# Towards an Agenda for Open Language Archiving

Steven Bird
*Charles Darwin University*
Australia
steven.bird@cdu.edu.au

Gary F. Simons
*SIL International*
United States of America
gary_simons@sil.org

***Abstract*—The Open Language Archives Community (OLAC) provides a comprehensive infrastructure that has allowed our community to index and discover language resources over the past 20 years. However, OLAC infrastructure has fallen behind as the digital libraries community has continued to evolve. New investment is required in order to move OLAC into the digital libraries mainstream. This paper reports on the first 20 years of OLAC and on an agenda leading to a more sustainable future for open language archiving.**

***Keywords—Russian as a heritage language; heritage speakers; language archives; oral history; bilingualism***

## I. Introduction

OLAC was founded in 2000 as an international partnership of institutions and individuals who are creating a world-wide virtual library of language resources by developing consensus on best current practice for the digital archiving of language resources, and developing a network of interoperating repositories and services for housing and accessing such resources. We take a language resource to be "any physical or digital item that is a product of language documentation, description, or development or is a tool that specifically supports the creation and use of such products" [29, p88].

OLAC infrastructure is built on Dublin Core metadata [14] and the Open Archives Initiative Protocol for Metadata Harvesting [17]. At the time of writing, OLAC catalogues over 440,000 items from 62 participating language archives (http://www.language-archives.org/archives). These items cover all of the living languages recognised by the ISO 639-3 standard (see Fig. 1; http://www.language-archives.org/documents/coverage.html). For the most recent month, we logged 8,600 record views on the OLAC site, with 2,172 click-throughs to individual archives (does not include traffic to the search service hosted at the University of Pennsylvania).

Users access the OLAC catalog in a variety of ways: via any search engine, since OLAC exposes everything as pages that Web crawlers can index; via faceted search which exploits the controlled vocabularies to give search with complete recall and precision (http://search.language-archives.org); via links from language-related sites like Ethnologue (https://ethnologue.com/language/aaa: see link to "Language Resources"); via services such as WorldCat, CLARIN, Linguist List which harvest OLAC metadata from the OLAC Aggregator (http://www.language-archives.org/cgi-bin/olaca3.pl ); by consuming the XML or RDF/XML nightly dumps of the entire OLAC metadata catalog (http://www.language-archives.org/xmldump/ListRecords.xml.gz; http://www.language-archives.org/static/olac-datahub.rdf.gz ); or by accessing the RDF/XML of any metadata record via HTTP content negotiation (http://www.language-archives.org/item/oai:paradisec.org.au:AA1-001)

Alongside this technical infrastructure, OLAC has a document infrastructure: defining OLAC metadata standards [23]; specifying processes around repositories [24]; and laying out the process for managing the document lifecycle through a Council and Board [25].

This paper reviews the first 20 years of OLAC and identifies new opportunities to support long-term growth and viability of open language archiving.

TABLE. 1. COVERAGE OF OLAC ITEMS INDEXED BY ISO 639-3 IN RELATION TO LANGUAGE SIZE

| Population range | Languages | OLAC has data | (%) | Items |
|---|---|---|---|---|
| 1-9 | 133 | 133 | 100 | 3,563 |
| 10-99 | 339 | 339 | 100 | 13,372 |
| 100-999 | 1,038 | 1,038 | 100 | 29,605 |
| 1,000-9,999 | 2,014 | 2,014 | 100 | 62,791 |
| 10,000-99,999 | 1,824 | 1,824 | 100 | 46,813 |
| 100,000-999,999 | 895 | 895 | 100 | 29,235 |
| 1,000,000-9,999,999 | 304 | 304 | 100 | 14,892 |
| 10,000,000-99,999,999 | 77 | 77 | 100 | 49,008 |
| 100,000,000-999,999,999 | 8 | 8 | 100 | 47,233 |
| Unknown | 277 | 277 | 100 | 7,909 |
| All living languages | 6,909 | 6,909 | 100 | 304,421 |
| Extinct languages | 626 | 599 | 96 | 7,247 |

## II. OLAC Vision for the Open Language Archiving

The original vision for OLAC was set out in a document entitled The Seven Pillars of Open Language Archiving [22]. According to this vision, the individuals who use and create language documentation and description are looking for three things: Data, information that documents or describes a language of interest; Tools, computational resources that facilitate creating or using language data; and Advice, help in knowing what data sources to rely on, what tools to use, and what practices to follow. Despite this need, potential users of language resources did not have ready access to the data, tools, and advice that they needed. We explored these shortcomings through a "gap analysis", as follows: some archives (e.g. Archive 1, in Fig. 1A) have a site on the Internet which the user is able to find, so the resources of that archive are accessible; other archives (e.g. Archive 2) are on the Internet, so the user could access them in theory, but the user has no idea they exist so they are inaccessible in practice; still other archives (e.g.

