OLAC and Serials: An Appraisal

Hugh Paterson III
i@hp3.me
University of North Texas & University of Oregon
Denton, Texas, USA

ABSTRACT

This paper reports on how journal articles are presented within the Open Language Archive Community's (OLAC) OAI-PMH aggregator for language resources. It discusses metadata record composition across data providers. The conceptual category of "Language resource" is a broad agglomeration including original creative works captured in handwritten, audio, and video mediums, annotations to the raw captures, and analysis of those annotations. Discovery of language resources is a challenge given the diversity of resource origins. Original creative works and annotations are products often available via archives while analysis, theory, and advice are often released via formal publishing venues such as journals. Scholars benefit from a view where resources from various release sources can be displayed with their inter-resource relationships, e.g., source material and analysis. Understanding how secondary journal materials are presented in OLAC records is a first step towards increasing the end-user utility of the OLAC aggregator.

CCS CONCEPTS

• **Applied computing** → **Document metadata**; • **Information systems** → **Digital libraries and archives**; **Database administration**.

KEYWORDS

OAI-PMH, Dublin Core, Journals, Serials, Catalogs, Open Language Archives Community, OLAC

1 INTRODUCTION

The Open Language Archives Community (OLAC) aggregator is a web service which combines and re-presents the catalogs of over 60 data providers [2]. It was originally conceived of as an aggregator for resources 'in and about languages' including references to *advice*, *data*, and *tools* [11]. It is unique among aggregators in that it offers a view, by language, of stewarded resources. This view is especially beneficial to language scholars and language users who seek out language resources for research and educational purposes. End-users benefit from visualizations presented during the discovery process which overtly connect original media resources demonstrating language-use to analysis and advice which is often contained within formally published resources discussing said media. In 2022, the OLAC aggregator contained nearly 449,000 entries [8]. Best estimates show that only 0.4 percent of those catalog records represent journal articles. This suggests that there is still much work left to do to implement the original vision laid out in the OLAC documents [11]. In 2022, an initial analysis was conducted on how journal articles and serial works were presented within aggregated records. This was done to prepare for ongoing work related to making more published resource records available via OLAC, thereby contributing to its original vision. The research objective was not to discover all the journal articles present within the OLAC record set, but rather to investigate the diversity in how they were recorded within the OLAC metadata application profile.

2 METHODS

The goal of this study, using data collected in 2022 and 2023, was to search OLAC records for the purpose of documenting how different data providers were reporting journal articles. To investigate records the OLAC-provided full-text, faceted search tool was used.[1] The search apparatus at OLAC is not case sensitive. Three investigative terms were chosen due to their semantic relationship in English. The terms were *journal*, *article*, and *serial*. The count for each term was recorded for each contributing data-provider. Counts are provided in Section 3. The returned results were then manually qualitatively assessed for relevance. The search terms used in this study overlap with terms-of-art within linguistics. For example, *serial* is used in the context of *serial verb construction*, and *article* is a term for a category of words which generally introduce a noun phrase such as the English words: *a*, *an*, and *the*. The manual review process produced a smaller set of records. Select examples from this smaller set are then discussed in Section 4.

2.1 Reproducibility

The methods employed in this study are not significantly complex and therefore easily reproducible. However, the exact results will vary as the aggregator collects more records. No data capture for the comprehensive set of search results was attempted. However, records discussed in Section 4 were captured, committed to a .git repository, and submitted to Zenodo [9]. While no back up copy of the searched records or comprehensive OLAC data dump from the time of the investigation exists, scholars may be interested in a comprehensive OLAC data dump from 2021 available via Zenodo [7].

2.2 Known Resources

The specific search method was chosen even though there is one data-provider, the journal *Language Documentation & Conservation* (LD&C), which provides over 1500 article records to OLAC. Additionally, SIL International's *Language & Culture Archives'* (L&CA) OAI-PMH feed provides records from several of SIL's serial publications as well as many records of journal publications by SIL affiliated authors. Data from these sources were not excluded from results, but the goal of the investigation was to find records which reference or represent serials across as many data providers as possible.

