LANG ARC 2023

2nd International Workshop on Digital Language Archives

# PROCEEDINGS OF THE INTERNATIONAL WORKSHOP ON DIGITAL LANGUAGE ARCHIVES: LangArc-2023

UNT

acm Association for Computing Machinery

IEEE

# Proceedings of the International Workshop on Digital Language Archives:

# LangArc-2023

Virtual Format
June 30 – July 1, 2023

Workshop Co-Chairs:
Oksana L. Zavalina, Shobhana L. Chelliah

Proceedings Chair:
Oksana L. Zavalina

University of North Texas
Denton, Texas

LANG ARC 2023

2nd International Workshop on Digital Language Archives

Digital Object Identifier (DOI): 10.12794/langarc2114294

# Welcome from the Workshop Co-Chairs

It is our pleasure to share with you the Proceedings of the 2nd International Workshop on Digital Language Archives (LangArc-2023)!  The proceedings include 10 peer-reviewed accepted submissions from Asia, Australia, Europe, and North America. The workshop, held as a virtual event on June 30, 2023, US Central time (June 30-July 1, 2023, Coordinated Universal Time UTC), is part of the ACM/IEEE Joint Conference on Digital Libraries 2023 https://2023.jcdl.org/ .

This interactive virtual workshop seeks to address a growing need. It explores a broad scope of issues related to digital language archives -- digital libraries that preserve and provide online access to language data. The objective of this workshop is to bring together researchers, practitioners, educators, and students from around the world who are currently working or are interested in working in different areas related to collecting, archiving, curating, organizing, and providing access to born-digital or digitized language data, and evaluation of digital language archives. The workshop will help foster collaborations among information professionals; library and information science, linguistics, data science, computer science, and humanities researchers; educators; representatives of language communities (including indigenous communities, refugees, speakers of under-resourced languages); and other interested audiences. The event is the second one in the series of regular workshops focused on the digital language archives. The 1st International Workshop on Digital Language Archives was held online on September 30, 2021, as part of the ACM/IEEE Joint Conference on Digital Libraries 2021.

We hope you find these proceedings interesting and useful and will consider attending or actively participating by authoring submissions for the upcoming meetings of the International Workshop on Digital Language Archives.

*Dr. Oksana L. Zavalina, Professor at the Department*
*of Information Science at the University of North Texas*
*Oksana.Zavalina@unt.edu*

*Dr. Shobhana L. Chelliah, Professor at the Department*
*of Linguistics at the Indiana University Bloomington*
*schellia@iu.edu*

# Table of Contents

# Bharatavani Project - Reviving Linguistic Diversity and Cultural Heritage in India: A Case Study

Narayan Choudhary, LR Premkumar, Chandan Singh, Shubhanan Mondal, Shivangi Priya, Beluru Sudarshan, P. Perumal Samy, Shailendra Mohan
Central Institute of Indian Languages Mysuru
Karnataka, India
n.choudhary@gov.in, lrprem90@gmail.com, chandansingh.ciil@gmail.com, subhanan.ciil@gmail.com, priyashivangi6@gmail.com, beluru.sudarshana@gmail.com, director-ciil@gov.in

## ABSTRACT

The Bharatavani project, launched in 2016 initiated by the Government of India, addresses the crucial need to preserve and promote indigenous languages and cultures. The paper presents an overview of the project, which focuses on recording socio-cultural and linguistic information about 121 Indian languages and making it accessible to a broader audience. The project leverages technological advancements to document significantly smaller and lesser-known languages and mother tongues in India, to raise awareness and maintain and promote the country's rich linguistic diversity. The Bharatavani project aims to bridge the digital divide and ensure equal access to knowledge and information by emphasising the importance of incorporating these languages into the digital sphere. Through the creation of e-content, the project offers multimedia resources, including text, audio, video, and images, through the online portal www.bharatavani.in and the Bharatavani Android App. This research highlights the significance of content generation, software development, and web portal creation for selected languages in the first phase, with subsequent plans for translation, online teaching-learning, and language teacher training in the second phase. By embracing the potential of technology, the Bharatavani project aspires to create a Knowledge Society in the digital era, enabling individuals across India to explore and celebrate their linguistic heritage.

## CCS CONCEPTS

• Applied computing → Computers in other domains → Digital libraries and archives • General and reference → Document types → Surveys and overviews • Social and professional topics → Computing / technology policy → Government technology policy • Human-centered computing → Accessibility→ Accessibility technologies.

## KEYWORDS

Digital Archives, Bharatavani, Ministry of Education, E-content, Indigenous Languages

**ACM Reference format:**

Narayan Choudhary, LR Premkumar, Chandan Singh, Shubhanan Mondal, Shivangi Priya, Beluru Sudarshan, P. Perumal Samy, and Shailendra Mohan. 2023. Bharatavani project - Reviving linguistic diversity and cultural heritage in India: A Case Study. In Proceedings of the 2nd International Workshop on Digital Language Archives (LangArc-2023), ACM/IEEE Joint Conference on Digital Libraries. USA, 3 pages. https://doi.org/10.12794/langarc2114300

## 1 INTRODUCTION

As a nation, India is known for its linguistic diversity and pluralism. Languages of India primarily belong to five linguistic families: Indo-European, Dravidian, Austro-Asiatic, Tibeto-Burman, and Samito-Hamitic [1]. According to the Census of 2011, there are 121 languages and 270 mother tongues with speakers' strength of 10 000 and above at the national level [3]. Twenty-two of these languages are included in the VIII Schedule of the Constitution of India, and there are 100 non-scheduled languages spoken by more than 10,000 speakers each. Moreover, there are several languages/mother tongues that are spoken by fewer than 10,000 persons each. The scheduled and non-scheduled languages and some other languages/mother tongues have writing systems.

In contrast, hundreds of languages/mother tongues remain oral. However, in the face of globalisation and the dominance of a few major regional languages, many regional and indigenous languages have been marginalised, risking the cultural heritage they embody. In response to this challenge, the Government of India launched the Bharatavani Project to preserve, promote, and propagate linguistic diversity by leveraging digital technologies. This research article provides an overview of the Bharatavani Project, its objectives, implementation strategies, and its impact on language documentation, promotion, and revitalisation. Additionally, it explores the project's significance in the broader context of linguistic heritage preservation and the challenges that need to be addressed for its sustained success.

Electronic content came rather late into Indian languages. In order to preserve knowledge about indigenous languages and cultures, it is crucial to record socio-cultural and linguistic information about these languages and make it accessible to a wider audience. It is essential to utilise technological advancements for documenting vernaculars, mainly smaller and lesser-known languages/mother tongues in India. This project can effectively raise awareness about these languages/mother tongues, thus contributing to maintaining India's rich linguistic diversity.

In India, the exclusion of mother tongues from formal education is closely tied to the perception of inferiority and reduced vitality [2] attributed to minor, minority, and tribal languages compared to dominant majority languages like English. Furthermore, the preference for English-medium education has marginalised other major regional and constitutional languages, significantly weakening them across all aspects of Indian society. Globalising information through the internet primarily occurs in English and a few scheduled languages, leaving hundreds of languages/mother tongues absent in the digital content realm. Moreover, content development in these languages/mother tongues, including their cultural components (textual, auditory, or visual), progresses slowly. As a result, a significant portion of the population needs help to utilise the information available in cyberspace to enhance their knowledge and foster social and economic growth. The need for technology to incorporate these languages/mother tongues into digital networks or provide translation services further compounds the issue.

To overcome this gap, the Ministry of Education, Government of India, launched the Bharatavani Project to address these challenges and safeguard India's linguistic heritage. The Bharatavani Project is implemented by the Central Institute of Indian Languages, Mysore, in the form of a web portal and mobile application where registered users can access books dealing with encyclopaedia, language learning materials, dictionaries & glossaries, textbooks, grammar and more in 121 Indian languages for free. Till March 2021, Bharatavani had accomplished the task of hosting more than 5500 resources in 92 Indian languages. Besides books, Bharatavani hosts multimedia content describing various literary and cultural aspects of Indic linguistic communities. Already the biggest single knowledge portal in the world, Bharatavani mobile application is now operational with 200+ digital dictionaries in multiple languages and subject combinations, the first of its kind in the world. Bharatavani has become the world's largest single-point hub of important indigenous content. The objective is to make indigenous knowledge resources in their respective languages available through a robust digital platform.

One of the project's key objectives is to develop e-content in different languages/mother tongues and showcase India's linguistic diversity in cyberspace. It is an integral part of the broader mission of creating a Knowledge Society in the Digital India. The project ran in two phases. In the first year, 18 Scheduled languages (Assamese, Bengali, Bodo, Dogri, Gujarati, Kannada, Kashmiri, Konkani, Malayalam, Manipuri, Maithili, Marathi, Nepali, Oriya, Punjabi, Santali, Tamil, Telugu) and 32 non-scheduled languages were be covered. In the second year, 68 non-scheduled (excluding Arabic, Afghani, English and Persian) Indian languages/mother tongues were covered.

The project revolves around an online portal called www.bharatavani.in and a corresponding Bharatavani Android App. These platforms deliver knowledge in multimedia formats, including text, audio, video, and images, for all languages in India. The project's initial focus is on content aggregation and developing a corner on the web portal for selected languages. The project aims to encompass translation, online teaching-learning, and online language teacher training in the second phase. Overall, the Bharatavani Project aims to leverage digital platforms to facilitate access to knowledge and information in and about all languages in India. By doing so, it seeks to embrace the country's linguistic diversity and contribute to advancing a Knowledge Society in the digital age.

## 2 OBJECTIVES OF THE BHARATAVANI PROJECT

The primary objectives of the Bharatavani Project include:

1. Creating a digital repository of linguistic resources: The project focuses on collecting, digitising, and archiving language-related materials, such as dictionaries, grammar, texts, and audio-visual content, to build a comprehensive repository.
2. Development of linguistic tools and technologies: The project emphasises the development of language learning tools, machine translation systems, text-to-speech synthesisers, and other language technologies to facilitate linguistic research and communication.
3. Dissemination of linguistic resources: Bharatavani aims to make linguistic resources accessible to a wide range of users through online platforms, mobile applications, and offline modes to promote language learning and research.

## 3 IMPLEMENTATION STRATEGIES

The Bharatavani Project employs a multi-faceted approach to achieve its objectives:

1. Language documentation and digitisation: Linguistic scholars and experts collaborate to document and digitise linguistic resources in various Indian languages, ensuring their preservation and accessibility.
2. Technology development: The project invests in research and development of language technologies to facilitate language learning, content generation, and translation services.
3. Language promotion and awareness: Bharatavani conducts workshops, seminars, and awareness campaigns to promote linguistic diversity and foster pride in regional languages among communities.

## 4 IMPACT AND ACHIEVEMENTS

Since its inception, the Bharatavani Project has made significant strides in language preservation and revitalisation. Some of the notable impacts and achievements include.

1. Preservation of endangered languages: By digitising and archiving endangered languages, the project has helped prevent the loss of linguistic knowledge and cultural heritage associated with these languages.
2. Language learning and dissemination: The project has provided digital platforms and mobile applications that enable users to learn Indian languages, facilitating cross-cultural understanding and communication.
3. Academic and research support: The availability of linguistic resources and technologies has enhanced linguistic research, enabling scholars to delve into the grammatical structures, dialects, and socio-cultural aspects of different languages.

## 4 CONCLUSIONS

Despite its successes, the Bharatavani Project faces several challenges that need to be addressed for its continued effectiveness. Indian language scripts are complex and challenging to connect with digitally generating software programmes. In contrast to English, where consonants and vowels have different representations in word creation, the alphabet, which consists of both, is written as a single unit for Indian languages. Indian scripts, sometimes known as abugida scripts due to this distinguishing trait, are challenging to create and incorporate into programmes. Also, making the Indic scripts compatible across devices is a major task for developing language technologies-based applications. In addition, for error-free performances, the unique glyphs used by Indian languages that cause issues with various devices must be fixed.

The Bharatavani Project is a significant initiative to protect India's linguistic diversity and preserve the cultural heritage embedded within these languages. The project has made commendable progress in language preservation, dissemination, and research by leveraging digital technologies and fostering community involvement. However, to ensure its sustained success, continued support, collaboration, and innovative strategies are necessary to overcome the challenges and propel the project towards a more inclusive and linguistically vibrant future for India.

## REFERENCES

[1] Abbi, Anvita (2018). A sixth language family of India: Great Andamanese, its historical status and salient present-day features. In The Dynamics of Language: Plenary and focus lectures from the 20th International Congress of Linguists (p. 134). Juta and Company (Pty) Ltd

[2] Giles, H. (1977). Towards a theory of language in ethnic group relations. In H. Giles (Eds.). Language, ethnicity and intergroup relations, (pp.307-348). London: Academic Press

[3] Office of the Registrar General of India. 2018. The Census of India 2011): Paper 1 of 2018- Language. Office of the Registrar General of India, New Delhi.

[4] Perumal P. Samy, Narayan Choudhary, L.R. Premkumar, Chandan Singh, Subhanan Mandal. 2021. Indian Languages in Digital Space: Sharing Knowledge through Web-portal and Mobile App. Central Institute of Indian languages, Mysore.

# Making Photographs in Language Archives Maximally Useful: Metadata Guidelines for Community and Academic Depositors

Shobhana L. Chelliah
Department of Linguistics
Indiana University Bloomington
United States of America
schellia@iu.edu

## ABSTRACT

Collections in language archives typically include photographs. The purpose of these photographs is to supplement linguistic information about materials, places, and people related to cultural activities that are being forgotten. Instruction on metadata creation for these photograph deposits must take into consideration the variety of depositors to and users of language archives. In addition to the use of existing controlled vocabularies, classification lists, or thesauri in metadata creation, we observe in metadata for photographs the need for open-ended descriptions of personal experience related to the objects, places, and things photographed.

## CCS CONCEPTS

• Applied computing → Computers in other domains → Digital libraries and archives • Applied computing → Document management and text processing → Document management → Document metadata • Information systems → Users and interactive retrieval; Multilingual and cross-lingual retrieval

## KEYWORDS

Language archives, photographs, metadata guidelines, language revitalization, community documentation, lexicography, dictionary

**ACM Reference format:**

Shobhana L. Chelliah. 2023. Making photographs in language archives maximally useful: metadata guidelines for community and academic depositors. In Proceedings of the 2nd International Workshop on Digital Language Archives (LangArc-2023), ACM/IEEE Joint Conference on Digital Libraries. USA, 3 pages. https://doi.org/10.12794/langarc2114301

## 1 RESEARCH PROBLEM

Collections in language archives typically include photographs. The purpose of these photographs is to supplement linguistic information about materials, places, and people related to cultural activities that are being forgotten. See for example *The Language Archive* (https://archive.mpi.nl/tla/islandora/object/tla%3A1839_00_0000_0000_0018_206F_2). The photograph is a key part of recovery of cultural information - a picture can supplement in invaluable ways words used to describe an object, person, or place or evoke memory of that object and its use, or a person and the relationship of that person to place or event.

The protocols for the digital preservation of photographs and the standard metadata needed for cataloging a photograph are well known (see for example, https://archivingforthefuture.teachable.com/ and https://vrc.uchicago.edu/guide-cataloging-your-images). An additional question posed by endangered language archival collections is the need for depositors where languages are in a state of attrition. Here, we ask two questions related to depositors and users. For depositors we ask, given that language documentation is undertaken by different stakeholders and researchers,

- What help can be provided in the selection and metadata collection process for photographs included in collections?
- What should depositors be aware of regarding the cataloging, cross-referencing and description of these items?
- How should they create metadata for these items? The traditional practices represented in the photographs may be linked by keywords, but speakers may not know those keywords because they have lost or are losing terms related to traditional practices. In this case, what additional information or guide is necessary to support access to culture and language documentation as provided in the photographs?

Using six anonymized depositor profiles, we first describe the differing goals and needs of depositors. We also consider how these needs and goals impact the kinds and number of photographs included in collections along with the metadata provided and how this metadata may provide access to community users of the archive. Previous guidelines on photograph metadata are not necessarily modeled for nonacademic community depositors and users. Take, for example, this explanation of descriptive metadata for relating files from *Archiving for the Future*: "…if a set of digital photographs of woven designs in fabric are meant to accompany a PDF document that describes weaving techniques, this information should be included in the descriptive metadata about these objects. Note that such relationships between files are also relevant structural metadata." [1] This statement is accurate and clear but with some reframing could be useful for a wider population.

## 2 BACKGROUND

Language documentation is a sub-discipline of linguistics based on awareness that all languages are valuable in the investigation of the extent and limits of cultural and linguistic diversity [2]. Linguists and communities speaking languages that have yet to be fully described face the danger of irrevocable loss of these sources of information as intergenerational transmission of language and culture is no longer guaranteed due to the pressures of world languages and additional factors such as displacement [3,4]. Documentary linguistics has placed a great effort on the creation of comprehensive digital, long-lasting records of at-risk or endangered language through language archiving. Language archives are the result of this effort.

The linguistic and cultural information in language collections may be used by those who robustly know and use the language, those who are heritage speakers with only a few words or phrases in their repertory, or someone with linguistic fluency between these situations ([5] and [6] on levels of proficiency). Depositors may have an academic connection to the language and be experts in the interpretation and use of the materials deposited (e.g., a verb paradigm, a traditional narrative with morphological glossing) or be non-academics involved in community or individual revitalization efforts [7]. It is against this backdrop that we look at metadata and guidance for depositors and users of photographs in language archives.

## 3 DEPOSITOR AND USER PROFILES

As part of our mission at the Computational Resource for South Asian Languages archive (CoRSAL: https://corsal.unt.edu/), we provide for would-be depositors the workshops and one-on-one sessions on metadata creation. Based on these, I list six scenarios of photograph deposits and challenges to metadata creation.

### 3.1 Non-community academic depositors

Our non-community academic depositors often include photographs with particular texts, such as traditional narratives, to provide additional tools for interpreting and contextualizing those texts. For example, the collection may include a traditional narrative about a particular type of building, so the depositor may include pictures of the building, its rooms, and materials used to construct the room. In this case, the depositor must be aware that in most digital repositories each picture will be entered as an independent item in a digital collection, and that the metadata would be the only way for a user to access these items together. Thus, the depositor must link the text in all its forms (audio, video, and transcriptions) with all the photographs.

Another type of photograph included by non-community academics are pictures of events such as story-telling festivals and literacy workshops. Some reasons it makes sense for a resource of an endangered language to include pictures of such events are to document who in the community is involved in language work, which elders' speech is represented, and which varieties (such as which village) are represented. The metadata for such items would be useful if they included relevance to the revitalization process and links with relevant audio and video (such as narratives provided by elders in the pictures).

A third category of photograph we have seen in the deposits of academic depositors is project staff at conferences or on field trips. Here, we provide our depositors with the following rule of thumb: if when creating metadata, the photograph does not link to any culturally or linguistically relevant place, person, or thing, consider excluding the photograph from the archival collection and including it on the project website.

### 3.2 Community depositors

Community members may contribute photographs to cultural heritage sites such as the Boro and Dimasa Heritage Digital Archive (https://bododimasaarchive.org/digital-heritage). Photographs are a very common upload to such sites but as we learned when listening to depositors at a workshop for this platform, metadata creation can be difficult. On the one hand, it seemed to the depositor unnecessary to include information that was well known to the community and on the other it was almost overwhelming as there was so much to be said about each item. In most cases, depositors wrote a few words and left it at that. Here, some training is needed for community members to examine why they included a photograph (what is the larger cultural significance, how is the photograph relevant to an event being documented, or how does the photograph have personal significance), what elements would need to be understood by future generations (including key ethnographic descriptors as described in [8]), and their own experience with or position vis-à-vis the item, place, event, or person. Again, the depositor needs to be aware that the photograph will be an independent item and in order to link with related items, the depositor must keep track of keywords and descriptions to repeat these in items that need to be linked together. One method of metadata creation we are now investigating is for the depositor to turn on a recorder, say the file name of the photograph, and record answers to a list of questions on the who, what, which, when, where, and how about the photograph along with room for why- that is, why has the depositor chosen to include this photograph in their collection. We feel that this low-stress method will free the community depositor to be more informative. A challenge here will be to acquire a contact language that is familiar to both the depositor and the curator.

Several CoRSAL depositors are expert photographers and videographers. We refer, for example, to the Daniel Tholung deposit of photographs in the Lamkang Language Resource digital collection. This deposit includes many pictures of dancers in traditional attire. The metadata directs viewers to specific aspects of the items pictured and provides names of items, ways they are created, who wears them and when. Items and metadata such as these can be used as a guide for community depositors about what to aim for in terms of completeness and relatedness of items.

## 3.3 Lexicographer depositors

Another category of depositors to CoRSAL are dictionary creators who include photographs illustrating words in a dictionary and deposit both the lexical database and the photographs for archiving. The result for a digital collection is potentially hundreds of individual photographs all related to a single dictionary project. The metadata creator needs to be aware that some users will browse the collection without knowledge of the overarching lexicography project, so the photographs must be relevant as a standalone piece. For this reason, the description of the item could include the full dictionary entry including pronunciation, translation and description and refer to related sound files where available.