Archive 3) are not even on the Internet. There are potentially hundreds of archives (Archive $n$) that the user should know about. Finally, tools and advice reside in many places, and are not indexed in a way that allows users to discover them, or relate the available tools to the available data.

OLAC was established in order to address these issues. According to the vision, OLAC would do this by offering four things: Gateway, a single portal through which users can access all available data, tools, and advice; Metadata, uniform descriptions of all available data, tools, and advice; Reviews, peer evaluations of available data, tools, and advice; and Standards, processes and protocols that enable the operation of the gateway and ensure the quality of metadata and reviews. We then articulated an overall solution having the structure shown in Fig. 1B.

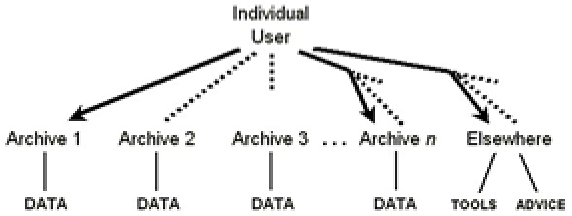

FIG. 1A. The Vision of the Open Language Archives Community: Gap Analysis for People Attempting to Access Language Archives

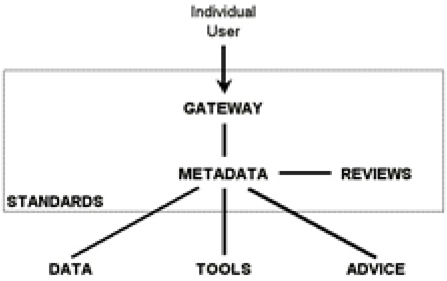

FIG. 1B. The Vision of the Open Language Archives Community: The Seven Pillars of Open Language Archiving

We have fleshed out this vision in further detail: requirements, for users, creators, archivists, developers and sponsors [21]; a survey of the state of the art in digital language documentation and description [3]; a later analysis with best practice recommendations [6]; and a white paper on establishing the infrastructure for open language archiving on the framework of Dublin Core Metadata and the Open Archives Initiative [4], subsequently implemented and reported in a series of publications [5, 19, 26, 27].

## III. TAKING STOCK OF OLAC TODAY

The present state of OLAC can be summarised as follows. The OLAC document process [25] has been established and used by the community in a series of workshops over several years to create many OLAC documents (http://www.language-archives.org/documents.html). The repositories and metadata standards have functioned continuously for 20 years [23, 24]. The community has used the OLAC Process to develop and refine vocabularies for linguistic data type [2]; discourse type [16]; contributor roles [15]; and linguistic field [1]. The most significant of these vocabularies is linguistic data type, though it only has three items corresponding to the Boasian trilogy [7]: lexicon, primary text, and language description. We specified the OLAC Language Extension [28] which standardised OLAC metadata to use ISO 639-3 codes for representing the names of languages. We developed a MARC to OLAC crosswalk [12]. We compiled best practice recommendations for the use of OLAC Metadata [31]. We established guidelines for metadata quality and provided automatic evaluation of quality and a quarterly report emailed to the repository coordinator, in order to motivate effort to improve metadata quality [18], an area where OLAC has been considered exemplary [11]. We have articulated sustainability conditions for language resources [29], chiefly, the conditions that ensure a resource will be usable—it is discoverable, available, interpretable, and portable. We have established and continue to maintain core infrastructure hosted at the Linguistic Data Consortium, and a search service hosted at the University of Pennsylvania Library (http://search.language-archives.org). Library.

Alongside these contributions of OLAC is the response from the community, including over 5,000 publications that cite OLAC (https://scholar.google.com/scholar?q="OLAC"+language). There is evidence that OLAC is enabling research that accesses language resources (e.g. [13]), and that OLAC is supporting ongoing scholarship on language archiving itself (e.g. [8, 9]).

Aside from these successes, there are various ways in which OLAC has not yet achieved our aspirations for long-term sustainability: the OLAC Council and Board have fallen inactive; the software infrastructure has not been refreshed in over a decade, and it is being maintained by volunteers and could fail catastrophically at any time (the website and search functionality would still operate, but new content coming from participating archives would not be harvested); of the 62 registered archives, 27 have not been updated in the past five years, and an overlapping 19 archives are failing to harvest. Also, the original vision for OLAC identified potentials which have not yet been realised: the indexing of tools and advice (cf. [34]), and using a formal document process in defining best practices in language archiving beyond resource description and discovery (cf. [6, 32]). More fundamentally, OLAC has not had the resources to keep up with current best practices of the digital libraries community. Funding has always been project-based. Advice from program managers has been that we add a research piece and compete for research funding, or that we objectively quantify the value of OLAC and seek infrastructure funding.