[1] http://search.language-archives.org

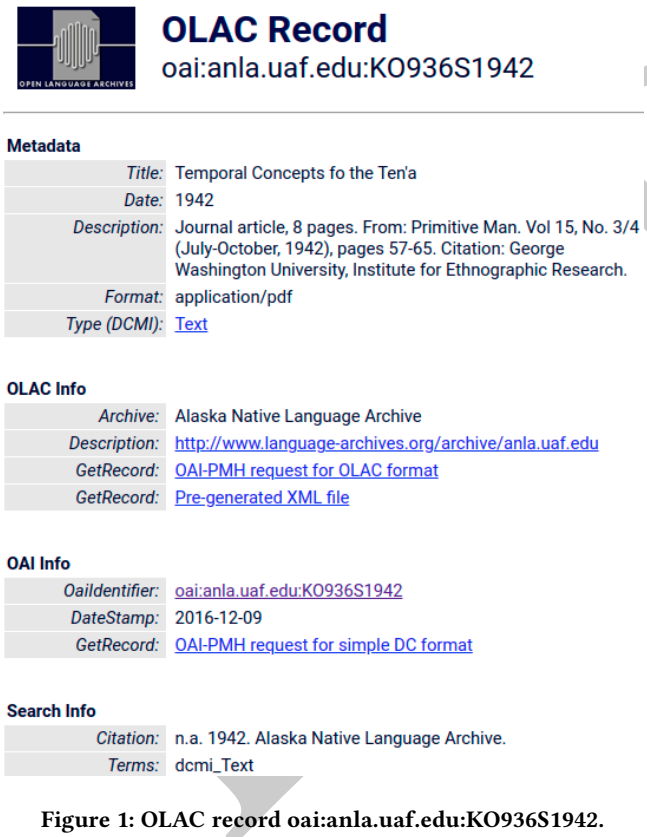
3 DATA

3.1 Summary Tables

For the sake of space the data is only partially presented in this paper. Over three hundred records were viewed in the investigation. Two summary tables are provided via Zenodo [9]. Table A provides the quantitative results by OLAC data-provider for each of the search terms. Table B presents a short summary of the kinds of things recovered from each of the data providers for that search term.

3.2 Examples

The three example records replicated here were drawn from the investigation. Their full XML records are available via Zenodo [9]. Figure 1 presents a record for a journal article cataloged by the *Alaska Native Language Archive*. Figure 2 presents a record for a journal article by the L&CA. In this case SIL International is the publisher of the journal through their Dallas, Texas, based publishing unit. Figure 3 is the record of a journal article published by LD&C via the University of Hawai'i Press.



OLAC Record
oai:anla.uaf.edu:KO936S1942

Metadata

Title:	Temporal Concepts fo the Ten'a
Date:	1942
Description:	Journal article, 8 pages. From: Primitive Man. Vol 15, No. 3/4 (July-October, 1942), pages 57-65. Citation: George Washington University, Institute for Ethnographic Research.
Format:	application/pdf
Type (DCMI):	Text

OLAC Info

Archive:	Alaska Native Language Archive
Description:	http://www.language-archives.org/archive/anla.uaf.edu
GetRecord:	OAI-PMH request for OLAC format
GetRecord:	Pre-generated XML file

OAI Info

OaiIdentifier:	oai:anla.uaf.edu:KO936S1942
DateStamp:	2016-12-09
GetRecord:	OAI-PMH request for simple DC format

Search Info

Citation:	n.a. 1942. Alaska Native Language Archive.
Terms:	dcmi_Text

Figure 1: OLAC record oai:anla.uaf.edu:KO936S1942.