# 4 CONCLUSION

Photographs can be an important tool in language and culture revitalization efforts and therefore are included in language archives. Depositors need guidelines on how to select and describe their photographs. Users need to know which photographs are best viewed together through relevant and transparent keywords. The onus of knowing keywords beforehand and details of traditional practices to support searching, cannot be placed on the user since, in the revitalization context, exact practices may be only weakly known. Rather, metadata must be overly specific and maximally linked to provide access. The main metadata takeaways are that:

1. Those depositing photographs used to illustrate dictionaries should fully fill out the metadata fields with the information in the dictionary entry so that the photograph tells a story as an independent item. In addition to this, the photograph should be linked back to the dictionary as a whole.
2. Those depositing pictures of events should note in the description the significance of the event for the goal of the collection as a whole. Each picture should tell a different part of the story so that there are no duplicates. Photographs not relevant to the goal of the collection should be omitted.
3. Photographs are a useful mechanism for enriching verbal descriptions of cultural practices. Depositors can be guided on how to express cultural details including the personal significance of the photographed items, persons, and places. Training is needed in the use of keywords to link related photographs.
4. Sample community deposits can be a great way to demonstrate how community depositors can create useful metadata for photographs.

# REFERENCES

[1] Kung, S. S., Sullivant, R., Pojman, E., & Niwagaba, A. (2020). Archiving for the future: Simple steps for archiving language documentation collections [OER]. https://archivingforthefuture.teachable.com/. CC BY-SA 4.0 license.

[2] Krauss, M. (1992). The World's Languages in Crisis. Language, 68 (1), 4-10.

[3] Bradley, D., & Bradley, M. (2018). Language Endangerment. Cambridge:

[4] Chelliah, S.L. (2021). Why Language Documentation Matters. Springer Briefs in Linguistics. Dordrecht: Springer Academic Press

[5] Chelliah, S.L, & de Reuse. W.J. (2011). Handbook of Descriptive Linguistic Fieldwork. Dordrecht: Springer Academic Press.

[6] Crowley, T. (2007). Field Linguistics. A Beginner's Guide. Oxford: Oxford University Press.

[7] Burke, M., Zavalina O.L., Chelliah, S.L., & Phillips, M.E. (2022). User needs in language archives: Findings from interviews with language archive managers, depositors, and end-users. Language Documentation and Conservation 16, 1-24.

[8] Franchetto, B. (2006). Ethnography in language documentation. In J. Gippert, N. Himmelmann, & U. Mosel (Eds.), Essentials of Language Documentation (pp. 183-212). De Gruyter Mouton. https://doi.org/10.1515/9783110197730.183 .

# Exploration of Metadata Practices in Digital Collections of Archives with Arabian Language Materials

Saleh Aljalahmah
Basic Education College
The Public Authority for Applied Education and Training
Kuwait
sh.aljalahmah@paaet.edu.kw

Oksana L. Zavalina
Department of Information Science
University of North Texas
United States of America
Oksana.Zavalina@unt.edu

## ABSTRACT

A high proportion of materials held by archives in Arabian Gulf and included in digital collections are oral histories, manuscripts, and other language content. As metadata is important for resource discovery, this study aimed to develop understanding of the current state of metadata practices in digital collections of archival institutions in the Arabian Gulf region. It also explored perspectives (including attitudes and possible barriers) for development of large-scale regional portals that would facilitate discovery of Arab digital archives (including language collections) by aggregating metadata. This research project used semi-structured interviews of the managers of 4 out of 5 digital language archives in Kuwait. Results provide insights into perspectives of metadata interoperability among archives and suggest the need for metadata training, and documenting metadata creation guidelines. Findings contribute to evaluating the feasibility of and planning for future functional regional aggregations of cultural heritage digital collections.

## CCS CONCEPTS

• Information systems → Database administration • Applied computing → Computers in other domains → Digital libraries and archives • Applied computing → Document management and text processing → Document management → Document metadata • General and reference → Evaluation

## KEYWORDS

Arabic language archives, information organization, information access, metadata interoperability

**ACM Reference format:**

Saleh Aljalahmah and Oksana L. Zavalina. 2023. Exploration of Metadata Practices in Digital Collections of Archives with Arabian Language Materials. In Proceedings of the 2nd International Workshop on Digital Language Archives (LangArc-2023), ACM/IEEE Joint Conference on Digital Libraries. USA, 4 pages. https://doi.org/10.12794/langarc2114295

## 1 INTRODUCTION AND LITERATURE REVIEW

Archives are cultural heritage institutions whose main function is to provide access to information. This access is enabled through information organization, which includes development and application of data content standards that guide metadata creation, for example, Describing Archives: A Content Standard (DACS). In information science and practice, the term metadata refers to bibliographic records that represent materials held by cultural heritage institutions (archives, libraries, and museums). Metadata is considered according to the type, function, domain, etc. (e.g., [4],[17]). Metadata must provide easy access and retrieval for the users, as well as support for the work tasks of collection managers [18]. Frameworks for evaluating metadata quality formulated metadata quality criteria (e.g., [7]).

Metadata schemes (e.g., Dublin Core, Encoded Archival Description, Machine Readable Cataloging: MARC, Open Language Archives Community Metadata: OLAC) include metadata element sets accompanied by metadata creation guidelines. Organization of information also entails development and application of the controlled vocabularies for names of persons and institutions (e.g., Union List of Artist Names, Thesaurus of Geographic Names), subject, genre, language, and other terms (e.g., Library of Congress Subject Headings, OLAC Discourse Type Vocabulary, Glottolog), and classification systems (e.g., Dewey Decimal Classification: DDC, etc.). In addition to these international standards for information organization, institutions often develop their own local metadata schemes, controlled vocabularies, and guidelines, or create adaptations of existing standards for their digital collections to better meet their target audience's needs.

To facilitate access and improve user experience, metadata records that represent materials held by archives, libraries, and museums are brought together in aggregations that serve as centralized points of access. Well-known examples of such aggregations include multinational (Europeana), and regional (e.g., Digital Library of the Caribbean). For a portal like that to function properly and support resource discovery, metadata aggregated in a portal needs to be interoperable. Metadata interoperability can be defined as "the compatibility of two or more systems such that they can exchange information and data and can use the exchanged information and data without any special manipulation" [8]. Quality of metadata has been evaluated and discussed in relation to its interoperability when brought together into aggregations of digital content (e.g., [20]).

Metadata interoperability is commonly achieved through development of mappings between different metadata schemes, and aggregation-wide metadata guidelines (e.g., metadata application profile) that metadata harvested into the aggregation needs to conform to (e.g., [9], [14]). The first step in this process is the background exploration of metadata practices and standards used by institutions that will likely participate in the portal as contributors. The first centralized portal – the Digital Library of the Middle East that aims to provide access to digital content from the Arabian Gulf archives, libraries, and museums – was launched in 2021.

Cultural heritage institutions in the Middle East have a long history of using international information organization standards. For example, Egyptian Organization for Standardization and Quality Control helped translate international information organization standards since 1957; Jordan Library and Information Association modified Dewey Decimal Classification system to better meet the needs of regional users since 1970s [10]. Arabian Gulf countries embrace the opportunity to share their knowledge with other countries through digital collections. The digitization movement in the region was pioneered by Qatar that launched its digital library in 2012 (Qatar National Library, no date). UNESCO special envoy for basic and higher education believes that digitization efforts could help the world better understand Arab culture [5].

Arabic language, along with e-government and information retrieval is among the most common research topics of the articles by authors from the Arabian Gulf published in top journals [19]. At the same time, there is shortage of investigations into the information organization practices in Arabian Gulf counties, especially in archives. So far, only one published paper and one poster abstract focused on the creation or adoption of metadata, metadata quality assurance, metadata interoperability in Arabian Gulf institutions, including one Kuwaiti archive that provides access to Arab language and culture data [2, 3]. No research has yet been published on aggregation of metadata in the portals that provide central point of access to collections of cultural heritage institutions, including archives. Our exploratory study examined the status of the organization of knowledge in archives in an Arabian Gulf country that can be characterized as digital language archives as they:

- have been developed largely by collecting oral histories after the Gulf war of 1990 when most Kuwaiti archival collections perished [1]
- provide access to digitized and/or born-digital Arab language materials.

## 2 METHOD

The research questions addressed by this study were:

- What are the techniques and approaches used in archives' information organization: metadata schema, data content standards, controlled vocabularies, content management tools, search options, metadata harvesting, etc.?
- What are the archive metadata managers' perceived readiness for and barrier to aggregating metadata in regional portals that would facilitate discovery of Arab digital archives (including language collections)?

This study focused on one Arabian Gulf country: Kuwait. Archives were selected based on criteria which included location of headquarters in Kuwait, and availability of one or more digital collections managed by the archive. At the time of data collection, only 5 archives in Kuwait met these criteria. Potential respondents were selected from the lists of employees available on their institutions' websites. Interviewing those employees of Kuwaiti archives who make decisions about information organization in digital collections allowed to identify similarities and differences, as well as opportunities and challenges for providing access to digital collections via large-scale centralized portals. The interview recruitment email was sent in two languages (English and Arabic) and invited to respond in the language of respondent's choice. The response rate constituted 80%, with representatives of 4 archives participating in the study.

We used email interviews as they allow participants to find time in their schedule to provide more thoughtful, reflective responses [13]. Previous studies on metadata-related topics relied on email interviews and found this approach effective [15]. The semi-structured interview questions were sent to the participants in both English and Arabic and participants had the freedom to choose the language of their answers.

## 3 FINDINGS AND DISCUSSION

The participants' views regarding large scale portals that would aggregate metadata records representing items in the Arabian Gulf countries' digital collections were very positive overall. All 4 participants believed this to be a necessary development. Respondents also raised some concerns, specifically emphasizing the associated costs. Despite Arabian Gulf countries' generally strong economies, governments are not necessarily ready to invest in such projects due to various reasons, including the post-pandemic shift of focus to other areas, and traditional lack of government support to archives. Participants also pointed out the lack of the workforce with the expertise and preparation necessary to design and implement projects like this in the Arabian Gulf countries, as there are no established large-scale aggregations in the region. Participants suggested that as a possible solution, which would however increase project costs, international experts with experience implementing and maintaining such portals could be hired. Relying on professionals from other countries is already an established practice.

The study revealed a variety of training levels among those responsible for metadata in digital collections. One respondent reported having a graduate degree in Library and Information Science, another in Library and Information Technology, and one more in Computer Science. One participant had an International Baccalaureate 2-year degree in unrelated field although took some archiving coursework In addition to formal education, one archive’s metadata manager reported that they were trained by more experienced colleagues/professionals, another one stated that they received personal on-the-job training, one participant was entirely self-trained in the archive metadata management tasks of their job, and the 4th participant reported a combination of self-training and attending workshops. Respondents had a wide range of years of experience in the field: one worked in archives since as early as 1989 and the most recently hired of our study participants started in 2008.

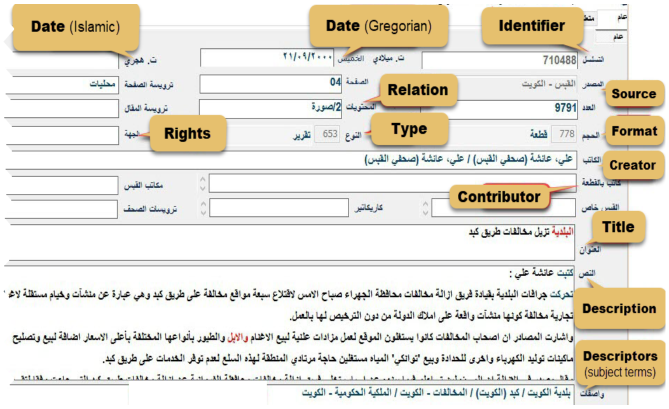
In the digital collections of two archives, searching is available in both Arabic and English languages. Two other archives only provide Arabic-language search capability. In 3 digital archives that participated in the study users can print and save/download digital items, as well as send them over email. Sharing on social media is also available in 2 archives. One archive implemented only the search function and no other navigation/interaction functions.

No single digital content management system was found to be used by more than one archive. Respondents were found to rely on VIRTUA, SQL, and Symphony. One archive developed its own in-house digital content management system. None of the participants reported that their metadata records are exposed for harvesting using Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

Three respondents reported using MARC 21 as the metadata scheme for their digital collections. One archive developed a local metadata scheme based on Dublin Core as shown in Figure 1 below, with 13 descriptive metadata elements based on 12 Dublin Core elements. None of the participants used the archive-specific metadata scheme Encoded Archival Description (EAD).

Two Kuwaiti digital language archives used standard Dewey Decimal Classification (DDC), and one developed the DDC-inspired local “Islamic Dewey” classification system. One archive relied on the outdated version 20 of the DDC released in 1989, while 3 newer versions have appeared since then (the latest in 2012) due to the limited budget not allowing for upgrade. One respondent’s’ answer to the question about classification scheme(s) used was invalid as they referred to the Anglo-American Cataloging Rules which is not a classification system.

Two study participants reported using subject headings lists developed by and for Arabian Gulf countries: El-Khazindar list of Arabic Subject Headings (Khazindar, 1983) or The Major List of Arab Subject Headings (https://www.amazon.com/-القائمة-الكبرى-لرؤوس-الموضوعات-العربية-ebook/dp/B07FXFCNNN/). Maknaz Expanded Thesaurus (http://en.maknaz.org/) developed in the Arabian Gulf and available based on subscription is used in one archive as a controlled vocabulary for genre terms and names, in addition to subject terms. Participants did not mention any other controlled vocabularies used and relied on free-text keywords for metadata fields other that those representing aboutness, genre, and creators or contributors.



عام
متعلق
عام
التسلسل 710488
ت. ميلادي الخميس ٢١/٠٩/٢٠٠٠
ت. هجري
المصدر القبس - الكويت
الصفحة 04
ترويسة الصفحة محليات
العدد 9791
المحتويات 2/صورة
ترويسة المقال
الحجم 778 قطعة
النوع 653 تقرير
الجهة
الكاتب علي، عائشة (صحفي القبس) / علي، عائشة (صحفي القبس)
كاتب بالقطعة
مكاتب القبس
القبس خاص
كاريكاتير
ترويسات الصحف
العنوان البلدية تزيل مخالفات طريق كبد
النص كتبت عائشة علي :
تحركت جرافات البلدية بقيادة فريق ازالة مخالفات محافظة الجهراء صباح الامس لاقتلاع سبعة مواقع مخالفة على طريق كبد وهي عبارة عن منشآت وخيام مستقلة لاع
تجارية مخالفة كونها منشآت واقعة على املاك الدولة من دون الترخيص لها بالعمل.
واشارت المصادر ان اصحاب المخالفات كانوا يستغلون الموقع لعمل مزادات علنية لبيع الاغنام والابل والطيور بأنواعها المختلفة بأعلى الاسعار اضافة لبيع وتصليح
ماكينات توليد الكهرباء واخرى للحدادة وبيع 'توانكي' المياه مستغلين حاجة مرتادي المنطقة لهذه السلع لعدم توفر الخدمات على طريق كبد.
واصفات بلدية الكويت / كبد (الكويت) / المخالفات - الكويت / الملكية الحكومية - الكويت

**Figure 1: Metadata record example**

To the question about availability of local metadata creation guidelines that are used to document and guide the metadata practices at their institutions, two interview participants responded negatively (with one of them commenting that they follow the guidelines of an existing standard: MARC 21). One interviewee reported having metadata creation guidelines restricted for internal use only on the site of the archive. None of the participants mentioned following the metadata guidelines found in the international standards for archival description: ISAD[G] and ISAAR[CPF].

## 4 CONCLUSION

This exploratory study’s findings will be useful as a benchmark for future research. Results show that some international standards of information organization have been adopted or adapted for the regional needs by those archives in Kuwait that include digital collections of Arabic-language materials and therefore fit the definition of a digital language archive. Our study found lack of metadata creation guidelines documenting local practices in application of the standards or locally developed information organization tools, and lack of participation in metadata harvesting with OAI-PMH or equivalents, which indicate potential challenges for interoperability of metadata. Other potential challenges include limited technical skill sets and financial resources available to support aggregations of archival digital collections.

Future research is needed to investigate metadata quality in digital archive collections with Arabian language content. Future studies will also examine the status of the organization of knowledge in other digital language archives in Kuwait, as more such archives are developed, and in other Arabian Gulf countries: Bahrain, Oman, Qatar, Saudi Arabia, and United Emirates. The observed metadata quality and metadata-related practices in these countries might affect the establishment of large-scale portals that include metadata records from entire region.

To ensure functionality of aggregations the demand for which is growing, Arabian digital language archives will need to take several steps towards ensuring metadata interoperability. This includes adopting and applying unified metadata creation guidelines. Generating crosswalks that show equivalences between metadata fields used by different archives in the region, as well as mapping between different controlled vocabularies used, is another important step worth considering.

# REFERENCES

[1] Ahmed, S. (2018). “Seeking information from the lips of people”: oral history in the archives of Qatar and the Gulf region. Archival Science, 18(3), 225-240.

[2] Aljalahmah, S., & Zavalina, O. L. (2021). Information representation and knowledge organization in cultural heritage organizations in Arabian Gulf: A comparative case study. Journal of Information and Knowledge Management, 20(2), 1-20. https://doi.org/10.1142/S0219649221500507

[3] Aljalahmah, S., & Zavalina, O.L. (2021). A case study of information representation in a Kuwaiti archive. In iConference 2021 Proceedings. http://hdl.handle.net/2142/109683

[4] Baca, M. (2016). Introduction to metadata. 3rd edition. Getty Publications. Retrieved from http://www.getty.edu/publications/intrometadata/

[5] Bade, R. (2010, May 27). Digital library project aims at the Middle East. Roll Call. Retrieved from https://www.rollcall.com/2010/05/26/digital-library-project-aims-at-the-middle-east/

[6] العربية - المجلد الأول [The Major List of Arab Subject Headings (2018): Arabic Edition]. Retrieved from https://www.amazon.com/-القائمة-الكبرى-لرؤوس-الموضوعات-العربية-ebook/dp/B07FXFCNNN/

[7] Bruce, T. R., & Hillmann, D. I. (2004). The continuum of metadata quality: defining, expressing, exploiting. ALA editions.

[8] CC:DA (ALCTS/CCS/Committee on Cataloging: Description and Access). (2000). Task Force on Metadata: Final report, June 16, 2000. Retrieved from http://downloads.alcts.ala.org/ccda/tf-meta6.html .

[9] Chan, L. M., & Zeng, M. L. (2006). Metadata interoperability and standardization–a study of methodology part I. D-Lib magazine, 12(6), 1082-9873.

[10] Eid, S. (2019). Library metadata standards and Linked Data Services: An introduction to Arab and international organizations. Journal of Library Metadata, 19(3-4), 163-185.

[11] Holloway, M. F. (1959). Patterns of library service in the Middle East. Library Trends, 8 (October), 192-208.

[12] Khazindar, I. A. (1983). Qāʾimat ruʾūs al-mawḍūʻāt al-ʻArabīyah: [Alkhazindar List of Arabic Subject Headings]. al-Kuwayt: Yuṭlabu min Dār al-Baḥūth al-ʻIlmīyah.

[13] Meho, L. I. (2006). E-mail interviewing in qualitative research: A methodological discussion. Journal of the American Society for Information Science and Technology, 57(10), 1284-1295.

[14] Park, J., & Childress, E. (2009). Dublin Core metadata semantics: An analysis of the perspectives of information professionals. Journal of Information Science, 35(6), 727-739.

[15] Park, J. R., & Tosaka, Y. (2015). RDA implementation and training issues across United States academic libraries: An in-depth e-mail interview study. Journal of Education for Library and Information Science, 56(3), 252-266.

[16] Qatar National Library. (No date). Qatar Digital Library. Retrieved from https://www.qdl.qa/en/about

[17] Riley, J. (2010). Seeing Standards: A Visualization of the Metadata Universe. Retrieved from http://jennriley.com/metadatamap/seeingstandards.pdf

[18] Riley, J. (2017). Understanding metadata. Washington DC, United States: National Information Standards Organization (http://www.niso.org/publications/press/UnderstandingMetadata.pdf), 23.

[19] Robinson, B. (2016). Addressing bias in the cataloging and classification of Arabic and Islamic materials: Approaches from domain analysis. In A.B. Click et al. (eds.), Library and Information Science in the Middle East and North Africa (pp. 255-269), Munich: De Gruyter Saur.

[20] Shreeves, S. L., Knutson, E. M., Stvilia, B., Palmer, C. L., Twidale, M. B., & Cole, T. W. (2005). Is quality metadata shareable metadata? The implications of local metadata practices for federated collections. In Hugh A. Thompson (ed.), Proceedings of the Twelfth National Conference of the Association of College and Research Libraries (pp. 223-237), Chicago, IL: Association of College and Research Libraries. Retrieved from http://hdl.handle.net/2142/145.

# Why it Can be Difficult to Make Historic Language Recordings Accessible: A View from a Corpus of Historic Dialect Recordings

Christian Huber
Phonogrammarchiv
Austrian Academy of Sciences
Vienna, Austria
christian.huber@oeaw.ac.at

## ABSTRACT

There is a growing demand to make historic linguistic field recordings accessible not only to the scientific community but also to the language communities as well as the interested public. However, when dealing with a corpus of historic language recordings, a number of challenges must be faced before dissemination issues can even be addressed. The present paper reports the experiences made in preparing a corpus of historic Austrian dialect recordings from the Phonogrammarchiv's holdings and the real-life issues encountered in the process and discusses what needs to be done with such a corpus before something can be done with that corpus.