Since the founding of OLAC, the space for defining best practices in language archiving more broadly has been filled by the establishment of DELAMAN—the Digital Endangered Languages and Musics Archives Network (https://www.delaman.org/). We have initiated a process that is bringing OLAC under the governance structure of DELAMAN, with a narrowed scope of "developing consensus on best current practice for the interoperable description of archived language resources."

## IV. Towards an Agenda for Open Language Archiving

Much remains to be done across the space of language archiving [33]. In considering the opportunities offered by OLAC in particular, we begin with what OLAC already offers: a community that has grown up around the participating archives; a suite of documents that define OLAC's operation; a process for updating these documents; an archive registration process; an aggregation infrastructure; a federated search service; a focus on documenting subject language and linguistic data type in language resource metadata; and automated encouragement for archives to improve metadata quality.

In looking to the future, we envisage improvements in coverage. There are significant collections not yet participating, both archives and special collections within libraries. However, it is evident that implementing a data provider for OLAC metadata is too high a bar for some organisations. Some archives only expose an index page per language, and instead need to expose metadata for the individual resources so that they can be indexed centrally. Finally, scholars need to be able to report language resources they discover in places that would never join OLAC (such as isolated texts in endangered languages).

We also envisage improvements in access. Many archives need to improve metadata quality so as to improve the discoverability of their holdings. At the time of writing, 22 out of 62 archives score below 70% on OLAC's metadata quality metric. Subject language is only used in 65% of records. Linguistic Data Type is used in a mere 21% of records. In addition, sub-communities could make OLAC more directly relevant for themselves, by cataloguing holdings according to their own system, e.g.: <dc:type>Sociolinguistic corpus</dc:type>, <dc:format>text/x-eaf+xml</dc:format>.

Finally, we envisage mainstreaming language archives, by replacing our parochial metadata format with a generic application profile, effectively steering OLAC and the cataloging of language resources into the library and information systems mainstream. Observing the trend in library automation toward Linked Data in cataloging, we have taken a first step by mapping OLAC metadata for Linked Data [30]. We envision OLAC's idiosyncratic metadata format being superseded by an application profile [10] for describing language resources. This would be anchored by a Language Resource Type vocabulary, enlarged from Linguistic Data Type to encompass the full range of resources held by language archives [20]. In this way, we hope to shift from an idiosyncratic community-specific infrastructure to a mainstream infrastructure that interoperates with the global Web of Data. At the same time, we would hope to influence mainstream cataloging practices to embrace the Language Resource Type vocabulary, along with ISO 639-3 for greater precision in language identification, so that their catalog records would conform to the application profile for language resources.

## Acknowledgment

We acknowledge the financial support of the US National Science Foundation during the early days of OLAC. We are grateful to Mark Liberman and Christopher Cieri at the Linguistic Data Consortium, and to Martha Brogan and Lauris Olson at the University of Pennsylvania Library for their support in hosting OLAC infrastructure. We are indebted to Helen Dry and Anthony Aristar of the Linguist List, to participants of several OLAC and EMELD workshops, and to OLAC's advisory board and council, for many discussions about the design and operation of OLAC. Special thanks to Haejoong Lee for maintaining OLAC software over many years.

## References

[1] Helen Aristar Dry and Michael Appleby. 2003. OLAC Linguistic Subject Vocabulary. http://www.language-archives.org/REC/field.html.

[2] Helen Aristar Dry and Heidi Johnson. 2002. OLAC Linguistic Data Type Vocabulary. http://www.language-archives.org/REC/type.html.

[3] Steven Bird and Gary Simons. 2000. A Survey of the State of the Art in Digital Language Documentation and Description. http://www.languagearchives.org/docs/survey.html.

[4] Steven Bird and Gary Simons. 2000. White Paper on Establishing an Infrastructure for Open Language Archiving. http://www.languagearchives.org/docs/white-paper.html.

[5] Steven Bird and Gary Simons. 2003. Extending Dublin Core metadata to Support the description and discovery of language resources. Computers and the Humanities 37 (2003), 375–388. http://arxiv.org/abs/cs.CL/0308022.

[6] Steven Bird and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. Language 79 (2003), 557–82.

[7] Franz Boas (Ed.). 1911. Handbook of American Indian languages. Smithsonian Institution Bureau of American Ethnology Bulletin, Vol. 40. Washington: Government Printing Office.

[8] Mary Burke and Oksana Zavalina. 2019. Exploration of information organization in language archives. Proceedings of the Association for Information Science and Technology 56 (2019), 364–367.

[9] Philipp Cimiano, Christian Chiarcos, John P McCrae, and Jorge Gracia. 2020. Discovery of language resources. In Linguistic Linked Data. Springer, 263–279.

