*Digital Curation PLan*

Digital Repository of the Midwest

August 2021

*Mikaela Slade*

*Dominican University*

*LIS 889 – Digital Curation*

Introduction

The Digital Repository of the Midwest is a group funded and supported by academic libraries across the Midwest. The purpose of the group is to offer preservation and access options to academic institutions. Another goal of DRoM is to provide outreach to individuals within these institutions by teaching about the importance of Digital Asset Management and offering support to those who wish for their work to be preserved.

DRoM offers long-term storage to creators as well as accessibility and authenticity assurance throughout the data’s lifecycle. Data is stored on secure servers at DRoM headquarters as well as physically in an off-site location and remotely on Cloud-based servers. DRoM follows the Dublin Core Metadata Initiative and includes as many DCMI terms as applicable. This standard was chosen due to its wide-spread use as well as staff’s familiarity and expertise using Dublin Core standards. DCMI is also a strong choice for data sharing and re-use, which will make the data available in these collections accessible to a wider audience.

This document will function as an example of three different types of content that DRoM plans to curate: a science collection, humanities collection, and a digital footprint assessment. The data within each of these collections is varied and different data requires different approaches for best preservation. Data preservation and curation is not one-size fits all nor is it a single event. These digital curation plans will detail the types of files preserved and why and a plan for inventory and assessment.

Staffing requirements for these projects include two Content Analysts to supervise data migration and authentication, three Metadata Specialists to extract metadata and manually enter information that cannot be automated, and one Data Curator who will oversee the final integration into the DRoM database.

e-Science Collection

*Summary*

The South Aegean Volcanic Arc project (SAVA) represents the collaborative work of a retired researcher who was recently awarded the Gold Medal Award for Distinguished Archaeological Achievement from the Archaeological Institute of America. Once hosted on Indiana University’s website (<https://www.indiana.edu/~sava/>) it has now been removed but is still accessible through more than a dozen captures spanning from 2008 to 2016 on the Internet Archive’s Wayback Machine: (<https://web.archive.org/web/20150604154137/http://www.indiana.edu/~sava/index.html>).

The South Aegean Volcanic Arc databank was originally created for the purpose of providing artifactual material provenance and became a multi-disciplinary project comprised of Classical Archaeology, Geological Sciences, and Information Technology. The collection has rock and sediment data from fieldwork in the Saronic Gulf portion of the South Aegean Volcanic Arc, plus the island of Santorini and specifically focuses on the distribution of Aeginetan Ware, a ceramic important to the Greek Bronze age.

The Digital Repository of the Midwest has partnered with Indiana University to better preserve the research originally available through Indiana University’s website. Although still available archived online, DRoM seeks to save the datasets and incorporate them into their own digital repository so that it can be easily searched, accessed, and cited by those academic institutions that support DRoM. DRoM believes that preserving high quality data such as SAVA is a valuable pursuit and supports future research.

*Inventory*

Types of data:

* Numerical data
  + Geochemical
  + Geological
* Image data
  + Ceramic
  + Map
  + Petrographic
* Acknowledgements
  + Bibliography
  + Contributors
  + Funding

Formats of data:

* Text -current text will be saved as .pdf
* Image (.jpeg) -.tiff is preferred format but not available in this case
* PowerPoint (.ppt) -with a .pdf version as backup
* Adobe Flash animation – since the interaction with the animation is not a key feature, these features will be screen recorded and saved as a .mp4 file
* Web ARChive (WARC) format is recommended by the Library of Congress and the Smithsonian for archiving websites

*Strategy*

The strategy for this collection is to migrate the data currently available on the archived website into more sustainable formats. The website is a product of its time and is not as easy to navigate as it once was. Additionally, the site utilizes defunct formats such as Adobe Flash which makes it inaccessible to most modern audiences. Current data will be transferred to DRoM’s database where it will be accessible to the academic libraries that subscribe to it. Numerical and image data will be saved but not website data because that is where the data is that can be re-used and referenced by current researchers.

As much metadata as possible will be extracted through automated methods and then reviewed for accuracy. Individual objects will receive unique identifiers as well as an MD5 hash for checksum verification. Objects within this collection were already divided into datasets which provides pre-existing metadata such as location and descriptions.

e-Humanities Collection

*Summary*

New Harmony was the site of two communities founded on utopian and anarchist ideals in the early 1800s. The first community was settled by George Rapp who then sold the land to Robert Owen. The site attracted educators, scientists, and idealists and became a center of culture for many years. The Digital Repository of the Midwest has acquired the collection from a researcher who no longer needs the information and wants his work to be archived long-term. DRoM sees this as a valuable opportunity to preserve Midwest history and offer it to other researchers for future use. The Owens family research offers fascinating insights into historical, political, religious, and even geological aspects of early 19th century Indiana.

This collection is comprised of a variety of documents, images, and other data about the Owens family, a historical group which settled the New Harmony community. A significant portion of the collection is the Robert Dale Owen Correspondence and Personal Papers, which are scanned images of the originals. These images have descriptions located in a separate .txt file and an .xlsx file which could make it easier to attach metadata to these images. Ultimately, it would be better if this data was attached to the images themselves instead of separately but at least that information is still there in some form.