4 DISCUSSION

Thirty-one of the sixty-plus OLAC data providers have records within the search parameters. There is a significant amount of diversity in the structure of records representing or referencing journal articles. Several re-occurring inconsistencies persisted across

OLAC Record
oai:sil.org:40239

Metadata

Title:	Translating "Messiah," "Christ," and "Lamb of God"
Contributor (author):	King, Phil
Date (W3CDTF):	2005
Description (URI):	https://www.sil.org/resources/archives/40239
Extent:	pages 1-27
Format (IMT):	application/pdf
Identifier (URI):	https://www.sil.org/resources/archives/40239 https://doi.org/10.54395/jot-tc4pm
Is Part Of (URI):	oai:sil.org:40276
Language:	English
Language (ISO639):	eng
Publisher:	SIL International
Relation:	Is Part of Series: Journal of Translation 1(3)
Subject:	Translation Relevance Theory
Subject (OLAC):	translating_and_interpreting
Type (DCMI):	Text

OLAC Info

Archive:	SIL Language and Culture Archives
Description:	http://www.language-archives.org/archive/sil.org
GetRecord:	OAI-PMH request for OLAC format
GetRecord:	Pre-generated XML file

OAI Info

OaiIdentifier:	oai:sil.org:40239
DateStamp:	2022-11-17
GetRecord:	OAI-PMH request for simple DC format

Search Info

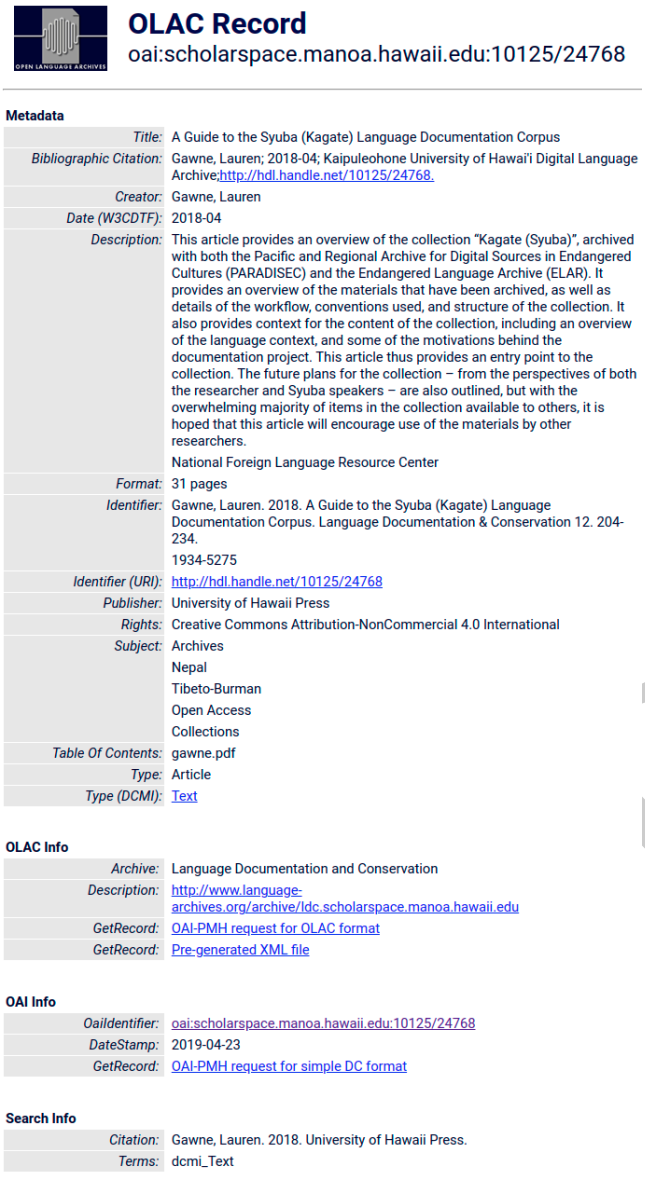
Citation:	King, Phil. 2005. SIL International.
Terms:	area_Europe country_GB dcmi_Text iso639_eng olac_translating_and_interpreting

Figure 2: OLAC record oai:sil.org:40239.

records related to the completeness and appropriate semantics of metadata element usage. In the following sub-sections I briefly address the usage of the description field, source relationships, and part-whole relationships. Significant other inconsistencies involved the following elements and are the subject of ongoing investigation: `dcterms:bibliographicCitation`, `dc:title`, `dc:contributor`, `dcterms:format`, `dcterms:extent`. These inconsistencies disrupt end-user continuity for the OLAC discovery experience.

