Where Have All the Collections Gone?

Hugh Paterson III

Open Language Archives Community Scholar

Poster presented at the 15th Annual Society of American Archivists Research Forum

July 21st, 2021

1. Introduction

The Open Language Archives Community (OLAC) aggregator currently compiles 443,217 records from 65 providers. Participating archives each provide Dublin Core metadata via an OAI feed. Based on the needs of both linguists and language community members, Watson et al. (2016) note that usability requirements are not met by language-archive records. Burke and Zavalina (2019 & 2020) established that record composition for the free-text description field is used in various ways across the 65 OLAC-participating archives they evaluated. Some of those free-text description fields indicated that the record which was indicated to be an item/artifact was in fact more a collection of items or artifacts. However, collection records should have a different composition from individual artifact records because they each have distinct scopes. With this in mind, different record types should have distinct evaluation criteria, when compared with artifact records. For example, collection records should link to the records of items in the collection, and thereby support the browsing of collections within an aggregator (Zavalina 2011). Unexplored in the literature are how record providers are utilizing distinct collection records. Existing evaluations of OLAC records do not take record types into account (Hughes 2004).

The current study explores the use of the DCMIType “Collection” to indicate collection records among OLAC-participating archives. Across the OLAC records, 850 use the DCMIType “Collection” and only 7 providers even use the “Collection” DCMIType. By using the DCMIType “Collection” and relating artifact records with collection records via the Dublin Core “hasPart” property, more about the original context of the collection is transferred from the host institution to the OLAC aggregator. When properly displayed this can lead to increased utility in browsing environments.

2. Open Language Archives Community Contributors and DCMITypes

65 Data providers share data catalogues of language and culture resources via a central aggregator. 443,217 Records are shared via OAI PMH using the Dublin Core, Dublin Core Terms, and OLAC name spaces. 369,520 Records include a DCMIType in their record.

7 Providers use DCMIType “Collection”. 850 OLAC Records use the DCMIType “Collection”.

DCMIType “Collection” indicates:

Archival Collections Aggregate Works

3. OLAC Collections

<table>
<thead>
<tr>
<th>Institution</th>
<th>Reported Collections</th>
<th>Archival Collections</th>
<th>Aggregate Works</th>
<th>Mislabelled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech and Language Data Repository (SLDR/ORTOLANG)</td>
<td>23</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bavarian Institute for Speech Science (BAS)</td>
<td>53</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Graduate Institute of Applied Linguistics Library</td>
<td>155</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multilingual Learning and Teaching Corpus Exchange</td>
<td>49</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pacific Collection at the University of Hawai’i at Manoa</td>
<td>272</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hamilton Library</td>
<td>168</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collections of Ortiz Oracu Numeriques (CoCOn)</td>
<td>163</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)</td>
<td>99</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. Data Providers not Contributing Collection Records

When ranked by number of records provided to OLAC, the top 25 contributors provide over 98% of the records. Only 4 providers among these 25 provide collection records. The top three data providers do not provide any collection records. Together the largest three archives represent over 66% of all records.

5. OLAC Non-Collections

If it is possible to estimate the number of missing collection records for archival collections? Rough estimations indicate that there are over 7,816 records which should have the DCMIType “Collection” applied, and 3,086 “mislabeled” collection records.

<table>
<thead>
<tr>
<th>Institution</th>
<th>Should Be Labelled as a Collection</th>
<th>Percentage of OLAC Records</th>
<th>OLAC Contributor Rank Based on Number of Records</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Rosetta Project: A Long Now Foundation Library of Human Language</td>
<td>59</td>
<td>3.79%</td>
<td>1</td>
</tr>
<tr>
<td>Endangered Languages Archive</td>
<td>4,005</td>
<td>41.14%</td>
<td>2</td>
</tr>
<tr>
<td>The Language Archive</td>
<td>1,467</td>
<td>11.17%</td>
<td>3</td>
</tr>
<tr>
<td>Land University Humanities Lab corpusserver</td>
<td>2,285</td>
<td>3.27%</td>
<td>6</td>
</tr>
</tbody>
</table>

6. Conclusions

These data support claims that:
1. User interfaces to language archives present user friction (because whole-part/collection-component relationships are not effectively communicated to OLAC web-users);
2. Record descriptions are not consistent (collection description information is compressed into artifact records);
3. Language archives are not using consistent frameworks for collection description such as DACS.

Over all we should expect a greater number of records with the DCMIType “Collection” within the OLAC aggregator. These collection records should link to the records for the constituent parts of the collection. We should anticipate that there is a broad range of content for which the DCMIType “Collection” is appropriate. “Collection” records would include records for archival collections, records of periodicals (but not their articles, unless they were aggregate works), and aggregate works. This leads me to ask: “where have all the collections gone?”

7. References