## CCS CONCEPTS

• Applied computing → Computers in other domains → Digital libraries and archives • Applied computing → Document management and text processing → Document management → Document metadata • Security and privacy → Human and societal aspects of security and privacy → Privacy protections • Applied computing → Law, social and behavioral sciences → Anthropology → Ethnography

## KEYWORDS

Historic dialect recordings, historic language corpora, metadata structuring, granularisation, geodata, archiving, digitisation, ethical issues, legal issues

**ACM Reference format:**

Christian Huber. 2023. Why it can be difficult to make historic language recordings accessible: A view from a corpus of historic dialect recordings. In Proceedings of the 2nd International Workshop on Digital Language Archives (LangArc-2023), ACM/IEEE Joint Conference on Digital Libraries. USA, 4 pages. https://doi.org/10.12794/langarc2114302

## 1 INTRODUCTION

The Phonogrammarchiv of the Austrian Academy of Sciences has been engaged in making linguistic recordings from its inception [1], its first recording of an Austrian dialect of German dating from 1901 [2]. Over the decades, a collection of several thousand recordings of German dialects of Austria and adjacent areas has been created [3][4]. However, historically grown collections of language recordings pose challenges that are rarely discussed, as they do not arise in modern corpora that are generated within a specific research context and infrastructure. In such collections, the recordings were made not only at different times, but also with different objectives, according to different methods, with different recording technologies, and using different documentation practices [5]. Therefore, before such corpora can be exploited in linguistic or other research, one must deal with questions of data organisation as well as the preservation of their sonic content.

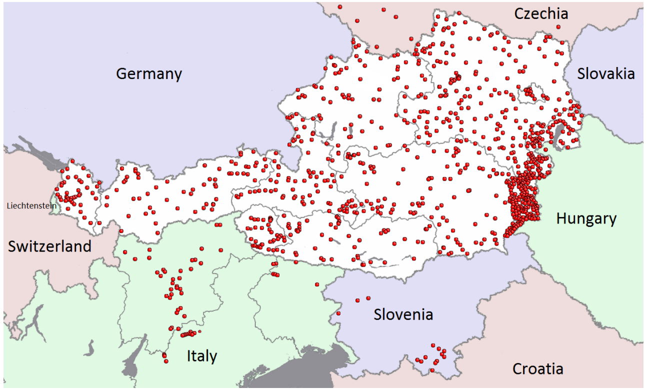


**Figure 1. Locations of documented dialect points (audio recordings) (© OpenStreetMap contributors)**

In accordance with the then-prospective budget, we selected approximately 2450 recordings of spontaneous language on magnetic tape (and some digital audio tape cassettes) from, roughly, 1000 places and 2500 speakers (see Figures 1 and 2), covering almost five decades (early 1950s to mid-1990s). In a cooperation of the Phonogrammarchiv with the Austrian Science Fund Special Research Programme F60 *German in Austria* and the *Austrian Centre for Digital Humanities* that started in 2016, we digitised these recordings and provided a structured and searchable description building on the Phonogrammarchiv's database and aim to annotate them utilising the corpus-linguistic structures developed in the *German in Austria* programme, and finally to present the results in a common platform.

## 2 DIGITISING THE TAPES

Traditional analogue sound carriers, e.g., wax cylinders or magnetic tape, are subject to natural decay. Once the carrier can no longer be played, the recordings on it are lost forever. Therefore, perishable sound documents must be digitised as long as the carriers can still be properly played in order to preserve the recorded contents in the long term. Digital audio data are no longer bound to an individual data carrier but can be losslessly copied as often as desired. In this way they can be electronically preserved for a virtually indefinite period of time.

At the start of the project, less than half of the recordings had already been digitised. The remainder was contained on around 400 tape reels that were digitised to 24bit/96kHz .wav files and subsequently segmented, so that each recording is now available as a separate file. We also discovered that among the previously digitised materials, a considerable number of digital copies of tapes had not been segmented, or only incompletely so, and other tapes had been digitised only partially. We therefore had to include the completion of these tasks in our workflow.

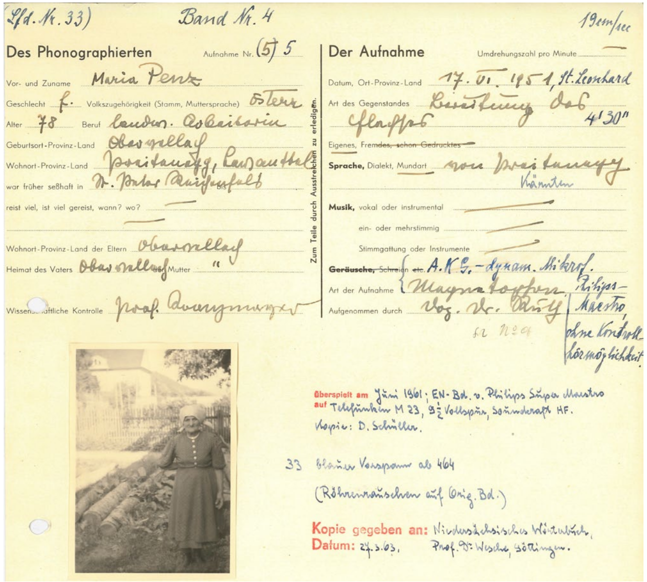
**Figure 2. Fieldwork in Carinthia (1951) (©Phonogrammarchiv)**

## 3 METADATA

The original historical archival documentation consists of data sheets on paper for each recording (for a long time handwritten, later typewritten) that were already available in a scanned format (.pdf files; for an example see Figure 3). Metadata include, e.g., the archive signature, the date and place of the recording, its duration, the recordees' names and social data, the involved fieldworker(s), recorded languages/varieties or musical forms, topics and other content-related indications, a time protocol detailing the contents of the recording, and technical metadata (e.g., technical equipment involved, track positions, tape speed).

### 3.1 Metadata enrichment

For handling the metadata, we utilised the pre-existing, very fine-grained, structures provided by the Phonogrammarchiv's relational database, and the metadata entries already available in it. However, these entries were often incomplete and in need of granularisation. When the Phonogrammarchiv introduced the electronic documentation of recordings in a database around 1990, there were already tens of thousands of recordings with archival documentation on paper. To save time and to have all recordings represented in the database quickly, most often only some basic metadata had been entered. An important task in the project was therefore to enrich the electronic metadata pertaining to our corpus based on the available analogue documentary materials (to be typed out or subjected to optical character recognition), and also to correct possible errors.



(Lfd. Nr. 33) Band Nr. 4 19cm/sec

Des Phonographierten Aufnahme Nr. (5) 5

Vor- und Zuname Maria Penz

Geschlecht f. Volkszugehörigkeit (Stamm, Muttersprache) österr.

Alter 78 Beruf landw. Arbeiterin

Geburtsort-Provinz-Land Obervellach

Wohnort-Provinz-Land Preitenegg, Lavanttal

war früher seßhaft in St. Peter Reichenfels bis

reist viel, ist viel gereist, wann? wo? —

—

Wohnort-Provinz-Land der Eltern Obervellach

Heimat des Vaters Obervellach der Mutter "

Wissenschaftliche Kontrolle Prof. Kranzmayer

Zum Teile durch Ausstreichen zu erledigen.

Der Aufnahme Umdrehungszahl pro Minute —

Datum, Ort-Provinz-Land 17. VI. 1951, St. Leonhard

Art des Gegenstandes Bearbeitung des Flachses 4'30"

Eigenes, Frem~~des, schon Gedrucktes~~

Sprache, Dialekt, Mundart von Preitenegg Kärnten

Musik, vokal oder instrumental —

ein- oder mehrstimmig —

Stimmgattung oder Instrumente —

Geräusche, ~~Schreien etc.~~ A.K.G.-dynam. Mikrof.

Art der Aufnahme Magnetophon Philips-Maestro, ohne Kontroll-hörmöglichkeit

Aufgenommen durch Doz. Dr. Graf

62 N° 9

Überspielt am Juni 1961; EN-Bd. v. Philips Super Maestro
auf Telefunken M 23, 9½ Vollspur, Sonderkraft HF.
Kopie: D. Schüller.

33 blauer Vorspann ab 464

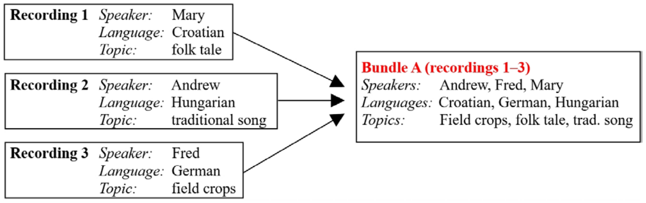
(Röhrenrauschen auf Orig. Bd.)

Kopie gegeben an: Niedersächsisches Wörterbuch,
Datum: 27.3.63. Prof. Dr. Wesche, Göttingen.

**Figure 3. Archive protocol of recording B 33 from 1951 (excerpt) (©Phonogrammarchiv)**

### 3.2 Granularisation

However, when switching to electronic documentation in the 1990s, it had also been decided to set up the database in such a way that it does not document individual recordings but only bundles of recordings: the metadata of the individual recordings made by a fieldworker on the same day were collapsed and lumped together into a single general bundled entry composed of the metadata of all recordings in the bundle, thereby dissociating the metadata from the actual recordings to which they pertain, as schematically shown in Figure 4. In such bundle entries the metadata are no longer associated with individual recordings but only with the bundle as such. Thus, from *Bundle A* in Figure 4 it can no longer be told whether *Mary*, or *folk song*, or *Croatian*, or any of the other entries, pertains to *recording 1*, *2*, or *3*.



**Recording 1** *Speaker:* Mary
*Language:* Croatian
*Topic:* folk tale

**Recording 2** *Speaker:* Andrew
*Language:* Hungarian
*Topic:* traditional song

**Recording 3** *Speaker:* Fred
*Language:* German
*Topic:* field crops

**Bundle A (recordings 1–3)**
*Speakers:* Andrew, Fred, Mary
*Languages:* Croatian, German, Hungarian
*Topics:* Field crops, folk tale, trad. song

**Figure 4. Lumping together metadata in a bundle entry (schematically)**

Therefore, search results can be severely contaminated, since a particular search criterion does not return individual recordings in the search results but only bundles of recordings that contain one or more recordings to which the search criterion applies. In addition, the search results cannot specify which recordings these are, and the search may also return a number of recordings to which the search criterion does not apply. Similarly, a combined search, e.g., a search involving two search criteria, may return bundles in the search results in which one or more recordings correspond to one of the search criteria at a time but with no recording to which both criteria apply. Since with bundles containing more than one recording, the search criterion may apply to minimally one and maximally all recordings in the bundle in the search results, the original protocols on paper must be consulted to determine the precise recording(s) to which the search criterion applies. Thus, a huge number of recordings cannot be unambiguously found by a search in the database, and the database often returns search results that do not conform to the search criteria.

In our corpus, roughly 50% of the recordings were included in such metadata bundles. Since sometimes up to 20 speakers (each representing the local variety of a different village) were recorded on a single day in the field, we were faced with a number of very complex bundles. To make the electronic documentation usable for any search-related purposes and corpus exploitation tasks, it was therefore necessary to granularise all metadata bundles and re-associate all pieces of metadata with those individual recordings to which they actually pertain. Since the problem is not restricted to our corpus but extends across the Phonogrammarchiv's database, we decided that the procedures to achieve this must be applicable to the database in general. For practical reasons we created an excerpt of the Phonogrammarchiv's database that contains only the data sets relevant to our corpus. Later, these data sets will be re-transferred and will replace the original entries.

In the next step we granularised all bundle entries composed of the metadata of several recordings into as many single-recording bundles as there were recordings in the bundle, together with extending the bundle signature by a delimiter followed by internal consecutive numbering (schematically shown in Figure 5).

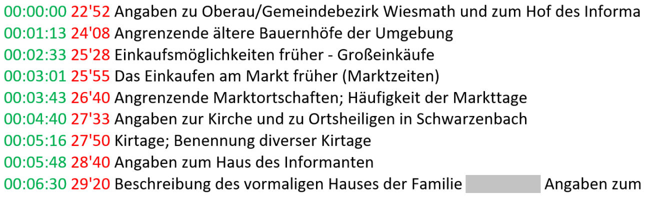
| *multi-recording bundle signature* | | *single-recording bundle signatures* | *original recordings (archive numbers)* |
|---|---|---|---|
| 19520919.N001<br>(= B 181–B 197) | => | 19520919.N001#001 | (= B 181) |
| | | 19520919.N001#002 | (= B 182) |
| | | 19520919.N001#003 | (= B 183) |
| | | *(etc.)* | |

**Figure 5: Granularisation of multi-recording bundles**

With the help of a matrix tool, each piece of metadata from the original bundle entry was then assigned to the single-recording bundle to which it pertains. Since in the original multi-recording entries all links between the metadata and the respective recordings were lost, this reassignment of metadata had to be done manually by falling back on the original hand- or typewritten documentation.

### 3.3 Timelines in protocols

Since the timelines in the original protocols of recordings (indicating what happens when in a recording) often do not start at the beginning of the respective recording but at the beginning of the tape reel containing it (which usually contains several other recordings), we had to correct the time markers in about 900 protocols and align them with the sound files (as, e.g., in Figure 6), later to be linked to the sound files in the database.



00:00:00 22'52 Angaben zu Oberau/Gemeindebezirk Wiesmath und zum Hof des Informa
00:01:13 24'08 Angrenzende ältere Bauernhöfe der Umgebung
00:02:33 25'28 Einkaufsmöglichkeiten früher - Großeinkäufe
00:03:01 25'55 Das Einkaufen am Markt früher (Marktzeiten)
00:03:43 26'40 Angrenzende Marktortschaften; Häufigkeit der Markttage
00:04:40 27'33 Angaben zur Kirche und zu Ortsheiligen in Schwarzenbach
00:05:16 27'50 Kirtage; Benennung diverser Kirtage
00:05:48 28'40 Angaben zum Haus des Informanten
00:06:30 29'20 Beschreibung des vormaligen Hauses der Familie [redacted] Angaben zum

**Figure 6: Adapted and original time markers in a protocol**

## 4 GEODATA

Due to the large number of villages and towns covered in the corpus it was necessary to implement a uniform and unambiguous representation of geographical information using a controlled list of places and converting mentions of toponyms (recording site, a speaker's place of birth or residence, etc.) into references to entries in the list of places. A local authority, *Statistik Austria*, provided us with an up-do-date and official dataset of all towns in Austria, including their official administrative names and geodata as well as the larger administrative units (municipalities, districts, provinces). With the help of this data set, it was possible to set up a representation of place names in such a way that they are not only identified by their official designations and reference numbers (beside geographical coordinates) but also are embedded in the hierarchy of the respective administrative units, where each level is embedded under the next higher level (i.e., PLACE < MUNICIPALITY < DISTRICT < PROVINCE < STATE), with the option of also adding alternative names of a toponym (e.g., potential historical names, or its name in other languages), or other information.

## 5 DISSEMINATION: LEGAL AND ETHICAL QUESTIONS

While it is a noble goal to make historic dialect recordings accessible to all interested parties (researchers, communities, or also the interested taxpayer who often financed the fieldwork and archiving), legal regulations have still to be obeyed, and ethical questions must be considered.

The recordings in the corpus were generally made under the stipulation that they would be used only for research purposes but would not be made publicly accessible. Thus, the recordings at times also feature sensitive or rather personal content (identified as such by the fieldworkers, the informants themselves, or also archivists), and great care must be taken when considering what should be made accessible to whom, even if several decades have passed since the recordings were made.

On the legal side, it must be kept in mind that the recordings were made at a time before it became common practice to record an agreement with the speaker as to how a recording could be used. A crucial question is whether what a speaker utters on a recording surpasses the threshold of originality and is protected by copyright law. In most cases this question cannot be decided outside a court of law, and permission to publish a recording had to be obtained from the speakers or their legal successors. However, in most cases, the personal data given in the protocols is not sufficient to track down speakers or their heirs (e.g., no date of birth is mentioned but only the year of birth, or the age at the time of the recording). If speakers or their heirs are not known or cannot be located, the respective recordings might be registered as orphan works.

In some types of research, e.g., sociolinguistics, certain personal data may be relevant. Since according to the General Data Protection Regulation the protection of an individual's personal data expires with the individual's passing (see, e.g., [8]), the personal data of speakers who are known, or can safely be assumed, to be deceased (e.g., if a speaker had reached a higher age than the oldest living individual in Austria, or on Earth), could in principle be shared. However, ethical considerations may come into play here as well. While it is standard practice to anonymise (or rather, pseudonymise [7]) personal data in written accounts, sound recordings pose the problem of the human voice. Whether or not a speaker's voice counts as personal data is still a matter of debate (see, e.g., [8] vs. [9]).

For such reasons, making the recordings openly accessible is not a trivial matter, and affordable solutions generally applicable not only to isolated recordings but to larger portions of the corpus, or to the entire corpus, are not yet in sight.

## 6 CONCLUSION

The preparation of a corpus of historic language recordings can be laden with more complications than first meets the eye. The dire funding situation in Austria for such projects often requires dividing the work between several cooperation partners contributing their respective expertise, and requires the partners' goodwill, and much in-kind work. A high degree of flexibility is asked for, since a change of priorities on the part of a cooperation partner (or even one's own department) may soon have the consequence that the project's objectives cannot be achieved according to the original planning, so that alternative ways must be found. Thus, we hope to be finally able to tackle the issue of merging the Phonogrammarchiv's metadata with *German in Austria*'s corpus-linguistic structure and to start annotating selected recordings, which has been delayed for several years. So far, about 70 transcripts in various formats (from the 1970s) are available. For increasing the number of transcripts, we have meanwhile decided to also include contributions from parties external to the cooperation who work on recordings from the corpus in other contexts. It is clear, however, that it will still take time until a substantial number of fully described and annotated recordings become available.

## ACKNOWLEDGMENTS

Thanks go to both cooperation partners, to Gernot Katzlberger of *Statistik Austria* for making available the geodata of nearly 17300 places in Austria, and to Regula Sutter for her help in selecting the appropriate CCS concepts.

## REFERENCES

[1] Leo Hajek. 1928. Das Phonogrammarchiv der Akademie der Wissenschaften in Wien von seiner Gründung bis zur Neueinrichtung im Jahre 1927 (58. Mitteilung der Phonogrammarchivs-Kommission). *Sitzungsberichte der Akademie der Wissenschaften in Wien, philosophisch-historische* Klasse 207(3), Hölder-Pichler-Tempsky, Vienna, 1–22. Online: https://www.oeaw.ac.at/fileadmin/Institute/PHA/PDF/Hajek_1928.PDF (accessed June 8, 2023)

[2] Ph 105: Attergauer Dialekt https://catalog.phonogrammarchiv.at/sessions.php?id_sessions=3794&action=view&sortieren=signatur&vonBis=0-9 (accessed May 14, 2023)

[3] Maria Hornung. 1961. Tonaufnahmen im Dienste der Mundartforschung. Zum 60jährigen Bestehen des Phonogrammarchivs der österreichischen Akademie der Wissenschaften in Wien. *Zeitschrift für Mundartforschung* 28(2), 183–191.

[4] Wilfried Schabus. 1999. Die Bestände des Phonogrammarchivs an Sprachaufnahmen. *Das audiovisuelle Archiv* 45, 23–32. Online: https://www.oeaw.ac.at/fileadmin/Institute/PHA/PDF/schabus_1999.pdf (accessed June 8, 2023)

[5] Christian Huber and Benjamin Fischer. 2021. Digitising a corpus of Austrian dialect recordings from the 20th century. In *Digital Lexis and Beyond*, ed. by Ch. Katsikadeli, M. Sellner & M. Gassner, Verlag der ÖAW, Vienna, 38–65. DOI: https://doi.org/10.1553/OE_Phonogrammarchiv

[6] Wirtschaftskammer Österreich. 2023. EU-Datenschutz-Grundverordnung (DSGVO): Wichtige Begriffsbestimmungen. https://www.wko.at/service/wirtschaftsrecht-gewerberecht/EU-Datenschutz-Grundverordnung:-Wichtige-Begriffsbestimmu.html (accessed May 14, 2023).

[7] Caroline Schwabe. 2021. Was sind pseudonymisierte Daten? Pseudonymisierte Daten nach DSGVO. https://www.robin-data.io/datenschutz-akademie/wiki/pseudonymisierte-daten/ (accessed May 14, 2023).

[8] David Vasella. 2021. VK Berlin: Identifikation über die Stimme nicht möglich; Begriff der "Verarbeitung". https://datenrecht.ch/vk-berlin-identifikation-ueber-die-stimme-nicht-moeglich-begriff-der-verarbeitung/ (accessed May 14, 2023).

[9] Datenschutz.org. 2023. Biometrische Daten: Besondere Schutzwürdigkeit bei sensibelsten Daten! https://www.datenschutz.org/biometrische-daten/ (accessed May 14, 2023).