4.1 Description Field

Discontinuity in metadata semantics can be observed when comparing the three selected records for journal articles. The journal article



OLAC Record
oai:scholarspace.manoa.hawaii.edu:10125/24768

Metadata

Title:	A Guide to the Syuba (Kagate) Language Documentation Corpus
Bibliographic Citation:	Gawne, Lauren; 2018-04; Kaipuleohone University of Hawai'i Digital Language Archive;http://hdl.handle.net/10125/24768.
Creator:	Gawne, Lauren
Date (W3CDTF):	2018-04
Description:	This article provides an overview of the collection "Kagate (Syuba)", archived with both the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) and the Endangered Language Archive (ELAR). It provides an overview of the materials that have been archived, as well as details of the workflow, conventions used, and structure of the collection. It also provides context for the content of the collection, including an overview of the language context, and some of the motivations behind the documentation project. This article thus provides an entry point to the collection. The future plans for the collection – from the perspectives of both the researcher and Syuba speakers – are also outlined, but with the overwhelming majority of items in the collection available to others, it is hoped that this article will encourage use of the materials by other researchers.
	National Foreign Language Resource Center
Format:	31 pages
Identifier:	Gawne, Lauren. 2018. A Guide to the Syuba (Kagate) Language Documentation Corpus. Language Documentation & Conservation 12. 204-234.
	1934-5275
Identifier (URI):	http://hdl.handle.net/10125/24768
Publisher:	University of Hawaii Press
Rights:	Creative Commons Attribution-NonCommercial 4.0 International
Subject:	Archives
	Nepal
	Tibeto-Burman
	Open Access
	Collections
Table Of Contents:	gawne.pdf
Type:	Article
Type (DCMI):	Text

OLAC Info

Archive:	Language Documentation and Conservation
Description:	http://www.language-archives.org/archive/ldc.scholarspace.manoa.hawaii.edu
GetRecord:	OAI-PMH request for OLAC format
GetRecord:	Pre-generated XML file

OAI Info

OaiIdentifier:	oai:scholarspace.manoa.hawaii.edu:10125/24768
DateStamp:	2019-04-23
GetRecord:	OAI-PMH request for simple DC format

Search Info

Citation:	Gawne, Lauren. 2018. University of Hawaii Press.
Terms:	dcmi_Text

Figure 3: OLAC record oai:scholarspace.manoa.hawaii.edu:10125/24768.

record shown in Figure 2 is provided by the L&CA, for an article appearing in the *Journal of Translation*. The description field contains a URL. The field is qualified with an invalid qualifier: `dc:terms URI`. Neither OLAC documentation [12] nor the Dublin Core documentation [5] have any indication that the `dc:description` field can be qualified with a URI. In contrast, the description field in Figure 3 provides something like an abstract (the data-provider doesn't qualify the description). Both of these records contrast with the description field from Figure 1, which has various kinds of bibliographic content in the description.

Within the `dc:description` field of the record shown in Figure 1, one can find the article's contributor, genre type, extent, and most of the elements needed for a bibliographic citation. No content oriented description is provided in the description field. For readability the description field is replicated in Figure 4.

```
1  <dc:description>Journal article, 8 pages. From:
   ↪ Primitive Man. Vol 15, No. 3/4 (July-October,
   ↪ 1942), pages 57-65. Citation: George Washington
   ↪ University, Institute for Ethnographic Research
   ↪ .</dc:description>
```

Figure 4: Description field from Figure 1.

4.2 Source Relationships

An important element of this inquiry was to investigate how journal articles were related to the language resources which motivated their creation via overt metadata relationships. The record for the resource presented in Figure 3 is the classic example. The journal article described is a guide to a specific archival collection of language resources stewarded by the *Endangered Language Archive* (ELAR). ELAR happens to also be an OLAC data contributor. The OLAC record does mention in the description field that the collection is deposited at ELAR, but there is no hyperlink between the OLAC record and the OLAC record for the ELAR deposit, or even between the OLAC record for the journal article and the deposit profile on the ELAR website. The broader finding applicable to records from all data providers is that no records for journal articles contained, `dcterms:isReferencedBy`, `dc:source`, or `dcterms:References` relationships. These are the kinds of relationship fields in which one would expect to find declared links between publications and their source or supporting materials.

