption (**MTPD**) for each activity is not exceeded.

Critical System	RTO/RPO	Threat	Prevention Strategy	Response Strategy	Recovery Strategy

Desktop Computer	8 hours/ 4 hours **MTPD:** 1 week	Hardware failure, software failure	Backup computer files every day on server, install virus software, UPS	Hard reboot computer, run virus scan, switch to new computer	Fix/replace computer hard drive, reboot data from last backup
Digital Repository	4 hours/ 2 hours **MTPD:** 1 Day	Server failure	Secure equipment room, backup server on offsite storage	Switch over to backup server, validate UPS functions	Fix/replace primary server; move back to primary server from offsite storage
Digital Repository	4 hours/ 2 hours **MTPD:** 1 Day	Storage failure or Massive Storage Failure	Secure equipment room, backup server on offsite storage, UPS, maintain store of spare disks	Replace damaged hard drive for spare, restore Data	Obtain more spare disk drives
Public Catalog Collection	2 hours/ 2 hours **MTPD:** 1 Day	Server failure	Store data on server and backup daily	Switch over to backup server, validate UPS functions	Fix/replace primary server; move back to primary server
Public Catalog Collection	2 hours/ 2 hours **MTPD:** 1 Day	Hack/security breach	Store data on server and backup daily, install firewall	Switch over to backup server, validate UPS functions	Fix/replace primary server; move back to primary server

SOUTH AEGEAN VOLCANIC ARC (SAVA) E-SCIENCE DATABASE COLLECTION

1. Data Summary

The e-science collection described in this digital curation plan (DCP) is for Indiana University's South Aegean Volcanic Arc (Sava) Database project. The collection includes chemical analysis and other data about rocks and pottery of the South Aegean Volcanic Arc (SAVA).

Purpose of the data collection/generation and its relation to the Digital Archive Group

The purpose of the database is the acquisition of high-quality, quantitative rock and clay-rich sediment data that can be used for the provenance (acquiring a specific date) of artifactual materials such as ceramics and milestones, presumed to have originated in the Saronic Gulf portion of the South Aegean Volcanic Arc. Data was also collected from the islands of Aegina, Poros, Milos, the peninsula of Methana, and Santorini.

The Digital Archive Groups seeks to preserve history for the digital age in collaboration with other organizations. The SAVA collection data can no longer be stored on Indiana University's website since the researcher's retirement. TDAG will take all the data from Indiana University's site and archive it off-site to be ingested in a digital repository for historical reference.

2. Collection Inventory

Types and formats of data

Types of data:

- *Numerical Data:* This data includes information for its geochemical database, Aegina island geologic history, and geologic field work
- *Image Data:* Includes images of ceramics, maps, and petrographics.
- *Collaborators:* A list of collaborators for the SAVA collection, which includes biographies and list of collaborators' recent research works.
- *Acknowledgments:* The data includes a list of funders for the project and the years of their contribution. Also, there's a bibliography of works related to the database

Formats of the data:
- Data and metadata will be stored and transferred in a comma-separated values (CSV) format for numerical data.
- Many of the files are saved as PDFs and PowerPoint (PPT) format. Data will be saved and converted as a PDF/A (PDF Archival) file to preserve the integrity of the files.
- Images are saved as JPEG files on the website. Images will be saved as 16-bit TIFF archival files to reduce any file degradation. Web pages will be saved as a PDF files as well and stored with metadata data using the Webpage-to-PDF tool.

3. Data Preservation and Archiving

What data will be archived?

The type of data that will be archived includes, numerical data, list of contributors and donors will be converted to comma-separated values (CSV) format, web pages, reports and PowerPoint presentations

as PDFs. Images of ceramics, maps and petrographic will be converted to TIFF files for long-term storage. Interactive maps will not be archived due to the inflexibility of the format. Images of the maps will be stored for reference to the university.

