Where Have All the Collections Gone?

Hugh Paterson III

Unaffiliated Collaborative 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, Wasson 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 three OLAC participating archives they evaluated. Some of these free-text description fields indicated that the record which was indicated to be an item/artifact was in fact more like 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 reocrds 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

Categorization of Reported Collections

Institution	Reported Collections	Archival Collections	Aggregate Works	Mislabelled
The Sociolinguistic Archive and Analysis Project (SLAAP)	36	☑		
Speech and Language Data Repository (SLDR/ORTOLANG)	23	☑		
Bavarian Archive for Speech Signals (BAS)	53	☑		
Graduate Institute of Applied Linguistics Library	155			☑
Multimodal Learning and teaching Corpora Exchange	49	☑		
Pacific Collection at the University of Hawai'i at Mānoa Hamilton Library	272		☑	
COllections de COrpus Oraux Numeriques (CoCoON ex-CRDO)	163	☑		
Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)	99		☑	

4 Data Providers ***not*** Contributing Collection Records

When ranked by number of records provided to OLAC, the top 25 continbutors 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.

Institution	Number of Records	Percentage of OLAC Records	OLAC Contributor Rank Based on Number of Records
The Language Archive	149,763	33.79%	1
Endangered Languages Archive	93,687	21.14%	2
SIL Language and Culture Archives	49,494	11.17%	3
California Language Archive	14,959	3.37%	6
Lund University Humanities Lab corpusserver	12,266	2.77%	7

5 OLAC Non-Collections

Is it possible to esitmate the number of missing collection records for archival collections? Rough esitmations indicate that there are over 7,816 records which should have the DCMIType “Collection” applied, and 1,086 “missing” collection records.

Institution	Should Be Labelled as a Collection	
The Rosetta Project: A Long Now Foundation Library of Human Language	59	Alan Lomax Collection records
Endangered Languages Archive	4,005	Search for “Collection” in records
The Language Archive	1,467	Search for “Collection” in records
Lund University Humanities Lab corpusserver	2,285	Search for “Collection” in records

Institution	Missing Archival Collection Records	
Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)	454	URL counts
Endangered Languages Archive	418	Duplicate title tag differentiation
Kaipuleohone	186	Search for “Collection” in records or locate files titled “collection description”

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 consitutent parts of the collection. We should anticipate that there is a broad range of content for which the DCMIType “Collection” is appropiate. “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

- Burke, Mary, and Oksana Zavalina. 2019. “Exploration of Information Organization in Language Archives.” *Proceedings of the Association for Information Science and Technology* 56 (1): 364–67. doi:10.1002/pra2.30.
- Burke, Mary, and Oksana L. Zavalina. 2020. “Descriptive Richness of Free-text Metadata: A Comparative Analysis of Three Language Archives.” *Proceedings of the Association for Information Science and Technology* 57 (1):e429. doi:10.1002/pra2.429.
- Hughes, Baden. 2004. “Metadata Quality Evaluation: Experience from the Open Language Archives Community.” In *Digital Libraries: International Collaboration and Cross-Fertilization*, edited by Zhaoneng Chen, Hsinchun Chen, Qihao Miao, Yuxi Fu, Edward Fox, and Ee-peng Lim, 320–29. Lecture Notes in Computer Science 3334. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-540-30544-6_34.
- Paterson III, Hugh. 2021. OLAC Nightly Data Dump (XML) from 18 July 2021 [Data set]. Zenodo. doi:10.5281/zenodo.5112131
- Society of American Archivists. 2013. Describing Archives: A Content Standard. 2nd ed. Chicago, Illinois: Society of American Archivists. http://files.archivists.org/pubs/DACS2E-2013_v0315.pdf.
- Wasson, Christina, Gary Holton, and Heather S. Roth. 2016. “Bringing User-Centered Design to the Field of Language Archives.” *Language Documentation & Conservation* 10 (December): 641–81. http://hdl.handle.net/10125/24721.
- Zavalina, Oksana L. 2011. “Contextual Metadata in Digital Aggregations: Application of Collection-Level Subject Metadata and Its Role in User Interactions and Information Retrieval.” *Journal of Library Metadata* 11 (3–4). Routledge: 104–28. doi:10.1080/19386389.2011.629957.

Suggested Poster Reference:
Paterson III, Hugh J. 2021. “Where Have All the Collections Gone?” Poster presented at the 15th Annual Society of American Archivists Research Forum. 21 July, 2021.