4.3 Container Relationships

Relationships play a significant role in positioning journal articles within discovery systems. The metadata fields discussed in Section 4.2 facilitate discovery on the basis of related source context, but this is not the only important relationship to consider. The record in Figure 2 illustrates a different type of relationship which was only found in records by L&CA but is extremely important for the discovery of serial resources (code snippet show in Figure 5). This is the part-whole/whole-part relationship which is also sometimes known as the part-container relationship. Serials vary by how many levels of whole-part relationship they exhibit. Some serial patterns have optional components such as *volumes* in the pattern *Series-Book-(Volume)-Chapter*, while others have patterns with optional *issues* such as *Journal-Volume-(Issue)-Article*.

```
1  <dcterms:isPartOf xsi:type="dcterms:URI">oai:sil.org
   ↪ :40276</dcterms:isPartOf>
```

Figure 5: Part-whole relationship indication in record from Figure 2.

In contrasting the records illustrated in Figures 2 and 3, one can also see that Figure 3 with the article appearing in LD&C contains an *International Standard Serial Number* (ISSN) identifier.

This identifier is for the journal or serial and applies to the whole entity, rather than the part entity. The data feed for LD&C does not include any container records (e.g., volume, issue, journal), whereas the L&CA feed only includes volume records, and then only in some cases. The L&CA does not supply records for the whole journal/serial. The result is that L&CA records have a relation field with a complex container identifier in plain text, rather than a link to a full record. The absence of declared part-whole relationships across many records impacts the ability of metadata consuming services to dynamically create record and navigation interfaces.

5 CONCLUSIONS

The data show that there is significant diversity among the records representing serials. Even though there is a low volume level of records compared to the total number of records, resources appearing in serials have not been a traditional focus of the current OLAC community. The absence of any formal guidance via the OLAC metadata application profile to address serial publications including their part-whole and source-analysis components has left data providers to their own devices. Record consistency and completeness could be improved. Formally adding a best practice recommendation to the OLAC application profile which addresses relationship metadata would improve the ability for end-users to navigate complex relationships between resources cataloged and held by different institutions. Figure 6 illustrates a model which does not require the addition of any elements or vocabularies to the OLAC metadata profile. It simply lays out that green and gray boxes need individual records and need to contain relationships already provided via the foundation upon which OLAC is built [1]. This stands in contrast to numerous other claims regarding the insufficiency of Dublin Core to describe journal articles [4, 6, 13].

The diversity in how journal articles and other serial-contained resources are cataloged presents a challenge to the end-user discovery of language resources. One approach towards reaching coherent data-provider behavior across the OLAC community is to release an OLAC best practice recommendation for documents published in serials. This would bring a more consistent discovery experience to end-users.

Journal, volume, and issue container-records should be marked with the DCMIType *collection*. This study found that these container records were absent from OLAC in most cases, and where provided, fail to provide the DCMIType *collection*. Therefore, observations of their absence in this study support previously reported observations that collection records are under-reported in OLAC [10].

Other research [3], sourcing its data directly from OLAC data providers rather than the OLAC aggregator, suggests that records with longer descriptions are higher quality. Due to various semantic inconsistencies it is not clear that this obtains within the OLAC data set for journal articles. Further investigation is needed to determine if metadata crosswalks from data providers to OALC are inducing errors.

The utility of a digital library to its end-users is directly related to how it meets their knowledge acquisition goals. Significant in this process for users is how they engage with the discovery process. Therefore not only is the materiality of the user-interface important

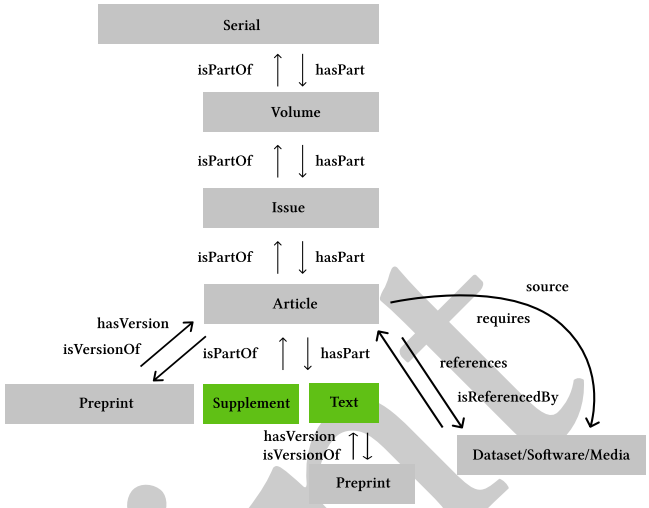


Figure 6: Dublin Core compliant model for serials in OLAC.

