# From Collections to Data
## Assessing Opportunities to Transform Archival Materials into Data at University of Washington Special Collections
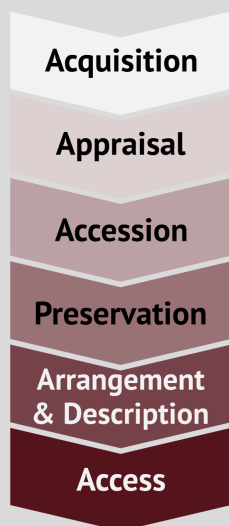
Hannah E. Robinson, MA

## Abstract

"From Collections to Data" analyzes University of Washington Special Collections (UWSC) holdings that can be used to develop, describe, make accessible, and support reuse of data derived from collection materials. UWSC has several collections that are potential candidates for computational research, but the organization does not currently have the means, personnel, or infrastructure to make the data accessible to users. The primary deliverable of this project is a report that suggests how UWSC can begin to provide access to datasets and comprises of three main sections: (1) background on the Collections as Data (CaD) movement through a case study analysis of cultural heritage institutions, (2) recommendations for developing a CaD initiative at UWSC, and (3) suggested next steps for implementing a CaD project in UWSC.

## Collections as Data

Collections as Data (CaD) is an Institute of Museum and Library Services project that "aims to foster a strategic approach to developing, describing, providing access to, and encouraging reuse of collections that support computationally-driven research." A main driving force behind the CaD program is the problem that cultural heritage institutions are digitizing thousands and thousands of materials and making them available online, but the materials are not much use for computational purposes. Users can browse individual digital representations of materials but they cannot download object metadata in large quantities. This inability to download metadata in bulk limits the research potential of digitized collection materials. UW Special Collections has the opportunity to increase discoverability and use through a CaD initiative.

## Standard Archival Workflow

- Acquisition
- Appraisal
- Accession
- Preservation
- Arrangement & Description
- Access

## Changes with a Collections as Data Project

**Arrangement & Description**
- Use globally unique persistent identifiers
- Apply standard metadata schemes like MODS, METS, CIDOC-CRM
- Engage community members in transcription, tagging, and description
- Provide public explanation of provenance including curatorial efforts made and materials restricted

**Access**
- Make materials available with an open license, when possible
- Provide data in non-proprietary formats
- Build access portals for datasets
- Include guides for how to use and interpret data

## Potential Collections for Data Projects

| Collection | Description of Data | Potential Curation Activities |
|---|---|---|
| Digitized oral history collections | 978 digitized audio files<br>645 have text transcripts | • Curate thematic sets<br>• Segment/tag audio<br>• Transcribe remaining oral histories |
| John Stamets photographs of Seattle construction projects | 33 TB of digital photographs | • Batch edit images to lower resolution<br>• Assign geographic metadata to images |
| Paul R. Brass papers on the India partition elections | Approx. 40 cubic feet of dot matrix printouts | • Digitize print outs<br>• Convert from PDF/image to tabular data |

## Next Steps

- Form partnerships with digital humanities scholars, other units and departments in UW Libraries, and technical experts
- Consult with community stakeholders and staff to select the best collection for a CaD pilot project
- Share knowledge with other archival institutions who have worked on data projects to gain insights
- Foster community, campus, and professional engagement with archival collections data