# Multiperspectivity and Neutrality in Language Archives

Tobias Weber
Graduate School Language & Literature
Ludwig-Maximilians-Universität München
Munich, Germany
weber.tobias@campus.lmu.de

## ABSTRACT

This paper discusses linguistic data and their creation with a focus on the human actions and decisions that shape them. The human factor and positionalities are often obscured by current practices in data handling. Obscuring the human agency does not only reduce the transparency of the research but also disenfranchises the humans behind the data. The language archive takes a dual role as a host for discourse and an agent within it. It cannot claim impartiality but adds perspectives to research data and narratives.

## CCS CONCEPTS

• Applied computing → Computers in other domains → Digital libraries and archives • General and reference → General literature • Security and privacy → Digital rights management; Data anonymization and sanitization; Information accountability and usage control; Social aspects of security and privacy • Human-centered computing → Interaction design theory, concepts and paradigms; Collaborative and social computing theory, concepts and paradigms.

## KEYWORDS

digital language archives, positionality, academic discourse, language data

### ACM Reference format:

Tobias Weber. 2023. Multiperspectivity and neutrality in language archives. In Proceedings of the 2nd International Workshop on Digital Language Archives (LangArc-2023), ACM/IEEE Joint Conference on Digital Libraries. USA, 4 pages. https://doi.org/10.12794/langarc2114303

## 1 INTRODUCTION

Humans take a central role in shaping language documentation projects and their outputs. These outputs are subsequently archived or disseminated to stakeholders and interest groups, e.g., scientific publications, text collections, pedagogical materials. In the process, the artefactually contained interaction between researcher(s) and their consultant(s) is also preserved and offers insights to the documentation project [26]. Following Windfeld Lund, all documentation has a communicative intent [34]; documentation efforts are communicative events that convey the relevance of an observation or interaction through time [12]. In linguistic data, there are at least two layers of communication, one directed to the researcher and the immediate audience by the speaker, the other to the envisioned audience of the documentation outputs, e.g., elders telling stories for future generations, researchers eliciting a phenomenon of particular interest to the research community. This constitutes an instance of audience design [6] whereby different communicative goals are served by a single interaction. This may occasionally cause difficulties in handling data, for example in cases where information was shared that should be kept within a defined community (e.g., secret knowledge, personal narratives), or where communicative intent was biased, or the situation is not clear to outsiders. The latter are prominently reported for older data, such as legacy materials [4, 28, 29], but can affect new data alike. There are different positionalities and perspectives in each archival deposit, and interaction with this discourse is facilitated through the archive. Thus, the focus of this paper is to discuss the role of memory institutions in academic discourse, beyond providing data.

## 2 HUMANS IN LINGUISTIC DATA

A major factor contributing to the commodification of language data is the structuralist view of language and society. While a structural approach to language is not problematic as such, i.e., aiming to describe languages systematically and generalise language use by individuals into abstract descriptions, this research requires a fundamental transformation of the data: Human agents often need to be replaced by variables describing their sociodemographic features. This initial step in data handling allows for clustering and characterisation of language use by different parts of the speaking communities, e.g., by regional distribution, age, gender, language biography. While this does not remove the consultant from the data record, the combination of features becomes more prominent, as it allows for the combination and comparison with other data points elicited from other consultants. It is, thus, the basis for generalisation and accounts of any language X or language use in/by a given context or a particular demographic. The latter also includes data collected automatically through crawling the internet, social networks, or digital databases (as used for research in computational linguistics and the computational social sciences).

In these contexts, analyses and associated data sets often aim for reproducibility, a goal also formulated for linguistic data [7, 11, 24]. Yet, these goals and standards, e.g., for FAIR data [32], are concerned with the use and reuse by academics and not primarily with facilitating access and use by community members. They stand at the opposite end of the spectrum and will be, bar citizen science, more interested in the stories told by their relatives or neighbours, mediated by their personal experiences and interactions with the people [16]. The language documenters' conduct and personal interactions with the community are more relevant than the technical details of the language data – and these data sets, independent of their quality or relevance for linguistic enquiry, may be a symbol of historical injuries for the communities, as often encountered with older data sets [4, 11]. In this view, the human factor in language documentation is more than the roles fulfilled by consultants, on which the documentation projects still crucially depend.

Through their decisions in setting the research agenda, the researchers may inadvertently reproduce injustices, e.g., from a colonial past [18]. They bear the responsibility for conducting ethical research and maintaining positive community relationships [2]; linguistic data sets must be transparent on the provenance of the data and the methods used in generating them[1]

. This can be facilitated by comprehensive descriptions of the data sets, i.e., a meta-documentation [3]. Yet again, a focus on standardised descriptions for the descriptive metadata, e.g., ontologies for metadata, can obscure the interactions between researcher and consultant.

As a consequence, the human influence in the research process is less evident and secondary to the data itself. Considering ethical principles for language work and community relationships, privacy protection and personality rights are of central importance. Once these prerequisites for sharing and making data accessible are agreed, it is furthermore important to acknowledge each contribution in the creation, transcription, annotation, and analysis of the data, ideally naming all individuals involved in a task [1, 23]. Ontologies may offer standardised description of contributor roles, but they do not reflect the decisions made by each individual – a detailed documentation of workflows additionally increases transparency and allows all involved to receive merit for their contributions [27, 29].

## 3 HUMANISING DATA – HUMANISING RESEARCH

In adopting the stance that humans should be regarded as more than a set of characteristics or associated data, the human factor in the creation and dissemination of language data remains visible. The danger of generating highly structured but anonymous data lies in reducing human roles in the creation, annotation, and analysis of linguistic data. Returning to the issue of reproducibility, language data is unique to the particular contexts of their creation. With different consultants, different researchers, other research objectives, or altered spatiotemporal settings, we expect to observe changes in the language (use). While it is possible to get identically reproduced language data, this is confined to highly standardised genres, e.g., related to customs and religion. And even within these standardised or formulaic language forms, it is possible to find variance and innovations introduced by individual speakers. For example, folk tales and songs exhibit different wording that can be analysed in terms of its historical development and geographical spread [14].

Apart from oral tradition, philologists have a long-standing tradition of studying differences between copies of manuscripts and other artefacts bearing written language data (e.g., inscriptions), even for standardised and canonised texts such as religious scriptures. While differences might appear negligible on a global scale, the study of variance between versions constitutes an independent research objective. In such projects, the aim is to explain divergent readings through the origins of each version, where human involvement is a central factor for explaining differences. Consequently, information on human agents and their involvement in the creation and use of a textual artefact is crucial to obtain – projected to the case of linguistic data and artefacts of research, a similar need for accounts of human actions can be observed for linguistic legacy data [11, 26]. Researchers working with orphan data and legacy materials often need to reconstruct information, despite the availability of metadata. The need to humanise linguistic data and to keep the human influence transparent is, thus, a prerequisite for maintaining understandable data sets. The necessary amount of information is difficult to estimate without knowing the research questions, yet examples from conversation analysis (and other fields based on an ethnomethodological approach) show that contextual clues that would not be routinely recorded in the 'thick' metadata shape the language data at hand [9, 30]. There might be no upper boundary for relevant metadata.

The call for complete accounts of the contexts underlying the data sets and the research generated with them is not restricted to the Humanities. Data provenance is also topical in natural and social sciences, linking humans and their actions to alterations in the data [20, 23]. Anonymous data of unclear provenance is more difficult to understand and replicate – it is less likely that data fabrication [2] and other instances of scientific misconduct are detected if contextual information is incomplete. If individual researchers (and consultants) are associated with data sets and their actions attributable to them, they have to accept the responsibility for their contributions. At the same time, they can claim full merit for their work [1, 27]. Thus, unless there are concerns about privacy and the protection of vulnerable groups, it is most ethical to keep human influence in the data visible, thereby allowing for the reconstruction of research contexts, data provenance, and the narrative of a research project.

[1] This can be understood as fulfilling the three principles outlined by Labov [17] and Wolfram [35]: the principle of error correction, the principle of debt incurred, and the principle of linguistic gratuity.

[2] Although this is a serious infringement of academic integrity and unethical, it is possible to learn about academic practices from these 'breaching experiments' [21, 22]

## 4 THE ROLE OF THE ARCHIVE

As discussed, technical metadata and descriptors cannot fully cover the contexts in the required detail. This also holds true for mere accounts of researchers, consultants, and other roles in the creation of data. And while language data contains traces of human actions and decisions that link them to their authors, the issue lies in the structural nature of these metadata. The story of the data, the research project, or a consultant is not self-contained – data do not talk for themselves. As with other instances of meaning making, understanding requires knowledge about the processes at work. A focus on interaction and actions and the rationale behind them allows peers in the present and future to comprehend and retrace the research trajectory, to find the 'human in the loop' [8], and, ultimately, creates transparency about the data and associated research. This is where an archive offers insights to the various positions and perspectives within the data and all accompanying metadata. An evident account of human action is more relatable to the layperson, including past and future consultants, than complex metadata accounts.

We may ask whether the archive itself a 'forum' or 'arena' for this discourse is just [15, 28], where the focus lies on presenting data and metadata. Yet, in the decision of what to present and how to portrait or collate artefacts, the archive, and any curator or archivist, becomes a part in the discourse. This has been noted by Derrida [10, p. 12], who outlines '[a]n eco-nomic archive in this double sense: it keeps, it puts in reserve, it saves, but in an unnatural fashion, that is to say in making the law (nomos) or in making people respect the law'. By adopting standards and ontologies, or in rejecting data sets [5], the archive takes an active position in academic discourse. It is not just a 'locus' [28] but an agent with certain positionalities and active contributions to discourse; the archive is both a player and the stage.

Thus, the archive does not just offer a starting point in the investigation of discourse surrounding data or an artefact, as in Foucault's archaeology of knowledge [13], but must itself be investigated in terms of its own positions – multiperspectivity is facilitated by the archive but also involves the archive and its own perspective. While there are strands in archival science that aim to investigate tacit assumptions and professional self-conceptions of archivists [33], the goal of multiperspectivity and open discourse must go beyond tracking archivists' decisions. The view of the archive as a neutral venue must be reconsidered, as it hosts a multiplicity of positions through the various embedded discourses and communicative goals in the data and adds layers of information through its own processes. The use of standards or standardised procedures can never replace accounts of human involvement in data creation and dissemination. Knowledge and 'truth' are shaped from the discourse in and on the deposits – it is co-located in the archive but decentralised in its form. A user contributes their own ideas, (research) questions, and perspectives through accessing the archive and enters the multiple discourses within. In this work, they reconstruct meaning from different sources and perspectives, thereby humanising research and adding new perspectives to old ones. The language archive needs to account for this exchange through and with itself, yet without claiming a neutral or unbiased stance.

## 5 CONCLUSION

This paper has emphasised different dualities and multiplicities regarding memory institutions and their positionality. There are different axes and perspectives in conceptualising and describing a language archive that either affect the institution itself or the data within it. Both can be tied to multiple contexts and functions, as the purposes for storing and accessing knowledge in an archive are manifold. The key to investigating these dual and multiple perspectives lies in keeping human decisions and actions in the archived data as well as in interaction with the deposits visible and retraceable. Since artefactual histories and research narratives are shaped by these (inter-)actions, special attention must be given to situations involving historical data sets (e.g., legacy data) or minoritised language communities; data in these contexts merit special ethical considerations and control for communities over their data [19, 25, 31, 36]. Ultimately, the archive itself cannot claim a neutral stance in this respect, as it, either through its statutes as an organisation or through the actions of associated humans, influences data preservation, presentation, and reuse. Likewise, language data are not neutral, and need to be investigated from a variety of perspectives. The archive can support this endeavour by allowing multiperspectivity in the archived data and in the work with deposits, thereby inviting discourse on and through language data within its various contexts.

## REFERENCES

[1] Helene N. Andreassen, Andrea L. Berez-Kroeker, Lauren Collister, Philipp Conzett, Christopher Cox, Koenraad De Smedt, Bradley McDonnell, and the Research Data Alliance Linguistic Data Interest Group. 2019. Tromsø recommendations for citation of research data in linguistics. https://doi.org/10.15497/rda00040 Research Data Alliance

[2] Peter K. Austin. 2010. Communities, Ethics and Rights in Language Documentation. Language Documentation and Description 7 (2010), 34–54. https://doi.org/10.25894/ldd226

[3] Peter K. Austin. 2013. Language documentation and meta-documentation. In Keeping Languages Alive. Documentation, Pedagogy, and Revitalisation, Mari Jones and Sarah Ogilvie (Eds.). Cambridge University Press, Cambridge, 3–15. https: //doi.org/10.1017/CBO9781139245890.003.

[4] Peter K. Austin. 2017. Language Documentation and Legacy Text Materials. Asian and African Languages and Linguistics 11 (2017), 23–44. http://hdl.handle.net/10108/89205.

[5] Samuel J. Beer. 2021. Interdisciplinary aspirations and disciplinary archives: Losing and finding John M. Weatherby's Soo data. Language Documentation and Description 21 (2021), 101–139. https://doi.org/10.25894/ldd19.

[6] Allan Bell. 1984. Language style as audience design. Language in Society 13, 2 (1984), 145–204.

[7] Andrea L. Berez-Kroeker, Lauren Gawne, Susan Smythe Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David I. Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nick Thieberger, Keren Rice, and Anthony C. Woodbury. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. Linguistics 56, 1 (2018), 1–18. https://doi.org/10.1515/ling-2017-0032.

[8] Steven Bird. 2020. Decolonising Speech and Language Technology. In Proceedings of the 28th International Conference on Computational Linguistics, Donia Scott, Nuria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, Barcelona, 3504–3519. https://doi.org/10.18653/v1/2020.colingmain.313

[9] Mary Burke, Hannah Tarver, Mark Edward Phillips, and Oksana Zavalina. 2022. Using existing metadata standards and tools for a digital language archive: a balancing act. The Electronic Library 40, 5 (2022), 579–593. https://doi.org/10.1108/EL-02-2022-0028

[10] Jacques Derrida. 1995. A Freudian Impression. Diacritics 25, 2 (1995), 9–63

[11] Lise M. Dobrin and Saul Schwartz. 2021. The social lives of linguistic legacy materials. Language Documentation and Description 21 (2021), 1–36. https://doi.org/10.25894/ldd12

[12] Konrad Ehlich. 2007. Textualität und Schriftlichkeit. In Was ist ein Text? Alttestamentliche, ägyptologische und altorientalistische Perspektiven, Ludwig Morenz and Stefan Schorch (Eds.). De Gruyter, Berlin and Boston, 3–17. https://doi.org/10.1515/9783110924336.3

[13] Michel Foucault. 2010 [1969]. Archaeology Of Knowledge. Vintage Books, New York

[14] Frog. 2013. Revisiting the historical-geographic method(s). In Limited sources, boundless possibilities: Textual scholarship and the challenges of oral and written texts, Karina Lukin, Frog, and Sakari Katajamäki (Eds.). The Retrospective Methods Network Newsletter, Vol. 7. University of Helsinki, Helsinki, 18-34.

[15] Isto Huvila. 2008. Participatory archive: towards decentralised curation, radical user orientation, and broader contextualisation of records management. Archival Science 8 (2008), 15–36. https://doi.org/10.1007/s10502-008-9071-0

[16] Ilya Khait, Leonore Lukschy, and Mandana Seyfeddinipur. 2022. Linguistic archives and language communities questionnaire: establishing (re-)use criteria. The Electronic Library 40, 5 (2022), 539–551. https://doi.org/10.1108/EL-01-2022-0012

[17] William Labov. 1982. Objectivity and commitment in linguistic science: The case of the Black English trial in Ann Arbor. Language in Society 11, 2 (1982), 165–201. https://doi.org/10.1017/S0047404500009192

[18] Wesley Y. Leonard. 2017. Producing language reclamation by decolonising 'language'. In Language Documentation and Description, Wesley Y. Leonard and Haley De Korne (Eds.). Vol. 14. EL Publishing, London, 15–36. http://www.elpublishing.org/PID/150

[19] David Nathan. 2014. Access and accessibility at ELAR, an archive for endangered //doi.org/10.25894/ldd161 languages documentation. Language Documentation and Description 12 (2014), 187–208. https://doi.org/10.25894/ldd172

[20] Thomas Pasquier, Matthew K Lau, Ana Trisovic, Emery R Boose, Ben Couturier, Mercè Crosas, Aaron M Ellison, Valerie Gibson, Chris R Jones, and Margo Seltzer. 2017. If these data could talk. Scientific Data 4, 1 (Dec. 2017), 170114. https: //doi.org/10.1038/sdata.2017.114

[21] Helen Pluckrose, James Lindsay, and Peter Boghossian. 2021. Understanding the "Grievance Studies Affair" Papers and Why They Should Be Reinstated: A Response to Geoff Cole. Sociological Methods & Research 50, 4 (2021), 1916–1936. https://doi.org/10.1177/00491241211009946

[22] Ian Reilly. 2020. Public Deception as Ideological and Institutional Critique: On the Limits and Possibilities of Academic Hoaxing. Canadian Journal of Communication 45, 2 (2020), 265–285. https://doi.org/10.22230/cjc.2020v45n2a3667

[23] Anne E. Thessen, Matt Woodburn, Dimitrios Koureas, Deborah Paul, Michael Conlon, David P. Shorthouse, and Sarah Ramdeen. 2019. Proper Attribution for Curation and Maintenance of Research Collections: Metadata Recommendations of the RDA/TDWG Working Group. Data Science Journal 18 (2019), 54. https://doi.org/10.5334/dsj-2019-054

[24] Tobias Weber. 2019. Can Computational Meta-Documentary Linguistics Provide for Accountability and Offer an Alternative to "Reproducibility" in Linguistics?.In 2nd Conference on Language, Data and Knowledge (LDK 2019) (OpenAccess Series in Informatics (OASIcs), Vol. 70), Maria Eskevich, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek, and Milan Dojchinovski (Eds.). Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 26:1–26:8. https://doi.org/10.4230/OASIcs.LDK.2019.26

[25] Tobias Weber. 2021. Mind the gap: Language data, their producers, and the scientific process. In 3rd Conference on Language, Data and Knowledge (LDK 2021) (OpenAccess Series in Informatics (OASIcs), Vol. 93), Dagmar Gromann, Gilles Sérasset, Thierry Declerck, John P. McCrae, Jorge Gracia, Julia BosqueGil, Fernando Bobillo, and Barbara Heinisch (Eds.). Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 6:1–6:9. https://doi.org/10.4230/ OASIcs.LDK.2021.6

[26] Tobias Weber. 2021. Philology in the folklore archive: Interpreting past documentation of the Kraasna dialect of Estonian. Language Documentation and Description 21 (2021), 70–100. https://doi.org/10.25894/ldd18

[27] Tobias Weber. 2021. The Curation of Language Data as a Distinct Academic Activity: A Call to Action for Researchers, Educators, Funders, and Policymakers. Journal of Open Humanities Data 7, 28 (2021). http://doi.org/10.5334/johd.51.

[28] Tobias Weber. 2022. Conceptualising language archives through legacy materials. The Electronic Library 40, 5 (2022), 525–538. https://doi.org/10.1108/EL-02-20220029

[29] Tobias Weber. 2023. Internal and external social dimensions of linguistic legacy materials: the case of Kraasna South Estonian. Ludwig-Maximilians-Universität München, Munich. https://doi.org/10.5282/edoc.31860

[30] Tobias Weber and Mia Klee. 2020. Agency in scientific discourse. Bulletin of the Transilvania University of Braşov Series IV: Philology and Cultural Studies 13, 1 (2020), 71–86. https://doi.org/10.31926/but.pcs.2020.62.13.1.5

[31] Joshua Wilbur. 2014. Archiving for the community: Engaging local archives in language documentation projects. Language Documentation and Description 12 (2014), 85–102. https://doi.org/10.25894/ldd166

[32] Mark D. Wilkinson, Michel Dumontier, Ijsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 1 (Dec. 2016), 160018. https://doi.org/10.1038/sdata.2016.18

[33] Joseph Winberry and LaVerne Gray. 2022. From "Mesearch" to "Wesearch": The Role of Community in Developing Identity-Centric Research. In Proceedings of the Association for Library and Information Science Education. Annual Conference: ALISE 2022. Go Back and Get It: From One Narrative to Many. University of Illinois, Urbana-Champaign. https://doi.org/10.21900/j.alise.2022.1033

[34] Niels Windfeld Lund. 2010. Document, text and medium: concepts, theories and disciplines. Journal of Documentation 66, 5 (2010), 734–749. https://doi.org/10.1108/00220411011066817

[35] Walt Wolfram. 1993. Ethical considerations in language awareness programs. Issues in Applied Linguistics 4, 2 (1993), 225–257

[36] Anthony C. Woodbury. 2014. Archives and audiences: Toward making endangered language documentations people can read, use, understand, and admire. Language Documentation and Description 12 (2014), 19–36. https: //doi.org/10.25894/ldd161.

# A CARE- and FAIR-Ready Distributed Access Control System for Human-Created Data

Peter Sefton
Language Data Commons of Australia
The University of Queensland
Brisbane QLD Australia
p.sefton@uq.edu.au

Moises Sacal Bonequi
Language Data Commons of Australia
The University of Queensland
Brisbane QLD Australia
m.sacalbonequi@uq.edu.au

Simon Musgrave
Language Data Commons of Australia
The University of Queensland
Brisbane QLD Australia
s.musgrave@uq.edu.au

Jenny Fewster
Australian Research Data Commons
Monash University
Melbourne VIC Australia
jenny.fewster@ardc.edu.au

## ABSTRACT

The Language Data Commons of Australia (LDaCA) makes nationally significant language data available for academic and non-academic use, managing the data in a culturally, ethically, and legally appropriate manner guided by FAIR and CARE principles. Here, we describe the approach which we are taking to access control and a design for a distributed access control system which can look after the A-is-for-accessible in FAIR data while respecting the CARE principles. We also describe and demonstrate a pilot system based on that design, showing how data licenses that allow access by identified groups of people can be used by adding functionality, CILogon for non-institutional identification and REMS for managing access to resources, to the existing Australian Access Federation infrastructure.