to end-users but also the materiality of the objects for which they are searching.

REFERENCES

[1] Steven Bird and Gary F. Simons. 2004. Building an Open Language Archives Community on the DC Foundation. In *Metadata in Practice*, Diane Hillmann and Elaine L. Westbrooks (Eds.). American Library Association, Chicago.

[2] Steven Bird and Gary F Simons. 2021. Towards an Agenda for Open Language Archiving. In *Proceedings of the International Workshop on Digital Language Archives: LangArc 2021*, Oksana Zavalina and Shobhana Lakshmi Chelliah (Eds.). University of North Texas, Denton, Texas, 25–28. https://doi.org/10.12794/langarc1851171

[3] Mary Burke 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 (Oct. 2020). https://doi.org/10.1002/pra2.429

[4] Assumpció Estivill, Ernest Abadal, Jorge Franganillo, Jesús Gascón, and J. M. Rodríguez Gairín. 2005. Use of Dublin Core Metadata for Describing and Retrieving Digital Journals. In *International Conference on Dublin Core and Metadata Applications*. DCMI, Madrid, Spain, 137–140. https://dcpapers.dublincore.org/pubs/article/view/812

[5] Diane I. Hillmann. 2005. Dublin Core Qualifiers. In *Using Dublin Core*. DCMI, Chapter specifications, Section 5. https://www.dublincore.org/specifications/dublin-core/usageguide/qualifiers/

[6] Wayne Jones. 2001. Dublin Core and Serials. *Journal of Internet Cataloging* 4, 1-2 (Nov. 2001), 143–148. https://doi.org/10.1300/J141v04n01_13

[7] Hugh J Paterson III. 2021. *OLAC Nightly Data Dump (XML) from 18 July 2021*. https://doi.org/10.5281/zenodo.5112131

[8] Hugh J Paterson III. 2022. An OLAC Perspective on Services: The Forgotten Language Resources. In *Proceedings of DC-2022*. Dublin Core Metadata Initiative, Online, Pre–print. https://hughandbecky.us/Hugh-CV/talk/2022-services-the-forgotten-language-resources/DC_2022_Conference_Paper_Paterson_Revisions_pre-print.pdf

[9] Hugh J Paterson III. 2022. *Supporting Evidence For OLAC and Serials: Publication Support Tables*. https://doi.org/10.5281/zenodo.7049203

[10] Hugh J Paterson III. 2022. Where Have All the Collections Gone?: Analysis of OLAC Data Contributors' Use of DCMIType 'Collection'. In *Proceedings of the 15th Annual Society of American Archivists Research Forum, 21 July, 2021*. Society of American Archivists, Chicago, IL. https://www2.archivists.org/am2021/research-forum-2021/agenda#peer

[11] Gary F. Simons and Steven Bird. 2000. The Seven Pillars of Open Language Archiving: A Vision Statement. In *Workshop on Web-Based Language Documentation and Description*. Open Language Archive Community, Philadelphia, PA. http://www.language-archives.org/documents/vision.html

[12] Gary F. Simons, Steven Bird, and Joan Spanne (Eds.). 2008. *OLAC Metadata Usage Guidelines* (2008-07-11 ed.). Open Language Archive Community. http://www.language-archives.org/NOTE/usage-20080711.html

[13] Mike Taylor. 2010. Bibliographic Data, Part 2: Dublin Core's Dirty Little Secret. https://reprog.wordpress.com/2010/09/03/bibliographic-data-part-2-dublin-cores-dirty-little-secret/

Received 14 May 2023; revised 6 June 2023