[10] Karen Coyle and Tom Baker. 2009. Guidelines for Dublin Core application profiles. https://www.dublincore.org/specifications/dublin-core/profileguidelines/

[11] Diane Hillmann. 2008. Metadata quality: From evaluation to augmentation. Cataloging and Classification Quarterly 46 (2008), 65–80.

[12] Christopher Hirt, Gary Simons, and Joan Spanne. 2009. Building a MARC-to-OLAC crosswalk: repurposing library catalog data for the language resources community. In Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries. ACM, 393–394.

[13] Russell Hugo. 2015. Constructing online language learning content archives for under-resourced language communities. Technical Report. University of Washington.

[14] Renato Iannella and Rachel Heery. 1999. Dublin Core Metadata Initiative – Structure and Operation. http://dublincore.org/documents/dcmi-structure/

[15] Heidi Johnson. 2003. OLAC Role Vocabulary. http://www.language-archives.org/REC/role.html.

[16] Heidi Johnson and Helen Aristar Dry. 2002. OLAC Discourse Type Vocabulary. http://www.language-archives.org/REC/discourse.html.

[17] Carl Lagoze, Herbert Van de Sompel, Michael Nelson, and Simeon Warner. 2002. The Open Archives Initiative Protocol for Metadata Harvesting, Version 2.0. http://www.openarchives.org/OAI/openarchivesprotocol.html.

[18] Gary Simons. 2009. OLAC Metadata Quality. http://www.language-archives.org/NOTE/metrics.html .

[19] Gary Simons. 2014. The role of metadata in the infrastructure for archival interoperation. Language and Linguistics Compass 8 (2014), 486–494.

[20] Gary Simons. 2016. From Linguistic Data Type to Language Resource Type: Laying the groundwork for a metadata application profile.https:

//scholars.sil.org/sites/scholars/files/gary_f_simons/presentation/simons-language_resource_type_vocabulary.pdf.

[21] Gary Simons and Steven Bird. 2000. Requirements on the Infrastructure for Open Language Archiving. http://www.language-archives.org/docs/ requirements.html.

[22] Gary Simons and Steven Bird. 2000. The Seven Pillars of Open Language Archiving: A Vision Statement. http://www.language-archives.org/docs/ vision.html.

[23] Gary Simons and Steven Bird. 2001. OLAC Metadata. http://www.language-archives.org/OLAC/metadata.html

[24] Gary Simons and Steven Bird. 2001. OLAC Repositories. http://www.language-archives.org/OLAC/repositories.html

[25] Gary Simons and Steven Bird. 2002. OLAC Process. http://www.language-archives.org/OLAC/process.html

[26] Gary Simons and Steven Bird. 2003. Building an Open Language Archives Community on the OAI Foundation. Library Hi Tech 21 (2003), 210–218. http://www.arxiv.org/abs/cs.CL/0302021

[27] Gary Simons and Steven Bird. 2003. The Open Language Archives Community: An Infrastructure for Distributed Archiving of Language Resources. Literary and Linguistic Computing 18 (2003), 117–128

[28] Gary Simons and Steven Bird. 2008. OLAC Linguistic Data Type Vocabulary. http://www.language-archives.org/REC/type.html

[29] Gary Simons and Steven Bird. 2008. Toward a global infrastructure for the sustainability of language resources. In Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation. De La Salle University, Manila, Philippines, 87–100.

[30] Gary Simons and Steven Bird. 2020. Expressing language resource metadata as Linked Data: The case of the Open Language Archives Community. In Development of Linguistic Linked Open Data Resources for Collaborative Data-Intensive Research in the Language Sciences, Antonio Pareja-Lora, María Blume, Barbara C. Lust, and Christian Chiarcos (Eds.). MIT Press, 117–130....

[31] Gary Simons, Steven Bird, and Joan Spanne. 2008. Best Practice Recommendations for Language Resource Description. http://www.languagearchives.org/REC/bpr.html.

[32] Nick Thieberger. 2012. Using language documentation data in a broader context. In Potentials of Language Documentation: Methods, Analyses, and Utilization, Frank Seifart, Geoffrey Haig, Nikolaus Himmelmann, Dagmar Jung, Anna Margetts, and Paul Trilsbeek (Eds.). University of Hawai'i Press, 129–34.

[33] Nick Thieberger. 2017. What remains to be done: Exposing invisible collections in the other 7,000 languages and why it is a DH enterprise. Digital Scholarship in the Humanities 32 (2017), 423–434.

[34] Hans Uszkoreit, Brigitte Jörg, and Gregor Erbach. 2003. An Ontology-based Knowledge Portal for Language Technology. In Proceedings of ENABLER/ELSNET Workshop "International Roadmap for Language Resources". ELRA.