Preservation & Archive strategy

Storage:

The archive will be stored using a three-part storage system. Files will be stored on an internal hard drive in the office and another hard drive to be stored in a separate unit off-site. The third copy of the file will be stored in a cloud-based storage system for online viewing.

Access & Discovery:

We will provide easy access to the materials for the university using an in-house digital management system, which will allow the university to download the files in a TIFF, JPEG, PDF or CSV format. Files will be stored by the structure of the website: Numerical Data, Image Data, Collaborators, Acknowledgements, and Webpages.

Preservation Planning:

All non-standard formats shall be converted to an interoperable format for ease of access. A MD5 checksum and JHOVE validation and verification will be performed on the data to ensure files aren't corrupted before ingesting into the digital asset management system. Metadata completion will be checked for misspellings and all Dublin Core metadata standard fields are entered correctly.

4. Data Organization, Documentation and Metadata

Current metadata that exists on the site includes: numerical data, image data, collaborators, acknowledgements.

New Metadata Standards

During the assessment stage, we will use the website metadata (please see page 5 for file naming convention and metadata structure), and file type (PDF, JPG, etc.). Web page title must also be included in the metadata to coincide with which artifacts were on that particular web page.

5. Staffing & Technical requirements

The project will require the following staff members in order to complete the project: our current staff includes three digital curators, and the project will require a data entry specialist to complete the database and metadata of the digital objects.

- Two (2) **Data Curator/ Digital Asset Manager**
  - o Responsible for ingesting of files into internal DAM system and working in collaboration with management lead to produce metadata schema, file format conversions and fixity checks.
- One (1) **Data Entry Specialist**
  - o Will oversee the input of metadata to spreadsheet for ingest by the Data Curator

Technologies Required

The project will require the use of the free Webpage-to-PDF tool. In addition, all other requirements can be found on page 6-7 for technical requirements to complete this project.

OWEN'S FAMILY DIGITAL HUMANITIES COLLECTION

1. Data Summary

The research project described in this digital curation plan (DCP) is for the Owen's Family Digital Humanities Collection. This prominent family settled the New Harmony community, a utopian common in the 1800s in southern Indiana. The community was known as the Athens on the Wabash for its support of the arts and sciences and were a major contributor to 19th century knowledge and culture. The researcher is contributing a complete collection of his published papers and books.

Purpose of the data collection/generation and its relation to the Digital Archive Group

Indiana Historical Society wishes to archive the research conducted for the researcher for a new museum exhibit and archive about the Owen's family. In collaboration with TDAG, the Historical society hopes to organize all materials by subject and date to preserve the affect the Owens had on the New Harmony community throughout the 19th century.

2. Collection Inventory

Types and formats of data

Types of data:

The Owen's collection has various TXT, JPEG, HTML, and .DS_Store files split between two databases. Some of the files are available on Google Drive from the organization, and the rest are available in a zip file titled **Owen_Correspondence**. The zip file has JPG images of various hand-written letters from the Owens family in a sequential order.

Formats of the data:

- Excel files will be converted to tab-delimited files be stored and transferred in a comma-separated values (CSV) format.

- The simple TXT files will be converted as a PDF/A (PDF Archival) file to preserve the integrity of the files.

- Images are saved as JPEG files in the two databases. Images will be converted as 16-bit TIFF archival files to reduce any file degradation.

- HTML files will be saved in the digital asset management system as well as a PDF file of the website using the Webpage-to-PDF tool. The files should be kept in the same folder and structural metadata added to show the data is related.

What data will be archived?

The type of data that will be archived includes TXT files, which stores historical information about New Harmony and the Owns family. Excel files appears to be a database of the JPEG files found in the image-only database, Owen_Correspondence. The file should be stored in the same folder as the

correspondence images. All JPEGs will be stored as TIFF files to preserve the integrity of the file, and all HTML provided will be saved as a HTML file. The HTML generated website will also be turned into a PDF and stored along with the HTML file. We will not archive DS_Store files. The current approach is to not ingest the .DS_Store files, as they are not regarded as an artifact of the collection, more as an artifact of the system the collection came from.