## CCS CONCEPTS

• Applied computing → Computers in other domains → Digital libraries and archives • Applied computing → Arts and humanities • Security and privacy → Human and societal aspects of security and privacy

## KEYWORDS

Language archive, FAIR, CARE, access control

**ACM Reference format:**

Peter Sefton, Moises Sacal Bonequi, Simon Musgrave, and Jenny Fewster. 2023. A CARE- and FAIR-Ready Distributed Access Control System for Human-Created Data. In Proceedings of the 2nd International Workshop on Digital Language Archives (LangArc-2023), ACM/IEEE Joint Conference on Digital Libraries. USA, 5 pages. https://doi.org/10.12794/langarc2114304

## 1 INTRODUCTION

The Language Data Commons of Australia (LDaCA) focuses on preservation and discovery of distributed multi-modal language data collections under a variety of governance frameworks. This will include access control that reflects ethical constraints and intellectual property rights, including those of Aboriginal and Torres Strait Islander, migrant and Pacific communities. Regarding rights, our project is informed by the CARE principle (https://www.gida-global.org/care) for Indigenous data which also describe the level of respect which should be given to any data collected from individuals or communities.

Language archiving has received considerable attention in the last 20 years because of the importance of the practice in the documentary linguistics tradition originating with Himmelmann [4]. Discussions of access to language archives [1,3] concentrate on the need for access control, who should be involved in making decisions and how those decisions can be documented. Perhaps understandably, the details and implementation of processes at a technical level have received less attention. Two exceptions to this generalisation must be mentioned. Broeder et al. [2] present a technical architecture for access control in a federated repository system, and Nathan [6] discusses a system based on the roles which can be taken by those interacting with the archive, an approach which emphasises that technical solutions must be based on human behaviour. In this paper, we present a design for a distributed access control system which could look after the A-is-for-accessible in FAIR data while respecting the CARE principles; and describe and demonstrate a pilot system based on that design, showing how data licenses that allow access by identified groups of people can be used with an Australian Access Federation (AAF) pilot system (CILogon) to give the right people access to data resources. We suggest that our approach combines desirable features of the designs described by [2] and by [6].

## 2. BASIC PRINCIPLES

Our system must be able to implement data access policies with real-world complexity and one of our challenges has been developing a data access policy that works across a range of different collections of language data. Accessibility, the A of FAIR data [8], means that data is accessible to the right people and who is included in 'right people' varies from collection to collection and even within a single collection. Another challenge is to make sure that the information about access is sustainable; that is, the information is not locked in a specific software solution and can be easily reused when delivery systems change.

The key idea is to separate safe storage of data from its delivery. Each item in a repository is stored with licensing information in natural language and the repository defers access decisions to an Authorization system, where data custodians can design whatever process they like for granting license access.

## 3. LICENSES

A license in this context is *a natural language document* in which a copyright holder sets out the terms and conditions of use for data. Licenses *may* have metadata that describes them, e.g., a property to say that this is an open license and such metadata *about* a license can be used to automate decision making. If it is labelled as being an open license, then a repository can serve data and include that data, if it is labelled as "closed" or more aptly, "authorization-required" then repository software can perform an authorization step, which we cover in detail later.

In the world of research data generated by or about human participants, licenses can't always allow unauthenticated access and data redistribution, and they may permit distribution only to certain people, or classes of person. So, a license is a document that expresses conditions such as "Data can be used by other researchers", but unfortunately we don't have systems in the research-data ecosystem which can automatically identify a user as "a researcher" (see also [2]).

The access control system we have been prototyping is based on licenses. For any data object, we store a license with it, and we give the license an ID which is a URL we can use to identify it uniquely (see Figure 1). Figure 2 shows how a license is explicitly linked to the data using a metadata description standard known as "Research Object Crate" (RO-Crate) [7]. Each object in the repository is a crate, with a metadata file that describes the object and (optionally) its component files, including the data license. Every item in a repository has a license, which may be an open one like CC Share Alike or a custom license derived from the ethics and participant agreements for a study in the context of local laws and institutional policy.

Using this license, distributed access portals in our architecture can check against an authorization system for each request for data. The portals may host data with the same licensing but do not need to maintain access control lists.

## 4. AUTHENTICATION

When we first developed access controls for LDaCA in 2021 it was a requirement that data licensing and access control decisions be decoupled from each other, and from particular repository software. We could not find an available open-source system for managing license-based access to data, so our starting approach used groups as a proxy for granting licenses on the basis that all common user-directory services such as LDAP include the concept of user groups.

A proof-of-concept Github based system demonstrated that authorization can be delegated from a data repository service to an external service. For each of the licenses there was a Github group (organization). The data_repository, when requested to serve data would get the user to login using the Github Authentication services (no Github repositories were used), then check if the user was in the correct license group. Although this worked, there were no workflow options and it supported only a single logon service, which is not widely used in academia or by community groups.

The AAF were already working with other research groups on a service called CILogon (https://www.cilogon.org/). Like Github, this service has groups but also allows users to log in with a variety of Authentication providers, including research institutions, via the AAF as well as social logins such as Google and Microsoft (and our old friend Github).

Again, this worked, but the current version of CILogon does not have particularly easy-to-use ways for a license-holder to create groups. The AAF team made us aware of the Resource Entitlement Management System, (REM: https://github.com/CSCfi/REMS), which is an open source application out of Finland which has been used previously in at least one language data repository [5]. This software is the missing link for LDaCA in that it allows a data custodian to grant licenses to users. And it works with CILogon as an Authentication layer so we can let users log in using a variety of services.

At the core of REMS is a set of licenses which can be associated with Resources - in our design this is (almost always) a one-to-one correspondence, for example we would have a license "Sydney Speaks Data Researcher Access license" corresponding to a resource that represents ALL data with that license. These Resources can then be made available through a catalogue, and workflows can be set up for pre-authorization processes ranging from single-click authorizations where a user just accepts a license and a bot approves it, to complex forms where users upload credentials, and one or more data custodians approve their request and grant them the license (see Figure 3). Once a user has been granted a license then a repository can authorize access to a resource by checking with REMS to see if a given user holds the license. Users do not have to find REMS on their own - they will be directed to it from data and computing services when they need to apply for authorization. Figure 4 shows the interactions involved in accessing data once a user has been granted the license in REMS via a data portal which gives access to data in a repository or archive.

## 5. DISCUSSION

Access Control Lists (ACLs) are a popular approach to the problem we are addressing but we suggest that the more modular approach which we advocate has several advantages over ACLs. Firstly, ACLs need maintenance over time - people's identities change, they retire and die, so storing a list of identifiers such as email addresses alongside content is not a viable long-term preservation strategy. Rather, we will encourage data custodians to describe in words what are permitted uses for the data, and by whom, in a license, then allow whoever is the current data custodian to manage that access in a separate administrative system.

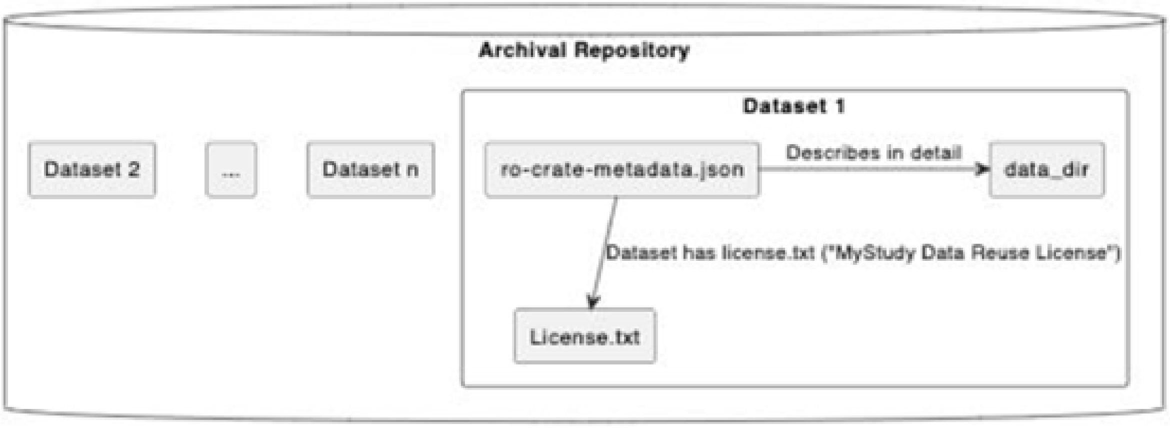


**Figure 1: Data packaging architecture**

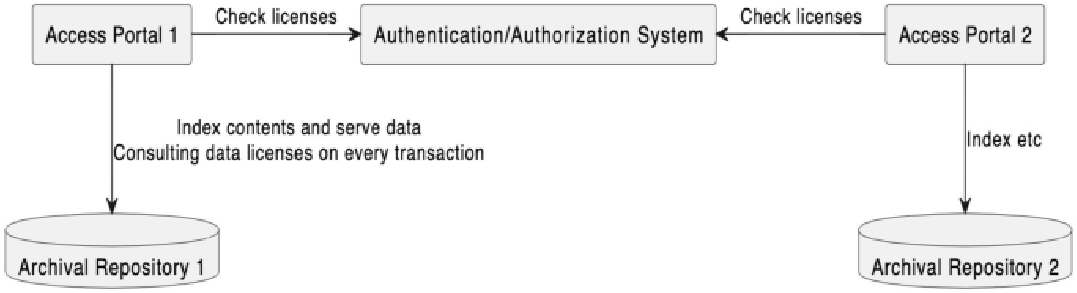

Access Portal 1
Check licenses
Authentication/Authorization System
Check licenses
Access Portal 2
Index contents and serve data
Consulting data licenses on every transaction
Index etc
Archival Repository 1
Archival Repository 2

**Figure 2: Relationships between repositories, portals, and the Authentication System**

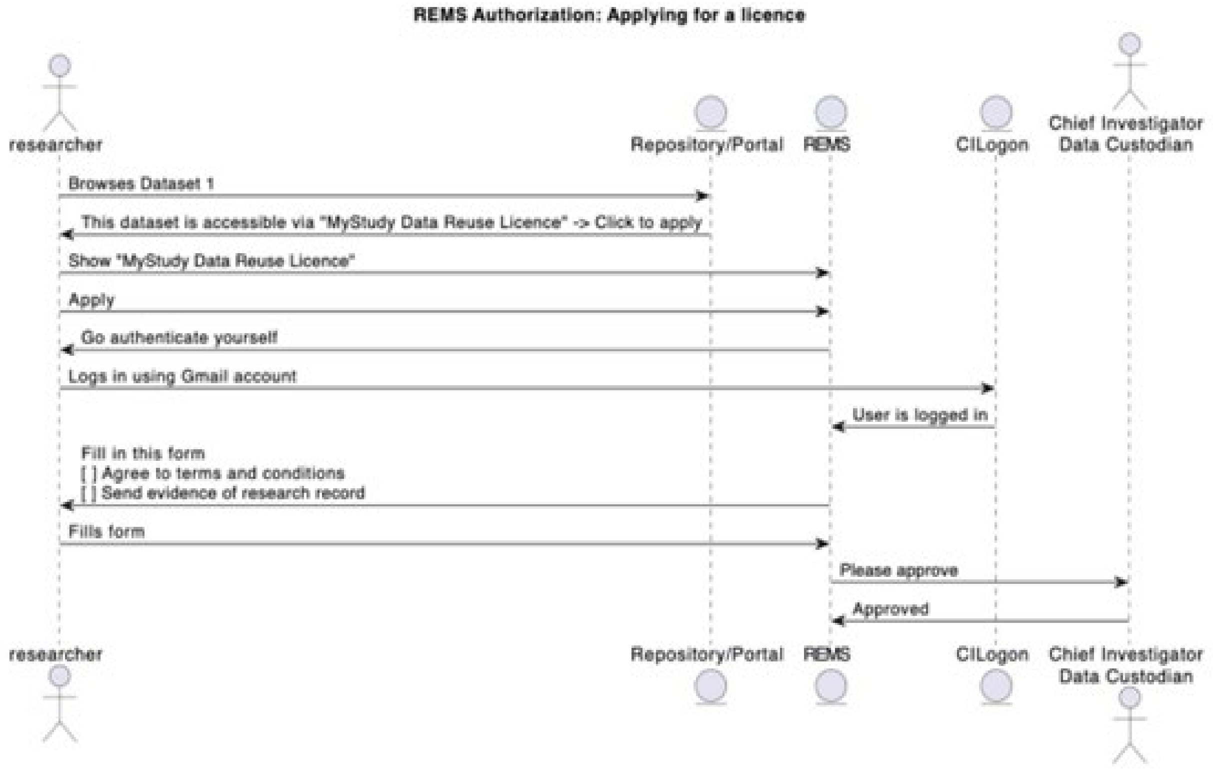


**Figure 3: Interaction diagram showing the flow involved in a user applying for a data license via REMS.**

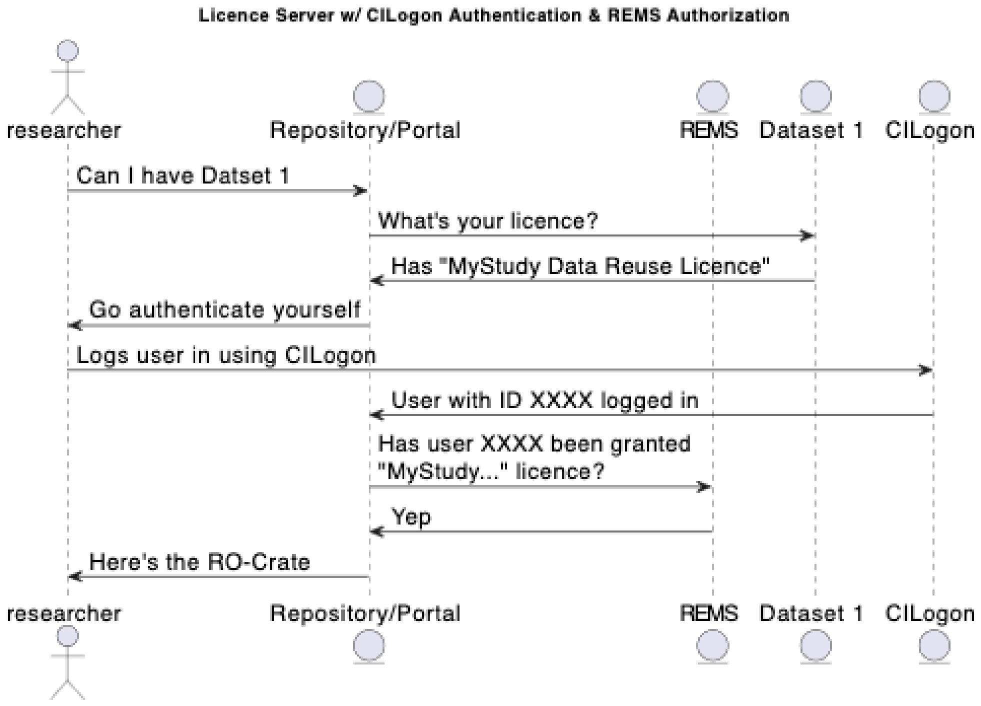


**Figure 4: The "access-control dance" for a user who has been granted a license in REMS**

Secondly, LDaCA data will be stored in a variety of places with separate portal applications serving data for specific purposes; if these systems all have in-built authorization schemes, even if they are the same, then we have the problem of synchronizing access control lists around a network of services. Thirdly, accessing data that requires some sort of authorization process is not a language or humanities specific problem, so working with an existing application that can handle pre-authorization workflows and access-control authorization decisions is an attractive choice and should allow LDaCA to take advantage of centrally managed services with relevant functionality. Fourthly, if complex access controls are implemented inside a system, then there is a risk that data becomes stranded inside that system and cannot be reused without completely re-implementing the access control. For example, imagine an archive of cultural material with complex access controls encoded into the business logic such as "this item is accessible only to male initiates". Applications like this need to store user accounts with attributes on both data and user records that can be used to authorize access. There is a high risk of data being stranded in a system such as this if it is no longer supported.

Our approach may seem to involve more work than an ACL based system. We believe that our emphasis on licenses as the basis for access control has advantages which outweigh the possibility of additional work (although we are not convinced that extra work will be needed in the long term). Reuse of data (the R in FAIR) means that users, including researchers and community members, should be able to download data for certain authorised purposes and activities. The license is the way that data custodians communicate to data users (and future administrators) what those purposes and activities are. A license, which is always packaged with data will allow:

- A user to inspect a five-year-old dataset in their downloads folder and work out what they are allowed to do with it.
- An IT professional to clean up laptop that has been handed in by (or seized from – it happens) a departing faculty member.
- A developer to re-create an access control replacing a decommissioned system.

We expect that the overhead of writing licenses will diminish greatly over time and standard clauses and complete licenses will be established.

It might seem that using REMS to administer access control means that we are locked into a specific software solution. This is not really the case; REMS is an app for establishing relationships between resources (licenses) and users. Both these components can be exported and used in another system for other purposes (e.g., auditing). In other words, if there is lock-in, it is temporary. But, because our process requires a governance step *first* in writing a license, then there is a statement of intent for re-building those processes later if needed - a step which is very likely to be missing in a system with built-in access control.

## ACKNOWLEDGMENTS

The Language Data Commons of Australia (LDaCA) project received investment from the Australian Research Data Commons (ARDC). The ARDC is funded by the National Collaborative Research Infrastructure Strategy (NCRIS).

## REFERENCES

[1] Andrea L. Berez-Kroeker and Ryan Henke. 2018. Language Archiving. In *The Oxford Handbook of Endangered Languages*, Kenneth L. Rehg and Lyle Campbell (eds.). Oxford University Press, 346–369. https://doi.org/10.1093/oxfordhb/9780190610029.013.18

[2] Daan Broeder, Freddy Offenga, Peter Wittenburg, Peter Van de Kamp, David Nathan, and Sven Strömqvist. 2006. Technologies for a federation of language resource archive. In *5th international conference on language resources and evaluation (LREC 2006)*, 2291–2294.

[3] Lisa Conathan. 2011. Archiving and language documentation. In *The Cambridge Handbook of Endangered Languages* (1st ed.), Peter K. Austin and Julia Sallabank (eds.). Cambridge University Press, 235–254. https://doi.org/10.1017/CBO9780511975981.012

[4] Nikolaus Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics* 36, 1 (1998), 161–196.

[5] Martin Matthiesen. 2015. REMS – Access Management at The Language Bank of Finland. In *DEIC Conference*. Middelfart, Denmark. Retrieved May 19, 2023 from https://gl.deic.dk/sites/default/files/uploads/PDF/Martin_Matthiesen_REMS_at_the_Language_Bank_of_Finland.pdf

[6] David Nathan. 2014. Access and accessibility at ELAR, an archive for endangered languages documentation. In *Language Documentation and Description, vol 12: Special Issue on Language Documentation and Archiving*, David Nathan and Peter K. Austin (eds.). SOAS, London, 187–208.

[7] Sefton, Peter, Ó Carragáin, Eoghan, Soiland-Reyes, Stian, Corcho, Oscar, Garijo, Daniel, Palma, Raul, Coppens, Frederik, Goble, Carole, Fernández, José M., Chard, Kyle, Gomez-Perez, Jose Manuel, Crusoe, Michael R., Eguinoa, Ignacio, Juty, Nick, Holmes, Kristi, Clark, Jason A., Capella-Gutierrez, Salvador, Gray, Alasdair J. G., Owen, Stuart, Williams, Alan R., Tartari, Giacomo, Bacall, Finn, Thelen, Thomas, Ménager, Hervé, Rodríguez-Navas, Laura, Walk, Paul, whitehead, brandon, Wilkinson, Mark, Groth, Paul, Bremer, Erich, Castro, Leyla Jael, Sebby, Karl, Kanitz, Alexander, Trisovic, Ana, Kennedy, Gavin, Graves, Mark, Koehorst, Jasper, Leo, Simone, Portier, Marc, Brack, Paul, Ojsteršek, Milan, Droesbeke, Bert, Niu, Chenxu, Tanabe, Kosuke, Miksa, Tomasz, La Rosa, Marco, Decruw, Cedric, Czerniak, Andreas, Jay, Jeremy, Serra, Sergio, Siebes, Ronald, de Witt, Shaun, El Damaty, Shady, Lowe, Douglas, Li, Xuanqi, Gundersen, Sveinung, and Radifar, Muhammad. 2023. RO-Crate Metadata Specification 1.1.3. (April 2023). https://doi.org/10.5281/ZENODO.3406497

[8] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino Da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 'T Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene Van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan Van Der Lei, Erik Van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 1 (March 2016), 160018. https://doi.org/10.1038/sdata.2016.18

# OLAC and Serials: An Appraisal

Hugh J. Paterson III
Department of Information Science
University of North Texas
United States of America
i@hp3.me

## 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 → Computers in other domains → Digital libraries and archives • Applied computing → Document management and text processing → Document management → Document metadata • Information systems → Database management systems → Database administration

## KEYWORDS

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

**ACM Reference format:**

Hugh J. Paterson III. 2023. Open Language Archives Community Aggregator and Serials: An Appraisal. In Proceedings of the 2nd International Workshop on Digital Language Archives (LangArc-2023), ACM/IEEE Joint Conference on Digital Libraries. USA, 5 pages. https://doi.org/10.12794/langarc2114296

## 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 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]. The 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. 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 the results, but the goal of the investigation was to find records which reference or represent serials across as many data providers as possible.

