Collection Size Descriptions as Archival Data: The Spectrum of <physdesc>

Sarah Buchanan and Haoyang Li {sarahab, haoyang.li}@utexas.edu

The University of Texas at Austin, School of Information

Visualizing Archival Data (VADA) in the Augmented Processing Table (APT)

The APT project studies how evidence is revealed through archival processing and archivists' creation of a descriptive document for an archival collection, known as a *finding aid*. Aggregation and study of elements in these finding aids can reveal trends regarding arrangement styles at the collection, repository, and region level as well as over time. Such findings can inform the teaching of arrangement in archival education as well as use and development of this standard with regard to data quality. Analyzing these finding aids through our system VADA, a visual analytic tool, facilitates research in archival discovery.

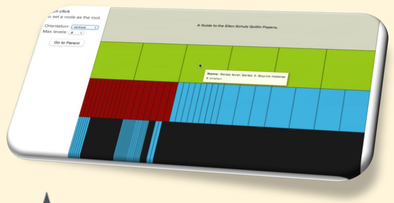
http://everest.ischool.utexas.edu/apt/

Abstract

The Physical Description element in EAD 2002, expressed as a <physdesc> tag, is one of 146 elements in this data structure standard. <physdesc> occurs within the high-level Collection Summary <did>, both of which are required elements in an EAD finding aid[1]. Overall this element captures both materiality and quantity, either as a collection whole or as separate extents. Encoding practices in <physdesc> permit the information to be presented as *plain text* or as parsed into four subelements[2]. The variation we discovered within <physdesc> thus impedes normalization of collection sizes (unit + value) into actionable data for quantitative analyses.

[1] Per: *RLG Best Practice Guidelines for EAD* (2002) and *EAD Best Practices at the Library of Congress* (2008).
[2] Four optional subelements: <dimension>, <extent>, <genreform>, <physfacet>.

Visualization of a collection arrangement:



Right: A sample of material types within <physdesc> which reveals many "outliers." Additionally:

- Unclear separation by comma, (semi)colon, ().
- Narrative sentences of unit descriptions.
- Mixtures of format types.
- Values expressed as decimals & fractions.

<physdesc> in EAD3, Gamma Release

Understanding "collection size" is complicated by the presence of variation both within the standard "box" unit as well as with many *outlier* units. The Physical Description element will remain in EAD3, but as an unstructured option that will only accept generic mixed content. A new <physdescstructured> element will adopt, rename, and add a fifth optional subelement[3]. These changes make motion toward consistency over flexibility, but the realities of permitted inconsistencies hinder the use, exchange, and analysis of EAD as structured data. Version compatibility and integration of the legacy data corpus (120K documents in ArchiveGrid; 8.7K in TARO) will be essential to support researchers discovering archival collections through EAD aggregators and federated searches.

[3] Five optional subelements: <quantity>, <unittype>, <dimensions>, <physfacet>, <descriptivenote>.

```
<physdesc label="Extent">
  <extent>Box (flat boxes, medium-sized flat boxes,
  storage boxes, small photo boxes, oversized document
  box, record storage boxes, archival boxes, manuscript
  box, card box, periodical box); Portrait, Reels of
  microfilm, Galley folders, Coins, Pages, ZIP disks,
  Posters, Artifact, Halftones, Negatives (Minox, Glass
  plate), Flat files, Scrapbooks, Trophy; Tubes; Ledgers;
  Cassette tapes; Diary; Woodblock; Videotapes. </extent>
</physdesc>
```

TCDL 2014 Austin * April 28-29

Texas Conference on Digital Libraries