Preservation & Archive strategy

Storage:

The archive will be stored using a three-part storage system. Files will be stored on an internal hard drive in the office and another hard drive to be stored in a separate unit off-site. The third copy of the file will be stored in a cloud-based storage system for online viewing.

Access & Discovery:

We will provide easy access to the materials for the historical society using an in-house digital management system. This will allow the organization to download the files in a TIFF, JPEG, PDF or CSV format. Files will be stored by subject and document type. An example of the subject type would be Owen's Correspondences, Website, and History of New Harmony. It's imperative that HTML files are stored with PDF copies of the fully-rendered site in a folder.

Preservation Planning:

All non-standard formats shall be converted to an interoperable format for ease of access. A MD5 checksum and JHOVE validation and verification will be performed on the data to ensure files aren't corrupted before ingesting into the digital asset management system. Metadata completion will be checked for misspellings and all Dublin Core metadata standard fields are entered correctly.

6. Data Organization, Documentation and Metadata

Current metadata that exists in the databases includes file names, HTML code. and an Excel file, which has a record of the files stored in the Owen_Correspondence database. The files from the first database does not appear to have any metadata attached. A metadata schema will need to be created. The main folder structure will be, *Owen's Correspondences*, *Website*, and *History of New Harmony*. In addition, each letter from the Owen's_Correspondence database requires transcribing from written

word to digital text. The text should be saved along with each JPEG file in the same folder. Also, text should be added to the description metadata.

New Metadata Standards

During the assessment stage, we will use an in-house folder structure to organize the artifacts (please see page 5 for file naming convention and metadata structure), and file type (PDF, JPG, etc). Files should include structural metadata to show a relation to an artifact in the folder. Example would be HTML file must show it's linked to the PDF version of it. Owen Correspondences must remain in the same folder, and we will use the Excel spreadsheet to add descriptive metadata to the files.

7. Staffing & Technical requirements

The project will require the following staff members in order to complete the project: our current staff includes three digital curators, and the project will require a data entry specialist to complete the database and metadata of the digital objects. The data entry specialist will also be required to transcribe the letters. The project is data intensive and will require at least two data entry specialists.

- One (1) **Data Curator/ Digital Asset Manager**
    - Responsible for ingesting of files into internal DAM system and working in collaboration with management lead to produce metadata schema, file format conversions and fixity checks.
- Two (2) **Data Entry Specialist**
    - Will oversee the input of metadata and transcribing of the Owens letters.

Technologies Required

The project will require the use of the free Webpage-to-PDF tool. In addition, all other requirements can be found on page 6-7 for technical requirements to complete this project.

DR. MAYA ANGELOU SOCIAL MEDIA COLLECTION

1. Data Summary

Purpose of the data collection/generation and its relation to the Digital Archive Group

Dr. Maya Angelou was a poet, educator, historian, best-selling author, actress, civil-rights activist, producer and director according to her Twitter and Facebook pages. Not only was she best known as an award-winning writer for books like "I know Why the Caged Bird Sings," but she was also a prolific user of Twitter during the later years of her life as well as posting semi-regularly on Facebook. As a poet, Maya Angelou found the Twitter site's 140-character limit perfect for short poems and musings, which she updated fairly often. Her Twitter account now has over 600 thousand followers, and her tweets are often re-tweeted every day by her fans. The Digital Archive Group would like to continue to document personal artifacts of prominent social figures. A partnership with the Maya Angelou foundation will achieve the goal of preserving this public figure's words of wisdom for the future. Angelou's words will also be used in a special exhibition to celebrate the life of Dr. Angelou.

2. Collection Inventory

Dr. Maya Angelou's foundation has an account on all the major social media channels. However, after further analysis, it appears that only Twitter and Facebook has original posts from the writer.