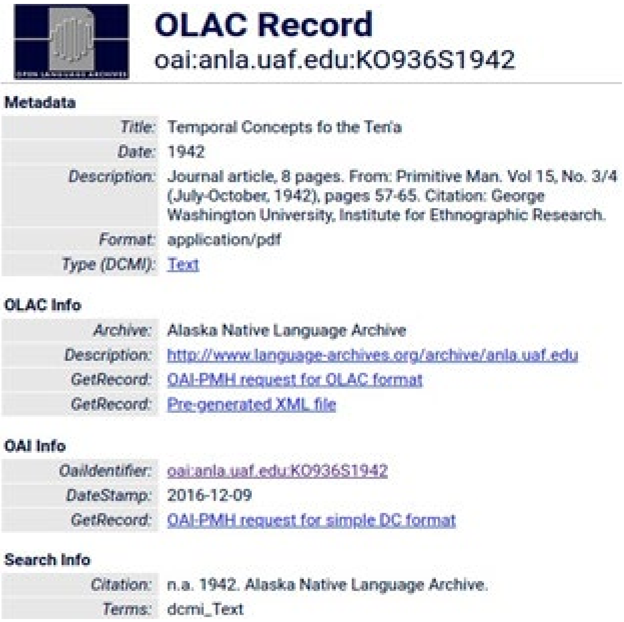
# 3 DATA

## 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 |

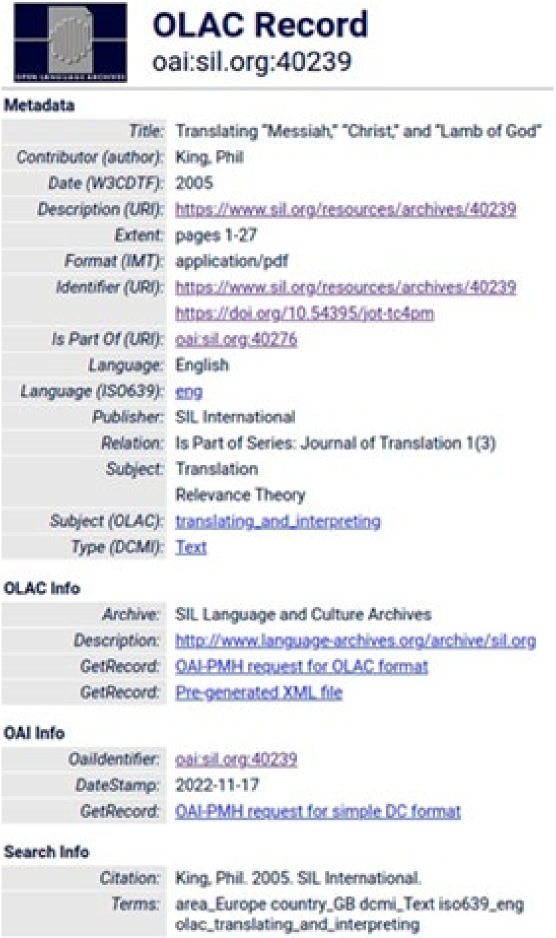
**OAI Info**

| | |
|---|---|
| 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.



**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<br>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<br>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:siLorg:40239.

## 3 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 in 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.

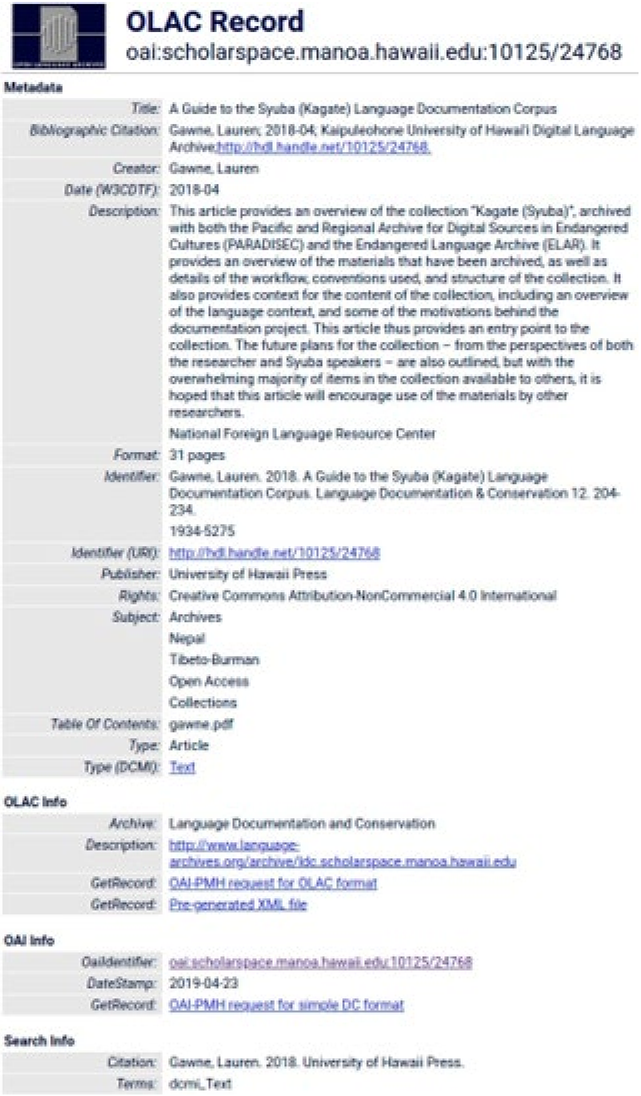
## 4.1 Description Field

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

## 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.



**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.<br>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.<br>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<br>Nepal<br>Tibeto-Burman<br>Open Access<br>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 |

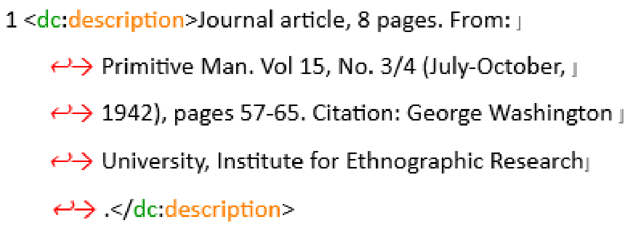
**OAI Info**

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

**Search Info**

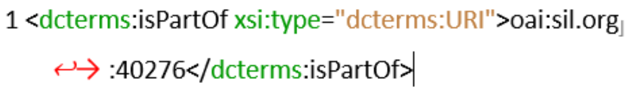
| | |
|---|---|
| Citation: | Gawne, Lauren. 2018. University of Hawaii Press. |
| Terms: | dcmi_Text |

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



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



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

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

### 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 based on 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.

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 CONCLUSION

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.

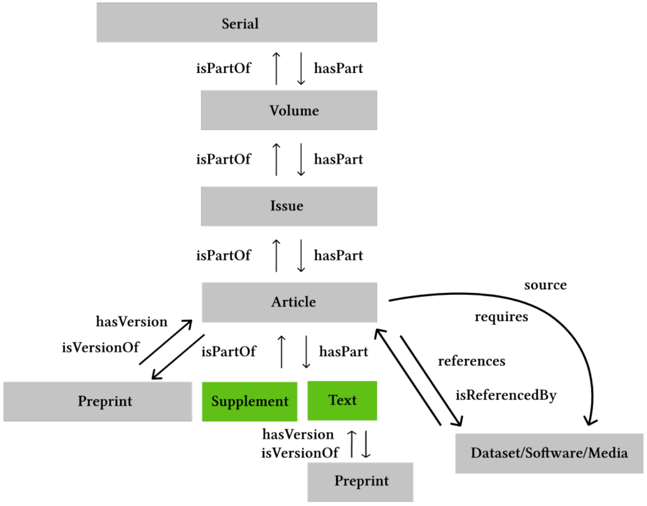


**Figure 6: Dublin Core compliant model for serials in OLAC.**

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 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 L. 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-servicesthe-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/

# Towards Making Shared Metadata Interoperable across the Open Language Archives Community

Hugh J. Paterson III
Department of Information Science
University of North Texas
United States of America
i@hp3.me

## ABSTRACT

This paper presents two methods for connecting aggregated records to their source institutional metadata profiles. The use case of the Open Language Archives Community (OLAC) application profile is considered and evaluated. The design purpose of OLAC is to share knowledge about language resources. To that end, the OLAC metadata application profile supports the exchange of metadata so that it can be aggregated and serve the needs of end-users. Uniformity in the semantic use of elements within the application profile provides the greatest utility for end-users. Discovering the source of semantic diversity remains a challenge. A first step in providing scholars access to the semantics of aggregated metadata is to publish the local metadata profiles used by institutions.

## CCS CONCEPTS

• Applied computing → Computers in other domains → Digital libraries and archives • Applied computing → Document management and text processing → Document management → Document metadata • Information systems → Database management systems → Database administration

## KEYWORDS

OAI-PMH, Metadata Semantics, Documentation, Open Language Archives Community

**ACM Reference format:**

Hugh J, Paterson III. 2023. Towards Making Shared Metadata Interoperable across the Open Language Archives Community. In Proceedings of the 2nd International Workshop on Digital Language Archives (LangArc-2023), ACM/IEEE Joint Conference on Digital Libraries. USA, 5 pages. https://doi.org/10.12794/langarc2114297

## 1 INTRODUCTION

Several scholars [13, 36, 37] note that stewards of language resources must strategize to engage with multiple audiences. That is, stewardship institutions need to consider multiple audiences and communication channels as they look to increase engagement with stewarded resources. This impacts cataloging (metadata record creation), resource discovery, and user interface design. The search and discovery process directly relates to successful stewardship. Undergirding search and discovery success is the issue of metadata quality. Yasser [38] summarizes Zeng and Qin [40] in describing the relationship between metadata quality and the ability of a digital library to meet its goals: "... poorly created metadata records result in poor retrieval and limit accessibility to collections, ultimately exercising a detrimental impact on the continuing adoption and use of a digital library. In consequence, problematic metadata is highly undesirable and needs to be understood for further action in developing remedial solutions."

The Open Language Archives Community (OLAC) has for twenty years [4, 5] provided a metadata application profile to archives and other data providers to help them meet their resource engagement goals. Recent research on language resource stewardship practices has reported on both user interface [39] and the content of description records [10]. These teams of scholars source their evidence directly via the web presentations of language resource stewards. Alongside these efforts, other work has focused on the display and presentation of records via the OLAC interface, which is often a derivative from the native metadata application profiles at institutions [29, 30]. The research that has analyzed OLAC records [29, 30] has focused on the semantics [22, 27] and usage of the metadata elements within records relative to the nature of the artifact being described. As such, it falls broadly into "metadata quality" research which investigates the accuracy, completeness, and consistency in records and across record sets [9, 23, 35]. This more recent work contrasts with previous models measuring metadata quality of OLAC records which used quantitative approaches to measure the number of elements provided per record [14].

Metadata accuracy and consistency has been addressed in large scale aggregation projects in a variety of ways, often including metadata utilities which attempt to regularize records for the benefit of end-users [12, 19, 24, 25, 31]. However, completeness can remain a challenge due to the variety of semantic options. Completeness is a measure of the totality of description versus the total possible description within the metadata schema based on an object's nature. Accessing source metadata schema documentation brings clarity to evaluation processes.

The issue of metadata quality is central to the idea of creating shareable and interoperable data which end-users will find useful in their searches [32]. Resource descriptions need to be interoperable not only at the syntactic level (Dublin Core Elements) but also at the semantic level, e.g., which elements are used and how the information values in the elements are derived. Consistent metadata quality—including semantics—is important to end-user experience. High-quality metadata is especially impactful when the engagement platform becomes dynamic or when visual representations are dependent on the content within the record. These are critical issues for the OLAC community to address if OLAC is to survive in the digital libraries' "mainstream" as Bird and Simons [4] envision.

## 2 OLAC METADATA

To fully appreciate the context of OLAC records, a deeper understanding of the processes by which OLAC records are generated is needed. For many OLAC data providers, the metadata records offered for aggregation are transformed (i.e., cross-walked) from a "native" or institutional metadata schema into the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) and Dublin Core-based OLAC metadata application profile [1–3, 33]. For example, SIL International's Language & Culture Archives uses a Dublin Core based application profile which has no public documentation but can be investigated via the open-source application RAMP [26].

Archive of the Indigenous Languages of Latin America (AILLA) has no public documentation on its website regarding its metadata schema, but it has been stated that it uses an IMDI based application profile [15]. IMDI (ISLE Meta Data Initiative) is a metadata profile which started out in Europe with projects such as DoBeS [7, 8, 17, 18]. It evolved into CMDI, a modular metadata schema used by various CLARIN entities [6, 11]. Some portion of the metadata schemas of several language resource stewards including The Endangered Languages Archive (ELAR) and The Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) can be seen in the LaMeta application's code. Broadly across aggregation efforts cross-walking metadata is a common architectural process data providers support to communicate with aggregators [16]. The conceptual process is illustrated in Figure 1.

The OLAC metadata application profile (OLAC-AP) provides specific access points for end-users to discover and engage with resource records. Data-provider- or institution-specific metadata application profiles may be designed to facilitate institution-specific user interfaces or reporting requirements. These may be in addition to, or in lieu of, access points provided via OLAC interfaces. The use of institution-specific metadata application profiles is not uncommon across digital library projects. This is the very reason that OAI-PMH was created, and that Dublin Core remains so pervasive across the digital libraries landscape—there is a clear need for (1) a generalizable super-set of metadata and (2) interface-building around generalized metadata.

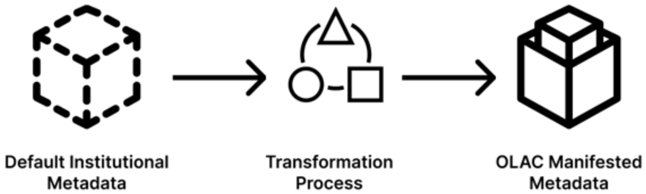


**Figure 1: Moving metadata from data-provider "local" formats to the OLAC format.**

The semantics of specific fields in institutional metadata application profiles may differ from the semantics of the most appropriate field in the OLAC-AP. Additionally, certain fields in the OLAC-AP may be inferred during the transformation process, e.g., the SIL Language & Culture Archives does not record the DCMI Type in their local application profile but generate this field for OLAC consumption based on several other factors. This means that inconsistencies or low-quality metadata may have several sources. Primary among these are signal noise via the transformation process and low-quality cataloging at the point of data origin. As metadata professionals look at OLAC metadata to evaluate record quality and interface utility for end-users, it is useful to consult the documentation for the transformation process and the institution-specific metadata application profiles. Additionally, institutional metadata application profiles and cataloging practices may evolve over time.

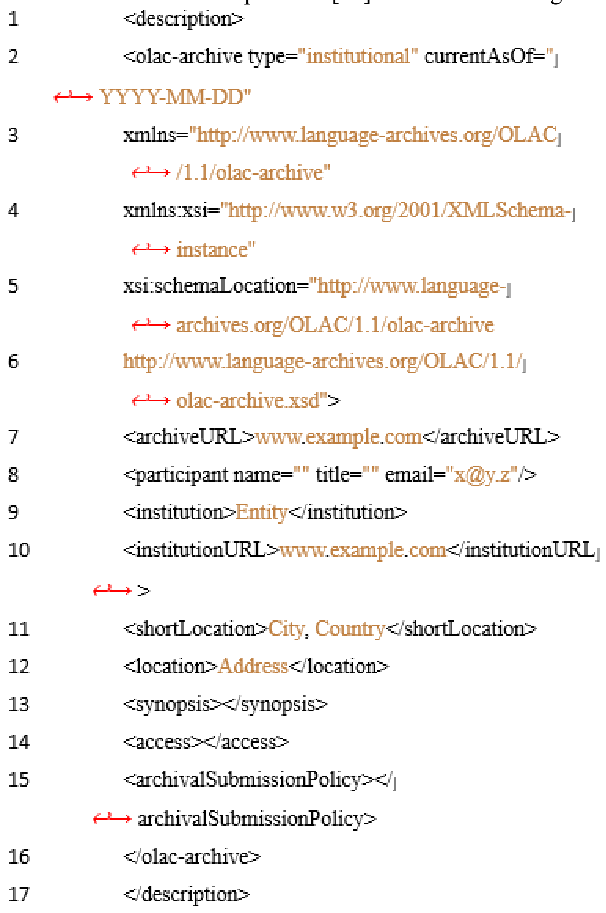
These changes may have different evolutionary cycles from metadata transformation processes. This can leave OLAC metadata in a discombobulated state while metadata is well-formed (to local standards) at data providers. However, many of the institution-specific metadata application profiles for OLAC data contributors are not accessible to the public and neither is documentation on the transformation process. The state of documentation access for OLAC data providers is not entirely out of the norm. Park and Tosaka [28], when investigating metadata aggregators, their application profiles, and the use of Dublin Core, observed that many data providers add to application profiles and frequently do not make their local metadata profiles public. They say: "the survey shows that the use of locally added homegrown metadata elements is allowed in nearly 70% of them. Only about one-fifth of local application profiles (19.6%) are made available online to the public. This means that not only is it difficult to create shareable metadata but also it is very difficult to have a quality assurance mechanism that is shareable beyond the local environment."

## 3 PROPOSAL

The rest of this paper discusses two ways in which records can be related to the cataloging schema used in their creation by the OLAC network of data providers. By granting access to metadata records via OLAC and access to the documentation for the metadata schemas at participating data providers, institutions support the flourishing of ethnolinguistic minority communities through metadata and language related artifacts, and they also support the scholarly networks which support them.

## 3.1 Modifying the data-provider description

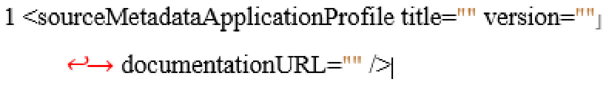
The first way in which records can be related to the cataloging schema is to add an XML element with the source schema in the data provider’s description record. The OAI-PMH implementation guidelines [21, §3.1] outline a series of optional containers. One container type provides information about a data provider. The OLAC-AP has implemented a container for providing identifying information about the data provider [34] as illustrated in Figure 2.



```
<description>
<olac-archive type="institutional" currentAsOf="YYYY-MM-DD"
xmlns="http://www.language-archives.org/OLAC/1.1/olac-archive"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.language-archives.org/OLAC/1.1/olac-archive
http://www.language-archives.org/OLAC/1.1/olac-archive.xsd">
<archiveURL>www.example.com</archiveURL>
<participant name="" title="" email="x@y.z"/>
<institution>Entity</institution>
<institutionURL>www.example.com</institutionURL>
<shortLocation>City, Country</shortLocation>
<location>Address</location>
<synopsis></synopsis>
<access></access>
<archivalSubmissionPolicy></archivalSubmissionPolicy>
</olac-archive>
</description>
```

**Figure 2: OLAC-AP structure for the description of a data provider.**

One approach to providing contextual information about the data provider’s native metadata application profile is to modify this section of the OLAC-AP to include the title, version, and location of access for the native metadata application profile used by the data contributor. Following the patterns in the existing documentation, something like what is illustrated in Figure 3 would work.



```
<sourceMetadataApplicationProfile title="" version="" documentationURL="" />
```

**Figure 3: Placement of OLAC-AP content within the OAI record structure.**

This method provides some basic access to the native metadata application profiles of OLAC data providers. This approach, however, has several drawbacks. For example, it does not specify at a record level which metadata schema or cataloging policy was current at the time a record was created. Cataloging policy can also affect metadata quality. However, if the proposed XML element were repeatable, then some change history would be accessible. A fourth attribute for dateActive="YYYY-MM-DD" would then indicate, in addition to the version number of the metadata profile, the dates a version was active. The method also does not address the change cycle in the metadata transformation technology if metadata is also transformed, which most is. A second repeatable element would be needed to track the metadata transformation technology life cycle. Good metadata application profile documentation should track changes, assigning version numbers to documentation versions and include dates of version changes within the documentation.

## 3.2 Record level association

Using OAI-PMH’s built-in record provenance feature [21, §3.4] provides a second solution for addressing the documentation of the native data provider metadata application profile. This approach would require a modification to the current OLAC database. Current OLAC architecture harvests records via OAI-PMH but only writes certain fields and attributes to the SQL database from which the User Interface is driven. The current architecture disregards any data provider information supplied within an OAI-PMH <provenance> container. Unlike the first proposed solution, the second solution applies at the record level. The OAI-PMH <provenance> container has specific elements useful for tracking both changes within the record, for example those conducted by metadata utilities after harvesting but prior to display, and sources of the record [20]. OLAC data providers can use the <provenance> container to acknowledge archival deposit curation activities.