*Inventory*

Types of Data

* Papers/books
  + History of New Harmony
  + History of Anarchism
  + Harmony Society Manuscript
* Website Information
  + .DS\_Store – this type of file is a Mac OS specific folder that contains attributes/metadata about the folder
  + newharmony.html
  + nh\_people.html
  + The folders newharmony\_files and nh\_people\_files contain images and gifs used on the website
* Correspondence scans
  + 74 folders
  + 916 items in total
* Images

Formats

* Text -current .doc or .txt files will be saved as .pdf where appropriate
* Image – most images are in .jpg form
* Website – two .html files
* XML spreadsheet

*Strategy*

Like the e-Science Collection, the e-Humanities collection will employ a data migration strategy. Most of the documents are already in readable formats that will easily migrate to DRoM’s database. However, the bulk of the collection (as previously mentioned) are scanned copies of handwritten documents from the correspondence and personal papers of Robert Owen. These do have short descriptions and titles in a separate file, but the decision remains whether to transcribe them. The cursive handwriting can be prohibitive to searching and reading but transcription is a tedious and time-consuming process. It is possible to employ programs such as Adobe’s Acrobat Pro to recognize text in PDFs but it is doubtful that it will be able to retrieve a lot of useable data from these scans. A possible approach is to upload the scans as is and employ volunteers to transcribe the letters so that they can be machine searchable in the future.

Not all the data in the collection will migrate to DRoM’s database. For instance, the notes on Roger D. Branigin as well as the email correspondence discussing funding timelines and dates are not relevant or important to the New Harmony collection and will be omitted. Additionally, the HTML data shows the code for a basic website, but the same information could be displayed as a PDF and still be searchable. The format of a website is not integral to the data itself and so other options such as emulation are not required.

Digital Footprint Collection

*Summary*

Alexandria Ocasio-Cortez is a member of the U.S. House of Representatives from New York City. She is known for being a grassroots politician and the youngest woman ever to serve in the U.S. Congress. Compared to her fellow congressmen, Ocasio-Cortez (or AOC as she’s often called) has a strong social media presence and often uses it to reach out to her constituents as well as others interested in American politics. In many ways, AOC is a digital celebrity and highlights how largely social media factors into the political sphere. It is also an indication of how younger generations desire authenticity and connection from their celebrities and public figures must consider their digital footprint as part of their overall brand.

Digital content is often created to be disposable and can disappear before its impact can be assessed. By saving an individual’s social media presence, researchers and others can look back and gather current attitudes and events as they were happening.

*Inventory*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Social Media Platform** | **username** | **Types of Content** | **# of followers\*** | **# of posts\*** | **Date Created** | **Last Active Date** |
| Twitter | @aoc | Text posts | 12.7M | 12,776 | 2010-04-28 | 2021-08-01 |
| Twitter  (Professional) | @repaoc | Text posts | 681,293 | **1,037** | 2018-12-29 | 2021-08-01 |
| Twitch | AOC | Video stream | 1M followers  3M channel views | 4 | 2020-10-20 | 2021-01-28 |
| Facebook  (Interest page) | @OcasioCortez | Text posts | 1.7M likes | Unknown | 2017-05-02 | 2021-07-31 |
| Facebook (Governmental Official page) | @repAOC | Text posts | 224,553 likes | Unknown | 2019-01-03 | 2021-07-30 |
| Instagram  (Personal Account) | @aoc | Photo/Video/Text | 8.7M | 465 | 2012-01-23 | 2021-08-01 |
| Instagram (Professional Account) | @repaoc | Photo/Video/Text | 1.4M | 71 | 2019-01-03 | 2021-08-01 |
| YouTube | Alexandria Ocasio-Cortez | Video | 78.2k subscribers | 54 | 2018-05-27 | 2021-03-12 |

\*as of 2021/08/01

*Strategy*

DRoM will utilize the third-party subscription-based service Archive-It (affiliated with the Internet Archive) to preserve content from Facebook, Instagram, and Twitter. Archive-It captures content at scheduled intervals. This service automates social media capture. According to Archive-It’s About Page: “Material archived via Archive-It is stored in not-for-profit data centers independently owned and operated by the Internet Archive and is available for users to download themselves for additional preservation and sharing.” The service also has built-in keyword and metadata creation tools so that the data and metadata can be stored in one location. As for YouTube and Twitch accounts, the best strategy is to migrate the data to DRoM servers by downloading the videos as MP4 files and extracting the metadata from the original sources.

A potential issue is duplicate captures since content is often created on one site and shared on multiple. For instance, recorded Instagram Live streams were also posted on her YouTube channel. Furthermore, most items that are shared on the professional accounts of AOC are also shared on her personal accounts. Due to this (and because it is likely that her professional accounts are not actively run by the Congresswoman but a team), DRoM will omit the capture of AOC’s professional accounts and focus on her personal ones.

AOC seems to favor Twitter as far as the sheer number of posts created there. However, a direct comparison of social media platforms is more complicated than what the above data shows. For instance, Instagram offers “stories” and live streams that allow for frequent video or photo updates that disappear after a day. Social Blade was a Social Media analysis site used to compile the data and Social Blade did not have data for stories. Social Blade also did not have data for Facebook posts because Facebook does not track or number your posts.