Types and formats of data

The data is relatively unstructured. We've grouped the data by social media, channel, followers, posts, and likes.

Social Media Channels			
Total from user:	**Followers**	**Posts**	**Likes/Reactions**
Twitter	612,6196	760	N/A
Facebook	5,496,690	1,251	5,669,616
Instagram	3711	44	N/A
YouTube	1268	31	0

3. Data Preservation and Archiving

What data will be archived?

The curation plan will include Maya Angelou's Twitter and Facebook posts. These were the sites she personally used from 2010 until 2014. There will be a clear distinction between which posts were made before and after her death. Any post made after May 2014 was from the family and Maya Angelou foundation. Although posts after 2014 weren't Angelou's original words, they are important artifacts to show the impact her work has had on the public. Her Twitter and Facebook posts should be archived for the world to see just like her poems. The Maya Angelou Foundation also has an Instagram and YouTube Channel, but Angelou hasn't personally contributed to these channels. We will note the social media channels here but will not focus on archiving them.

Preservation & Archive strategy

We used a Google search and the Maya Angelou foundation website to find her social media channels. Facepager and API tools were used to scrape the data. Raw data is available on Excel document titled, "MayaAngelou-SocialMediaData." The data will be parsed and added to the social media collection as a spreadsheet. We will also store images shared and create JPEGs of Facebook posts and Tweets made personally by Dr. Maya Angelou. The JPEGs of these images will be saved in a folder along with the CSV file of the recorded social media posts.

Preservation & Archive strategy

Storage:

The archive will be stored using a three-part storage system. Files will be stored on an internal hard drive in the office and another hard drive to be stored in a separate unit off-site. The third copy of the file will be stored in a cloud-based storage system for online viewing.

Access & Discovery:

We will provide easy access to the materials for the foundation using the Image Relay DAM system. This will allow the organization to download the files in a TIFF, JPEG, PDF or CSV format. Files will be stored by social media channel and organized by date. Files will also be saved as posts and comments.

Preservation Planning:

All non-standard formats shall be converted to an interoperable format for ease of access. A MD5 checksum and JHOVE validation and verification will be performed on the data to ensure files aren't corrupted before ingesting into the digital asset management system. Metadata completion will be checked for misspellings and all Dublin Core metadata standard fields are entered correctly.

8. Data Organization, Documentation and Metadata

Current metadata that exists in the databases includes Facebook Posts, Tweets, Likes, Retweets, Comments, and Followers. A metadata schema will need to be created. The main folder structure will be determined based on the social media channel. The database for each channel will be saved along with JPEGs of the posts by Dr. Maya Angelou.

New Metadata Standards

During the assessment stage, we will use an in-house folder structure to organize the artifacts (please see page 5 for file naming convention and metadata structure). and file type (PDF, JPG, etc). Files should include the file type, social media channel, and what kind of post it is from the social media channel.

9. Staffing & Technical requirements

The project will require the following staff members in order to complete the project: our current staff includes three digital curators, and the project will require the help of a data manager and data entry specialist to complete the database and metadata of the digital objects. The data entry specialist will also be required to transcribe the letters. The project is data intensive and will require at least two data entry specialists.

- One (1) **Data Creator**
  - The data creator will use the web scraping and analytics tools to retrieve the social media data.
- One (1) **Data Curator**
  - Responsible for ingesting of files into internal DAM system and working in collaboration with management lead to produce metadata schema, file format conversions and fixity checks.
- Two (2) **Data Entry Specialist**
  Will oversee the input of metadata and transcribing of the Owens letters.

Technologies Required

The project will require the use of the Facepager and API tool in order to extract the data. We will need a social media analytics tool to download all Likes, posts and images as well. The information will be collected as an Excel file and stored as a CSV file, and images downloaded from these tools. In addition, all other requirements can be found on page 6-7 for technical requirements to complete this project.