Collection as Data in Texas Digital Library Repositories:

Case Study and Practical Recommendations for Use

TEXAS
The University of Texas at Austin

Anne Morgan, Karla Roig, Katy Tuck

Introduction

This project provides practical recommendations for the use of Collections as Data in Texas Digital Library (TDL) repositories using the Texas Conference on Digital Libraries (TCDL) Proceedings as a case study. The recommendations were written by three graduate students at the University of Texas at Austin's School of Information as part of a semester-long group project with TDL. The purpose of this project is to increase understanding and improve methods of using Collections as Data for TDL members.

TDL and its member repositories use DSpace, an open source repository application. The TCDL proceedings are stored in the TDL DSpace repository. This collection includes presentations, posters and other materials from the annual Texas Conference on Digital Libraries. The proceedings were selected for use as a case study because the project contact at TDL indicated an interest in having the group work with this collection.

What is Collections as Data?

In "On a Collections as Data Imperative," Thomas Padilla (2017, p. 1) describes Collections as Data as "reframing all digital objects as data," defined as "ordered information, stored digitally, that is inherently amenable to computation." Collections as Data involves the application of computational methods in order to analyze and use collections of digital objects meaningfully. These computational methods include text mining, data visualization, mapping, image analysis, audio analysis, and network analysis (Padilla et al., 2019, May 20b, p. 2). Using Collections as Data is an example of a digital humanities project.

The following are potential uses of Collections as Data:

- Discover patterns in data
- Present data in a new, interesting way
- Attract new users and increase discoverability
- Create compelling visualizations of datasets to tell a story about your collection(s)

These potential uses point to two main benefits of using Collections as Data: increased useability and visibility of collections. Collections as Data moves "beyond traditional use," allowing "more flexible access" to collections (Wittmann et al., 2019, p. 49).

Methodology

The task presented to our group was the following: collect metadata from the TCDL proceedings, clean the data, and analyze it computationally. We approached our Collections as Data project by dividing it into a three-part process: metadata extraction, metadata cleaning, and metadata visualization.

Process

Data Extraction

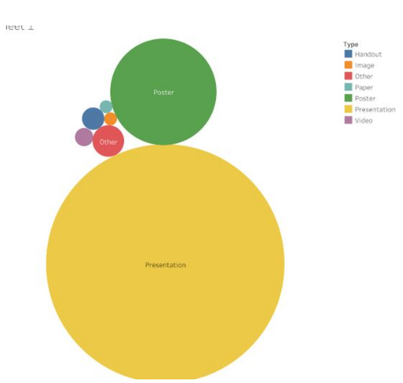
This process involved the extraction of the desired metadata from the OAI PMH Dspace data portal. The metadata was downloaded as XML files and then the metadata was imported using Excel and Notepad++ to divide the elements into their corresponding columns.

Data Cleaning

The process for cleaning the tabular data was done in Excel. To begin cleaning the data, we used the filters in order to sort and edit multiple cells containing specific elements at the same time and start shifting all elements to the right that don't correspond to the column subject heading until every element is in the correct place.

Data Visualization

For the data visualization process, we used Voyant Tools and Tableau Public. The image on the left is a word cloud created by importing the cleaned "Subject" metadata from the TCDL proceedings into Voyant Tools. This visualization presents the terms most used in the collection. The image on the right is a visualization of the "Type" metadata from the collection which shows the types most commonly presented in TCDL. These are two examples of simple and straightforward data visualizations.



List of Data Extraction, Editing, and Cleaning Tools

As part of our project and practical recommendations for use, we compiled a list of digital tools for Collections as Data projects.

- OpenRefine
- Notepad++
- Microsoft Excel
- Python
- DigiPres Commons
- Voyant Tools
- Tableau Public
- Gephi
- R
- ArcGIS
- StoryMarJS
- TimelineJS
- D3.js
- Babylon.js
- Carto

Recommendations

- Choose a collection you would like to explore. Ask the question: Why do you want to analyze this collection computationally?
- Start where you are and break the process into small, manageable tasks so that the work doesn't feel overwhelming.
- Identify the format of your data (XML, TIFF, PDF, Excel, etc.) and the level of technical expertise required to work with the format.
- Determine whether your data needs to be cleaned. If it does, preserve the contexts of data.
- Select the appropriate tools to analyze and visualize your data. Different stages of analysis (collection, cleaning, etc.) may require different tools.
- Provide open access to data and documentation.
- Seek out digital humanities and digital scholarship practitioners in the field and reach out to IT professionals within your institution for collaboration.

Selected References

Padilla, T. (2017). On a collections as data imperative. UC Santa Barbara. https://escholarship.org/uc/item/9881c8sv.

Padilla, T., Allen, L., Frost, H., Potvin, S., Russey Roke, E., & Varner, S. (2019, May 20a). 50 Things --- Always already computational: Collections as data. Zenodo. http://doi.org/10.5281/zenodo.3066237.

Shensky, M. (2020). Data & Donuts: Schedule. Retrieved November 21, 2020, from https://guides.lib.utexas.edu/data-and-donuts/schedule.

Wittmann, R., Neatrour, A., Cummings, R., & Myntti, J. (2019). From digital library to open datasets: Embracing a "collections as data" framework. *Information Technology and Libraries*, *38*(4), 49-61. https://doi.org/10.6017/ital.v38i4.11101.

Acknowledgements

Special thanks to Melanie Cofield, Head of Access Systems at UT Libraries and Digital Libraries Professor at the iSchool, to Lea DeForest, Communications Manager at Texas Digital Libraries, to Alex Suárez, Administrative Associate at Texas Digital Libraries, to David Bliss, Systems and Digital Archivist at UT Library's Stewardship Unit, and to Aleshka Blay, Real State Associate.