The two options presented need not be considered mutually exclusive. That is, they can be and likely should be used in concerts. The first one provides a general link to a presumably well documented metadata schema and the second one indicates record level provenance.

## 4 CONCLUSION

In considering the future of OLAC, Bird and Simons [4] say: “we hope to shift from an idiosyncratic community-specific infrastructure to a mainstream infrastructure that interoperates with the global Web of Data”. By identifying and linking to data-provider application profiles and implementing provenance tracking for archival records, OLAC further connects with the global availability of bibliographic records. OLAC data, and the narrative of the data providers, moves towards greater transparency and interoperability. Altering OLAC infrastructure to support record level provenance will build upon OLAC’s theme of openness. Well-formed provenance records can support a variety of scholarly activity metrics demonstrating scholarly effort.

Provenance recording and semantic inference can provide the mechanisms by which a metadata utility can engage with data providers to support their metadata curation processes at a scale they would not be able to achieve independently. Such a service changes the dynamics around involvement for data providers. Instead of simply providing metadata to OLAC, the ability to receive suggestions from a metadata utility can start to prompt data providers with record level nudges related to quality enhancements.

## REFERENCES

[1] Steven Bird and Gary F. Simons. 2001. The OLAC Metadata Set and Controlled Vocabularies. In Proceedings of ACL/EACL Workshop on Sharing Tools and Resources for Research and Education, Thierry DeClerck, Steven Krauwer, and Mike Rosner (Eds.). EACL-ACL; elsnet, Université de Toulouse, France, 7–18. https://www.aclweb.org/anthology/W01-1506

[2] Steven Bird and Gary F. Simons. 2003. Extending Dublin Core Metadata to Support the Description and Discovery of Language Resources. Computers and the Humanities 37, 4 (2003), 375–388. https://doi.org/10.1023/A:1025720518994

[3] 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.

[4] 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 L. Zavalina and Shobhana Lakshmi Chelliah (Eds.). University of North Texas, Denton, Texas, 25–28. https://doi.org/10.12794/ langarc1851171.

[5] Steven Bird and Gary F. Simons. 2022. The Open Language Archives Community: A 20-Year Update. The Electronic Library 40, 5 (2022), 507–524. https://doi.org/10.1108/EL-08-2022-0192.

[6] Daan Broeder, Menzo Windhouwer, Dieter van Uytvanck, Thorsten Trippel, and Twan Goosen. 2012. CMDI: A Component Metadata Infrastructure. In Proceedings of Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LR (A Worshop at LERC 2012), Victoria Arranz, Daan Broeder, Bertrand Gaiffe, Maria Gavrilidou, Monica Monachini, and Thorsten Trippel (Eds.). European Language Resources Association (ELRA), Istanbul, Turkey, 1–4. http://www.lrec-conf.org/proceedings/lrec2012/workshops/11.LREC2012%20Metadata%20Proceedings.pdf

[7] Daan Broeder and Peter Wittenburg. 2006. The IMDI Metadata Framework, Its Current Application and Future Direction. International Journal of Metadata, Semantics and Ontologies 1, 2 (2006), 119–132. https://doi.org/10.1504/IJMSO.2006.011008.

[8] Daan Broeder, Peter Wittenburg, and Onno Crasborn. 2004. Using Profiles for IMDI Metadata Creation. In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), Maria Teresa Lino, Maria Francisca Xavier, Costa Rute, and Raquel Silvia (Eds.). European Language Resources Association (ELRA), Lisbon, Portugal. http://www.lrecconf.org/proceedings/lrec2004/pdf/513.pdf

[9] Thomas R. Bruce and Diane I. Hillmann. 2004. The Continuum of Metadata Quality: Defining, Expressing, Exploiting. ALA Editions. https://ecommons.cornell.edu/handle/1813/7895

[10] 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

[11] CMDI Taskforce. 2016. Component Metadata Infrastructure (CMDI) Component Metadata Specification – Version 1.2. Metadata Specification CE-20160880. CLARIN. 37 pages. https://office.clarin.eu/v/CE-2016-0880-CMDI_12_specification.pdf

[12] Diane I. Hillmann, Naomi Dushay, and Jon Phipps. 2004. Improving Metadata Quality: Augmentation and Recombination. In DC-2004–Shanghai Proceedings. Dublin Core Metadata Initiative — a project of ASIS&T, Shanghai, China, 11-14 October. https://dcpapers.dublincore.org/pubs/article/view/770

[13] Gary Holton. 2012. Language Archives: They're Not Just for Linguists Any More. In Potentials of Language Documentation: Methods, Analyses, and Utilization, Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts, and Paul Trilsbeek (Eds.). Number 3 in Language Documentation & Conservation Special Publication. University of Hawai'i Press, Honolulu, Hawai'i, 111–117. http://scholarspace.manoa.hawaii.edu/handle/10125/4523

[14] Baden Hughes. 2004. Metadata Quality Evaluation: Experience from the Open Language Archives Community. In Digital Libraries: International Collaboration and Cross-Fertilization, Zhaoneng Chen, Hsinchun Chen, Qihao Miao, Yuxi Fu, Edward Fox, and Ee-peng Lim (Eds.). Number 3334 in Lecture Notes in Computer Science. Springer, Berlin; Heidelberg, 320–329. https://doi.org/10.1007/978-3-540-30544-6_34.

[15] Heidi Johnson and Arienne Dwyer. 2002. Customizing the IMDI Metadata Schema for Endangered Languages. In Third International Conference on Language Resources and Evaluation: International Workshop on Resources and Tools in Field Linguistics, Las Palmas 26 - 27 May 2002, Peter Wittenburg (Ed.). LERC, Las Palmas, Canary Islands, Spain. https://www.mpi.nl/lrec/2002/papers/lrec-pap-05-JohnsonDwyer.pdf

[16] Peter Kiraly and Marco Buchler. 2018. Measuring Completeness as Metadata Quality Metric in Europeana. In 2018 IEEE International Conference on Big Data (Big Data). IEEE, Seattle, WA, USA, 2711–2720. https://doi.org/10.1109/BigData.2018.8622487

[17] Alex Klassmann, Freddy Ofenga, Daan Broeder, and Roman Skiba. 2006. IMDI Metadata Field Usage at MPI. Language Archives Newsletter 8 (2006), 6–8. https: //pure.mpg.de/rest/items/item_60279/component/file_60280/content

[18] Alex Klassmann, Freddy Offenga, Daan Broeder, Romuald Skiba, and Peter Wittenburg. 2006. Comparison of Resource Discovery Methods. In Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06), Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk, and Daniel Tapias (Eds.). European Language Resources Association (ELRA), Genoa, Italy. http://www.lrec-conf.org/proceedings/lrec2006/ pdf/808_pdf.pdf

[19] Carl Lagoze, Dean Krafft, Tim Cornwell, Naomi Dushay, Dean Eckstrom, and John Saylor. 2006. Metadata Aggregation and "Automated Digital Libraries": A Retrospective on the NSDL Experience. In Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '06). IEEE, Chapel Hill, NC, USA, 230–239. https://doi.org/10.1145/1141753.1141804

[20] Carl Lagoze, Herbert Van de Sompel, Michael Nelson, and Simeon Warner. 2002. XML Schema to Hold Provenance Information in the "about" Part of a Record. In Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting (2.0 ed.). Open Archives Initiative. http://www.openarchives.org/OAI/2.0/guidelines-provenance.htm

[21] Carl Lagoze, Herbert Van de Sompel, Michael Nelson, and Simeon Warner. 2005. Guidelines for Optional Containers. In Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting (2.0 ed.). Open Archives Initiative. http://www.openarchives.org/OAI/2.0/guidelines.htm

[22] David Loshin. 2009. Identifying Master Metadata and Master Data. In Master Data Management, David Loshin (Ed.). Morgan Kaufmann, Boston, 129–142. https://doi.org/10.1016/B978-0-12-374225-4.00007-2

[23] David Loshin. 2011. Dimensions of Data Quality. In The Practitioner's Guide to Data Quality Improvement, David Loshin (Ed.). Morgan Kaufmann, Boston, 129–146. https://doi.org/10.1016/B978-0-12-373717-5.00008-7

[24] Joshua D. Lynch, Jessica Gibson, and Myung-Ja Han. 2020. Analyzing and Normalizing Type Metadata for a Large Aggregated Digital Library. The Code4Lib Journal 47 (Feb. 2020). https://journal.code4lib.org/articles/14995

[25] Andy Neale and Valentine Charles. 2020. MS68 Metis Strategic Recommendations M18: Aggregation Strategy. Europeana DSI-4. Europeana Foundation, The Hague, Netherlands. https://pro.europeana.eu/files/Europeana_Professional/Publications/Europeana%20DSI-4%20Aggregation%20Strategy.pdf

[26] Jeremy Nordmoe. 2011. Introducing RAMP: An Application for Packaging Metadata and Resources Offline for Submission to an Institutional Repository. In Proceedings of Workshop on Language Documentation & Archiving, SOAS, London, 18 November 2011., David Nathan (Ed.). SOAS, London, UK, 27–32. https: //www.sil.org/resources/archives/43211

[27] Jung-ran Park. 2006. Semantic Interoperability and Metadata Quality: An Analysis of Metadata Item Records of Digital Image Collections. Knowledge Organization 31, 1 (2006), 20–34.

[28] Jung-Ran Park and Yuji Tosaka. 2010. Metadata Quality Control in Digital Repositories and Collections: Criteria, Semantics, and Mechanisms. Cataloging & Classification Quarterly 48, 8 (Sept. 2010), 696–715. https://doi.org/10.1080/01639374.2010.508711

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

[30] 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

[31] Julien Antoine Raemy. 2020. Enabling Better Aggregation and Discovery of Cultural Heritage Content for Europeana and Its Partner Institutions. Master of Science. Haute école de gestion de Genève, Geneva, Switzerland. https://julsraemy.ch/ assets/doc/Mastersthesis_europeana_raemyjulien_FV.pdf

[32] Sarah L. Shreeves, Ellen M. Knutson, Besiki Stvilia, Carole L. Palmer, Michael B. Twidale, and Timothy W. Cole. 2005. Is "Quality" Metadata "Shareable" Metadata? The Implications of Local Metadata Practices for Federated Collections. In Proceedings of the Twelfth National Conference of the Association of College and Research Libraries, April 7-10 2005, Minneapolis, MN, H. A. Thompson (Ed.). Association of College and Research Libraries, Chicago, IL, 223–237. http://hdl.handle.net/2142/145

[33] Gary F. Simons and Steven Bird. 2003. Building an Open Language Archives Community on the OAI Foundation. Library Hi Tech 21, 2 (2003), 210–218. https://doi.org/10.1108/07378830310479848

[34] Gary F. Simons and Steven Bird (Eds.). 2008. OLAC Repositories (2008-07-28 ed.). Open Language Archive Community, Dallas, TX. http://web.archive.org/web/20230418175450/http://www.language-archives.org/OLAC/repositories.html

[35] Besiki Stvilia, Les Gasser, Michael B. Twidale, and Linda C. Smith. 2007. A Framework for Information Quality Assessment. Journal of the American Society for Information Science and Technology 58, 12 (Oct. 2007), 1720–1733. https://doi.org/10.1002/asi.20652

[36] Christina Wasson, Melanie Medina, Miyoung Chong, Brittany LeMay, Emma Nalin, and Kenneth Saintonge. 2018. Designing for Diverse User Groups: Case Study of a Language Archive. Journal of Business Anthropology 7, 2 (Nov. 2018), 235–267. https://doi.org/10.22439/jba.v7i2.5605

[37] Anthony C. Woodbury. 2014. Archives and Audiences: Toward Making Endangered Language Documentations People Can Read, Use, Understand, and Admire. In Special Issue on Language Documentation and Archiving, David Nathan and Peter K. Austin (Eds.). Number 12 in Language Documentation and Description. SOAS, London, 19–36. http://www.elpublishing.org/PID/135

[38] Chuttur M. Yasser. 2011. An Analysis of Problems in Metadata Records. Journal of Library Metadata 11, 2 (April 2011), 51–62. https://doi.org/10.1080/19386389.2011.570654

[39] Irene Yi, Amelia Lake, Juhyae Kim, Kassandra Haakman, Jeremiah Jewell, Sarah Babinski, and Claire Bowern. 2022. Accessibility, Discoverability, and Functionality: An Audit of and Recommendations for Digital Language Archives. Journal of Open Humanities Data 8, 10 (March 2022), 1–19. https://doi.org/10.5334/johd.59 .

[40] Marcia Lei Zeng and Jian Qin. 2008. Metadata. Neal-Schuman Publishers, New York. http://catdir.loc.gov/catdir/toc/ecip0816/2008015176.html.

# Ukrainian Archival Metadata in WorldCat: Exploratory Analysis

Vyacheslav I. Zavalin
School of Library & Information Studies
Texas Woman's University
United States of America
vzavalin@twu.edu

## ABSTRACT

The purpose of metadata is to enable information users to find, identify, select, obtain, and explore information resources. The largest global database of machine-readable bibliographic metadata, WorldCat includes over 500 million records that represent information resources in 483 languages. While most of these records describe individual officially published or released materials (print and electronic books and journals, VHS, CD, and DVD releases of documentary and feature films, officially distributed audio albums of songs and instrumental music), over 2.3 million of metadata records included in WorldCat represent archival materials of various kinds. Our study examined a sample of WorldCat records representing Ukrainian-language archival materials (including digital resources) that were not officially published or released with the goal to examine the extent to which these metadata records support the user tasks of find, identify, select, obtain, and explore.

## CCS CONCEPTS

• Applied computing → Computers in other domains → Digital libraries and archives • Applied computing → Document management and text processing → Document management → Document metadata • Information systems → Document representations; • General and reference → Evaluation

## KEYWORDS

information organization, information access, metadata evaluation, Ukrainian language archives, archival records, MARC21

**ACM Reference format:**

Vyacheslav I. Zavalin. 2023. Ukrainian Archival Metadata in WorldCat: Exploratory Analysis. In Proceedings of the 2nd International Workshop on Digital Language Archives, ACM/IEEE Joint Conference on Digital Libraries. USA, 4 pages. https://doi.org/10.12794/langarc2114298

## 1 INTRODUCTION AND BRIEF REVIEW OF LITERATURE

The International Federation of Library Associations and Institutions formulates 5 tasks of information users that metadata should support so that resources described by this metadata can be discovered and utilized. These user tasks include finding, identifying, selecting, obtaining, and exploring information [2]. The metadata record supports a user task if it contains information important for satisfying the associated user need. For example, to support the obtain user task, the metadata record that represents an online resource must include the URL where this resource is accessible.

Society of American Archivists' *Dictionary of Archives Terminology* defines archival materials broadly as "records in any format retained for their continuing value […]" (https://dictionary.archivists.org/entry/archival-material.html) and further explains that archival materials' synonym term archival records "connotes documents rather than artifacts or published materials […] may be in any format, including text on paper or in electronic formats, photographs, motion pictures, videos, sound recordings" (https://dictionary.archivists.org/entry/archival-record.html). Information access to archival materials – individual items and often collections – is provided through metadata that describes these items and collections, and allows the users to discover archival resources of interest to them, to support their learning, teaching or research of history, culture, language, etc.

International archival description standards ISAD[G] and ISAAR[CPF] as well as their U.S. implementation *Describing Archives: A Data Content Standard (DACS)* list typical attributes of archival resources and provide guidelines on how to identify and represent these attributes. For example, Part 1 of the DACS document suggests that every metadata record representing an archival resource must include 6 "identity elements": reference code, name and location of the repository, title, date, extent, and name(s) of creator(s) – if these names are known – in addition to elements from other element groups: scope and content, conditions governing access, and languages and scripts of the materials.

DACS guidelines apply to representing both individual archival items and collections of these items. In the archiving profession, collection-level descriptions that represent entire collections are very common traditionally, and as collections can be very large, often there is only a collection-level metadata record for an archival collection. The major metadata standard developed specifically to be used for representing archival collections is *Encoded Archival Description (EAD)*, currently in version 3, issued in 2019.

The databases that provide centralized access to archival metadata exist at state or country levels (for example, Texas Archival Resources Online), as well as at the regional and international level, with the largest such database ArchiveGrid aggregating over 7 million metadata records. Many metadata records follow the EAD standard: according to the ArchiveGrid Index Growth, as of April 2022 – the latest date for which information is available – the database contained almost 213 thousand EAD records. This represents approximately a 6-fold increase since February 2011 when ArchiveGrid first started including EAD records and reported having 35470 of them (https://researchworks.oclc.org/archivegrid/about/). Most archival metadata records that are included in the ArchiveGrid discovery tool, however, follow the Machine-Readable Cataloging (MARC) metadata standard developed in the library community in the late 1960s and widely used around the globe today. MARC, maintained by the United States Library of Congress, which also maintains EAD, is currently called *MARC21 Bibliographic Format*, and is updated at least twice a year to keep up with the technological developments, evolving user needs and expectations, and trends in organizing information.

As of April 2023, the largest global database of MARC21 metadata, WorldCat includes more than 547 million records that represent various information resources held at libraries, archives, museums, and other organizations in many countries (https://www.oclc.org/en/worldcat/inside-worldcat.html). As of the time of writing this paper, the WorldCat search for archival materials results in the notification that 2328586 metadata records in the database are for resources of this broad type.

The current WorldCat statistics indicate that this database represents resources in 483 languages. As of the time of writing this paper, the WorldCat search for Ukrainian-language resources results in the notification that 858712 metadata records in the database are for resources in this Cyrillic-script language that belongs to the East Slavic group of languages. Of these, 1953 WorldCat metadata records represent Ukrainian-language archival resources.

MARC21 records representing non-Latin script resources traditionally contain Romanized information that is generated by metadata creators by following the US Library of Congress Romanization tables (https://www.loc.gov/catdir/cpso/roman.html ). For example, to enter in the metadata record the Ukrainian-language title "Мавка. Лісова пісня" of the 2023 Ukraine-produced animated picture *Mavka. The Forest Song* (based on the famous 19th -century folklore-inspired poetic play by the Ukrainian author Lesia Ukrainka), a metadata creator would follow the Ukrainian transliteration table (https://www.loc.gov/catdir/cpso/romanization/ukrainia.pdf) to obtain the following Romanized text: "Mavka. Lisova pisnia".

In the past 15 years, the vernacular text (such as "Мавка. Лісова пісня") is sometimes being added to Romanized text in the WorldCat records representing non-Latin-script resources, including Ukrainian-language resources. The Smith-Yoshimura's report shows that as of December 2015, 8% of WorldCat records representing Ukrainian-language materials included the text in the original Cyrillic script in addition to Romanized text [5]. Toves at al. reported on intensifying these efforts for Ukrainian-language materials in recent years [7]. The WorldCat search conducted in May of 2023 demonstrates that the number of those metadata records representing Ukrainian-language materials that include any Cyrillic script data has increased to 112961 (or 13.45%).

Over the years, many studies with a focus on MARC21 metadata records have been published. This includes publications resulting from the study conducted in the 2000s that examined all (over 50 million) metadata records contained in WorldCat at the time [1, 3, 4] as well as a recent study that examined the data in various fields and subfields of a large sample of WorldCat records that were added to the database in 2020 [8]. Also, one study looked at the quality of the MARC21 records representing materials in Slavic languages [6]. However, no research has yet analyzed the WorldCat metadata records that represent Ukrainian-language archival resources, particularly in relation to the user tasks metadata is aimed to support. This study sought to address this gap.

## 2 METHOD

This study relied on the content analysis (qualitative and quantitative) of the sample of 339 MARC21 WorldCat records. The sample selection criterion was that the data in MARC21 field 008 subfields Ctrl and Lang indicates that the record represents a Ukrainian-language archival resource and that the keyword "Ukraine" is included in any other MARC21 field(s). The goal of the study was to explore the extent to which the user tasks of *find*, *identify*, *select*, *obtain*, and *explore* are supported in these metadata records.

The following questions were addressed by this study:

1. How do the records support the *find* user task by including:
   1.1. Vernacular data in the Cyrillic script?
   1.2. Fields representing creators of archival materials and what these archival resources are about?
2. How do the records support the *identify* user task by:
   2.1. Correctly marking the resource as an archival material?
   2.2. Including fields that represent the language?
3. How do the records support the *select* user task by providing:
   3.1. A descriptive summary for any archival resource?
   3.2. A list of items in the archival collection?
   3.3. A link to the archival finding aid?
   3.4. Information about duration of the sound recording or video recording in records representing individual archival resources (e.g., oral histories)?
   3.5. Information about file formats and system requirements for electronic resources?
4. How do the records support the *obtain* user task by providing:
   4.1. Access conditions/restrictions information?
   4.2. A URL for a digital resource or a call number for an analog resource?
5. How do the records support the *explore* user task by providing:
   5.1. Information on relations of the resource described by the record to persons, organizations?
   5.2. Provenance (aka custodial history) information?

## 3 PRELIMINARY FINDINGS AND DISCUSSION

The first research question addressed was 2.1. Our qualitative examination of the 339 records revealed that 34 of them (10.03%) did not meet the criteria of this study. They represented materials that are officially published, released, or distributed. These 34 records were excluded from further analysis, resulting in 305 records.

With regards to the *find* user task, finding Ukrainian-language archival resources by representation of its aboutness was overall well supported: 93.44% of records included the topical subject headings (with over 9 per record on average), 86.23% included representation of the geographic location the archival resource is about, and between 32.77% and 39.67% records included names of persons and/or institutions the resource is about. Similarly, discovery by creator name was supported at a relatively high level: 80.33% of records included the personal creator's name, and 9.51% included the name of the institution that created archival resource for a total of almost 90% of records. However, a very low level of inclusion of vernacular data was observed: only 4 records out of 305 included Cyrillic-script data in one or more fields: usually the title, and sometimes subject added entry name headings.

The *identify* user task was moderately supported overall. In addition to 10.03% of records marked as those that represent archival resources, in fact representing officially published, distributed, or released materials, only 57.7% of records included one or both of MARC metadata fields representing languages: 041 Language Code and/or 546 Language Note.

The findings regarding the support for the *select* user task were mixed. For example, 93.44% of records included a descriptive summary note (field 520); however, all these summaries were provided only in the English language. The contents note (field 505) was only included in 4.92% of records, yet the alternative field 555 Cumulative Index/Finding Aids, although only added to MARC21 standard in July 2022, was found to be already used in 29.51% of records in the sample. The system details note (field 538) that applies to representing any video or audio recording was only found in 1 record: 1.22 % of 82 records representing video recordings or mixed materials (i.e., archival collections) that include video recordings. A positive observation was that of these 82 records, 69 (84.15%) included information about the duration of the recording that is important for selecting an information resource.

The *obtain* user task was found to be weakly supported by the WorldCat MARC21 metadata records representing Ukrainian-language archival resources. Only 26.56% of records included information relating to copyright status (field 542), 12.79% of records included information about terms governing use and reproduction (field 540), and 9.51% of records included information on restrictions on access (field 506). At the same time, though, the URLs leading to either the digital archival resource itself, or the archival collection's finding aid, or online transcripts were included in 39.34% of records.

The study also revealed a low to moderate level of support for the *explore* user task, which emphasizes the representation of relationships in metadata. Biographical or Historical Data note (field 545) was included in over half (55.41%) of records, but information about ownership and custodial history (field 561) or about the immediate source of acquisition (field 541) only appeared in between 2.3% and 4.92% of records.

## 4 CONCLUSION

This exploratory study was the first one to examine MARC21 metadata records that represent Ukrainian-language archival materials, with the focus on supporting the user tasks. Additional qualitative metrics for WorldCat MARC21 archival resources metadata support of user tasks, beyond those examined in this study, would be beneficial, for example:

- Using the correct standard Romanization table in the romanization of data for non-Latin script materials
- Correctness of spelling of titles and names (to the extent this is possible to assess without examining the resource itself)
- Providing clear, non-misleading information in the record: e.g., representation of the language, digital/analog status, and other attributes of the archival resource.

The level of inclusion of vernacular Cyrillic-script data observed in this study was very low and significantly lower than the overall for Ukrainian-language resources that are represented with over 850 thousand metadata records in WorldCat (1.31% compared to 13.45%). This is concerning and warrants future studies to trace the trends in increasing access to archival materials for the native speakers of Ukrainian who are not proficient in English. Similarly, future studies need to examine trends in supporting *obtain* and *explore* user tasks, the level of support for which is currently insufficient, as shown by the findings of this study.

Future studies will examine the entire dataset of WorldCat MARC21 records representing Ukrainian-language archival materials, as well as ArchiveGrid MARC21 and/or EAD records that match this criterion.

## REFERENCES

[1] Eklund, A.P., Miksa, S.D., Moen, W.E., Snyder, G., & Polyakov, S. (2009). Comparison of MARC content designation utilization in OCLC WorldCat records with national, core, and minimal level record standards. Journal of Library Metadata, 9, 36-64.

[2] International Federation of Library Associations and Institutions. (2017). IFLA Library Reference Model: A Conceptual Model for Bibliographic Information. Retrieved from: https://www.ifla.org/files/assets/cataloguing/frbr-lrm/ifla-lrm-august-2017.pdf

[3] Moen, W.E., & Benardino, P. (2003). Assessing metadata utilization: an analysis of MARC content designation use. In Proceedings of the International Conference on Dublin Core and Metadata Applications (Seattle, WA, Sept. 28 - Oct.2, 2003). Retrieved from https://dcpapers.dublincore.org/pubs/article/view/745

[4] Moen, W.E., Miksa, S.D., Eklund, A., Polyakov, S. & Snyder, G. (2006). Learning from artifacts: Metadata utilization analysis. In Proceedings of the Joint Conference on Digital Libraries, June 11-15, 2006, Chapel Hill, NC.

[5] Smith-Yoshimura, K. (2015). Moving towards true multilingualism: Leveraging global cooperation through WorldCat. Retrieved from http://www.slideshare.net/oclcr/moving-towards-true-multilingualism-leveraging-global-cooperation-through-worldcat.

[6] Soglasnova, L. (2018). Dealing with false friends to avoid errors in subject analysis in Slavic cataloging: an overview of resources and strategies. Cataloging & Classification Quarterly, 56(5/6), 404-421.

[7] Toves, J., Tashlitskyy, R., & Soglasnova, L. (2021). The Ukrainian Kyrylytsia, restored: An automation project for adding the Cyrillic fields to Ukrainian records in OCLC WorldCat. East/West: Journal of Ukrainian Studies, 8(2), 307–320. https://doi.org/10.21226/ewjus626

[8] Zavalin, V., Zavalina, O.L., & Miksa, S.D. (2021). Exploration of subject representation and support of Linked Data in recently created library metadata: Examination of most widely held WorldCat bibliographic records. Library Resources and Technical Services, 65 (4), 154-165. Retrieved from https://journals.ala.org/index.php/lrts/article/view/7519.

# Language Archiving Training: A Case Study of a Metadata Course in Library and Information Science Graduate Program, 2020 - 2023

Oksana L. Zavalina
Department of Information Science
University of North Texas
United States of America
Oksana.Zavalina@unt.edu

## ABSTRACT

Since the early 21st century, funding agencies have been continuously supporting efforts aimed at language preservation and revitalization. This includes providing online access to unique and valuable collections of language data, which often originates from Indigenous and endangered language communities. Language materials are organized and represented in digital archives mostly by information professionals in the library, museum, and archival fields. However, a gap exists between the way these materials are organized and represented and the understanding of that data – and expectations towards the more functional ways of its organization and representation – by language preservation and revitalization researchers, and by members of language communities. Information resources collected by language archives have unique attributes of importance to their target user groups, and these attributes and their representation are not currently widely addressed by the formal training provided to information professionals. Similarly, specifics of these collections end-users' information needs are not currently examined in this training. In this case study, the project that seeks to address this training gap is presented and its preliminary results are evaluated.

## CCS CONCEPTS

• Applied computing → Computers in other domains → Digital libraries and archives • Applied computing → Document management and text processing → Document management → Document metadata • Social and professional topics → Professional topics → Computing education → Model curricula
• General and reference → Evaluation

## KEYWORDS

Language archiving training, metadata training, graduate education, Library and Information Science curriculum.

**ACM Reference format:**

Oksana L Zavalina. 2023. Language Archiving Training: A Case Study of a Metadata Course in Library and Information Science Graduate Program, 2020-2023. In Proceedings of the 2nd International Workshop on Digital Language Archives (LangArc-2023), ACM/IEEE Joint Conference on Digital Libraries. USA, 4 pages. https://doi.org/10.12794/langarc2114299

## 1 INTRODUCTION

Language archives provide access to various kinds of materials, including those unique for them (e.g., word lists), and those that other kinds of archives also frequently hold (e.g., notebooks, oral histories, recordings of community cultural events, etc.). The Open Language Archives Community (OLAC) specializes in providing access to language archival resources. Its Language Resource Catalog includes metadata records representing over 170 thousand resources held by over 60 archives around the world. For example, 36 of these metadata records represent individual items in archival collections related to Amdo Tibetan language, such as the *Medical Secretary and Doctor in Sokdzong (Sokdzong)* Amdo-Tibetan-language oral history transcript held by *COllections de COrpus Oraux Numeriques (CoCoON ex-CRDO)* archive in France. As can be seen in this example record, OLAC metadata records follow the Dublin-Core-based OLAC metadata scheme and make use of OLAC-developed controlled vocabularies.

OLAC is not the only centralized portal through which one may access metadata records representing language archive resources. As of June 2023, two well-known global metadata aggregators include a large number of records representing archival resources: over 2.3 million in the WorldCat database and over 7 million in the ArchiveGrid. It is not clear how many of these metadata records represent language archive resources, and more specifically digital ones: this is not one of the metrics published by developers of these aggregators. However, the estimates can be obtained by searching these databases for archival resources in a specific language.

For example, searching the ArchiveGrid by the phrase "Maori language" retrieves 60 exact matches. Some of these metadata records represent archival items (e.g., *Draft for unpublished second edition of the Grammar of the New Zealand language, 1827-1832 [by Kendall, Thomas]*). Most records though represent archival collections: for example, *United States, Indiana, Bloomington, Polynesian languages, 1949-1957* collection of word lists, dialect texts, speeches, grammatical statements with examples, and other text held by the University of Indiana Libraries.

Similarly, a WorldCat search combining the type of resource (archival material) and language (Cherokee) queries reveals that WorldCat currently includes 22 metadata records categorized as representing Cherokee-language archival resources. This includes some individual items such as *Cherokee Nation's record book* for 1902-1903 years. Most of the records though represent archival collections such as for example William West Long's *Cherokee Medicinal and Magical Texts, 1928-1936* collection of 2 notebooks (recorded in the Sequoyan syllabary) and 91 other items held by the American Philosophical Society Library.

Searches like the ones presented above demonstrate that language archive materials are largely held by libraries of various kinds. Thus, metadata to represent these resources is created by library professionals, sometimes in collaboration with those linguists who collected archival materials or with language community members [3] but often – as is the case with legacy materials – without. In general, providing access to legacy data in digital language archives presents several challenges, including those related to provenance, orphan data, and citation tracking [12].

There is clearly the need for providing the information professionals with training that would allow them to identify those attributes of resources in language archives that are important for the users (linguistics and language speakers, instructors, or learners) and to accurately represent these attributes in metadata. Until recently, such training was not provided. As a result, digital language archive materials are often made available to users in a less functional way (e.g., [1], [11]).

Research also demonstrates what specifically is missing in how digital language archival materials are represented. For example, interviews and observations revealed that, for many users, the Language element of metadata records, as well as and representations of the relationships between items (e.g., an audio recording and textual transcriptions or translations) are most important in their interactions with language archives, yet sometimes not represented [4, 6]. Most of respondents in Burke et al. [4] study noted that multilingual interfaces would enable more users to access digital language archive data. Some users also noted that maps displaying the geographic area where the languages are spoken would allow them to find materials easily.

The formal training in digital language archiving offered to information professionals needs to reflect these user preferences, as well as the best practices available. Some best practices for digital language archives were shared by teams of researchers and practitioners from different countries. For example, R and Takhellambam [10] presented the case study of the Sikkim-Darjeeling Himalayas Endangered Language Archive and discussed the collaborative digital language archive development in India. Two studies shared the experiences of the Computational Resource for South Asian Languages (CoRSAL) digital language archive: Dale [7] presented the approaches tested in the development of workflow for mediated archiving while Burke and colleagues [5] discussed the challenges and proposed solutions for name and subject representation in the digital language archive metadata.

Several researchers of digital language archives concluded that education for information professionals needs to cover the specifics of applying archiving techniques and tools in language archives. This paper reports on the project that seeks to address this curricular need, presents intermediate results of this project, and discusses next steps. This paper extends the early results report that was presented at the Association for Library and Information Science Education annual meeting in 2021 [13].

## 2 COURSE DESIGN

After the initial experiment teaching an interdisciplinary digital language archives metadata course to a combined class of linguistics and library and information science graduate students in the Spring of 2020, the decision was made that to maximize the benefits of this coursework, our team would need to develop the modules focusing on digital language archives and integrate them in relevant courses for information professionals and for linguists. The first candidate for inclusion of such a module was the advanced elective graduate course with the focus on digital library metadata that had participated in the initial experiment. During the Fall of 2020, we developed a digital language archives metadata learning module to integrate in the course and revised the other existing modules to draw examples from language archives in both lectures and assignments.

The new version of the course was tested in the Spring 2021 semester and was taught two more times since then: in Spring semesters of 2022 and 2023, with the cumulative enrollment of 42. To enroll in this course, students must successfully complete the core course in the fundamentals of information organization, followed by the introductory digital metadata course. In the immediate prerequisite (introductory metadata course), students develop knowledge and skills related to the application of major metadata standards, This includes use of data content standards, data value standards (controlled vocabularies), data encoding and transmission standards (XML, HTML, and to some extent MARC 21), and major metadata elements sets for metadata creation to describe items (Dublin Core DCTERMS, MODS, VRA Core 4.0) and collections (Dublin Core Collection Application Profile, Encoded Archival Description, MODS collection application profile, and use of VRA Core 4.0 collection record type). Learning materials of this prerequisite introductory metadata course discuss the user needs and their role in developing metadata element sets, controlled vocabularies, etc., and providing access to information, at the general level, with some examples.

This training prepares students to closely examine the metadata principles and tools in relation to digital language archives, in the advanced digital library metadata course which consists of 4 modules. With the course so far offered only in 16-week spring semesters, the class spends 4 weeks on each learning module. In the weekly class meetings, the teaching team presents material in an interactive way, with numerous illustrative examples, brainstorming and mini-exercise activities for students to help digest the content.

The course opens with Module *1. Metadata for Cultural Works and Specialized User Communities: Language Documentation Case Study* focused entirely on digital language archives. The learning objectives of this module are:

- Identify the needs of a specialized user community, types of materials of interest to these users, general and specific metadata standards that can be utilized in representing these materials for these audiences. Implement this knowledge in metadata work, including investigating relations between metadata elements and user tasks based on conceptual models, navigating controlled vocabularies, and selecting appropriate terms.
- Examine and evaluate current trends in metadata theory and practice, as well as perspectives of developing and applying metadata to provide effective information access for specialized user communities.

During the first week of a learning module, students participate in the class meeting (or review posted slides and recording) and select and read 2 items from the list of 20 or more relevant peer-reviewed professional and/or research publications prepared by the teaching team. These readings are then summarized and critiqued by each student in the discussion post. Students read each other's discussion posts and react to them, with the requirement to provide a substantial reaction to at least one of their classmates' discussions.

Each module has a major practical assignment, that a student completes in weeks 2-4 of the module. For the Module 1 that focuses on digital language archives, the practical assignment includes two parts. In Part 1 (Language Materials, their Users, User Tasks, and Metadata), students answer 4 blocks of questions based on their understanding and critical evaluation of the documentary linguistics workflow and types of materials collected by linguists, as well as user tasks and the specific ways in which metadata fields in a record address them as discussed in two models: the Functional Requirements for Bibliographic Records, and the IFLA Library Reference Model [8, 9]. In Part 2 (General and Specialized Controlled Vocabularies for Representing Resources in Language Collections to Facilitate Information Access), students make use of 16 data value standards (including 4 OLAC controlled vocabularies) to find and examine authority records or other controlled vocabulary entries for terms, names, and codes relevant for representing digital language archive materials.

In the remaining 3 learning modules – Metadata Quality, Metadata Interoperability, and Metadata as Linked Data – examples from digital language archives are used as much as possible, to keep engaging students with the issues related to digital language archives throughout the course. *Module 2. Metadata Quality* also has a significant digital language archiving component. In its practical assignment, students collect and analyze a sample of metadata records from a collection in the CoRSAL digital language archive based on three major metadata quality criteria (accuracy, completeness, and consistency) defined in Bruce and Hillmann [2]. Students compare results of this evaluation to those for a metadata sample in another (non-language-focused) collection that is hosted by the same institution and relies on the same metadata scheme.

## 3 LEARNING EFFECTIVENESS

This advanced graduate metadata course with the content focusing on digital language archives is overall well received by students, with student satisfaction scores in 2021-2023 ranging between 4.1 and 5.0 on a 5-point scale (response rate 50% - 90%). Here we present some preliminary results of basic quantitative evaluation of students' performance in the two modules with significant digital language archives content components.

To measure effectiveness of the digital language archiving learning in this course, we developed the following targets:

1. Individual target: each student receives at least 85% of possible cumulative points for 3 assessments. For that evaluation, we selected both assessments in Module 1 (discussion forum, and practical assignment) and a practical assignment in Module 2.
2. Class target: at least 90% of students meet the individual target.

In Spring 2021 when our grading was lenient because this was the first semester this course was offered in its current form – all 12 enrolled students (100%) met the individual target, with the average score of 95.66% and the median of 96.61% of possible cumulative points. In Spring 2022, 15 out of 20 enrolled students (75%) met the individual target. However, the average and median scores were quite high: 88.07% and 90.48% of cumulative possible points. Also, when only looking at Module 1 that solely focused on digital language archives, the Spring 2022 results were higher: 85% of students met the individual target, with the average of 92.35% and the median of 92.94%. In the most recent semester (Spring 2023), 90% of enrolled students met the individual target. The average percentage of possible cumulative points achieved by the student was 91.41%, and the median was 91.8%.

Our next step would be to conduct a more detailed analysis of digital language archives learning effectiveness using the available data. For example, we would investigate which type(s) of questions on the practical exercise in the digital-language-archives-focused Module 1 students tend to perform better and worse on. This would allow us to assess the implications for further development and improvement of training materials. Also, detailed examination of student feedback on accuracy and completeness of metadata representing items in the CoRSAL archive collections obtained as part of Module 2 exercise will help identify areas of improvement for CoRSAL metadata.

## 4 CONCLUSION

Our report presents a case study of the graduate course that begins to bridge the gap in information professionals' understanding of digital language archives users and their needs, materials included, and metadata needed. It will be useful for other educators working on addressing this curricular need.

Overall, the results meet our expectations yet the observation that digital language archives learning effectiveness was lower in the semester with the highest so far (yet still reasonable) enrollment of 20 students warrants further monitoring.

The course in question focuses on metadata, so some other important aspects of the digital language archives are outside of its scope, and either were not covered or did not have a practical assignment (or its component) addressing them. As more of the relevant courses for librarians and archivists are starting to integrate content that develops knowledge and skills necessary to successfully manage digital language archives, future studies will need to compare the instructional approaches, course materials, and results, with the goal of improving such training.

## REFERENCES

[1] Al Smadi, D. et al. (2016). Exploratory user research for CoRSAL [language archive]: report prepared for the Computational Resource for South Asian Languages. University of North Texas. Retrieved from https://digital.library.unt.edu/ark:/67531/metadc1707416/

[2] Bruce, T. R., & Hillmann, D. I. (2004). The continuum of metadata quality: defining, expressing, exploiting. ALA editions.

[3] Burke, M. (2021). Collaborating with Language Community Members to Enrich Ethnographic Description in a Language Archive. In Proceeding of LangArc-2021 (1st International Workshop on Digital Language Archives), 18-21. https://doi.org/10.12794/langarc1851172

[4] Burke, M., Zavalina, O. L., Chelliah, S.L., & Phillips, M. E. (2022). User needs in language archives: Findings from interviews with language archive managers, depositors, and end-users. Language Documentation & Conservation, 16, 1-24. Retrieved from https://scholarspace.manoa.hawaii.edu/handle/10125/74669

[5] Burke, M., Tarver, H., Phillips, M.E., & Zavalina, O. (2022). Using existing metadata standards and tools for a digital language archive: a balancing act. The Electronic Library, 40 (5), 579-593. https://doi.org/10.1108/EL-02-2022-0028

[6] Burke, M., Zavalina, O. L., Phillips, M. E., & Chelliah, S. (2021). Organization of knowledge and information in digital archives of language materials. Journal of Library Metadata, 20(4), 185-217. https://doi.org/10.1080/19386389.2020.1908651

[7] Dale, M. (2022). Creating workflow for mediated archiving in CoRSAL, The Electronic Library, 40 (5), 568-578. https://doi.org/10.1108/EL-02-2022-0027

[8] International Federation of Library Associations and Institutions. (2008). Functional Requirements for Bibliographic Records. Retrieved from https://cdn.ifla.org/wp-content/uploads/2019/05/assets/cataloguing/frbr/frbr_2008.pdf

[9] International Federation of Library Associations and Institutions. (2017). Library Reference Model. Retrieved from https://repository.ifla.org/bitstream/123456789/40/1/ifla-lrm-august-2017_rev201712.pdf

[10] R., K.N. & Takhellambam, M. (2022). A collaboratory model for creation of digital language archives in India. The Electronic Library, 40 (5), 594-606. https://doi.org/10.1108/EL-02-2022-0030

[11] Wasson, C., Holton, G., & Ross, H. (2016). Bringing user-centered design to the field of language archives. Language Documentation and Conservation, 10, 641-671. Retrieved from http://hdl.handle.net/10125/24721

[12] Weber, T. (2022). Conceptualising language archives through legacy materials. The Electronic Library, 40 (5), 525-538. https://doi.org/10.1108/EL-02-2022-0029

[13] Zavalina, O.L., & Chelliah, S.L. (2021). Exploring language archiving education for information professionals and interdisciplinary collaboration to support information access. Proceedings of the Association for Library and Information Science Education Annual Conference: ALISE 2021. Seattle: ALISE. Retrieved from: https://www.ideals.illinois.edu/items/118795.