Best Practices for Information Architecture, Organization, and Retrieval in Digital Language Archives within University Institutional Repositories

Robert E. Vann
Department of Spanish
Western Michigan University
Kalamazoo, USA
robert.vann@wmich.edu

***Abstract*— This report presents a case study about building a working digital language archive in a hosted university institutional repository. Best practices in language documentation regarding information architecture, organization, and retrieval are considered in relation to university library commitments to resource acquisition/preservation and online cataloging/delivery systems. Despite challenges, findings suggest that constructing digital language archives in university institutional repositories may offer viable collaborative solutions for researchers unable to find suitable, pre-existing archives in which to deposit their language documentary materials. The report concludes that, in such situations, the ability to satisfy best practices may respond to the strengths/weaknesses of particular software implementations as much as it reflects the design team's vision, as theory and method in language documentation increasingly become matters of library and information science.**

Keywords—institutional repository, university library, digital language archive, language documentation theory, best practices, information architecture, information organization, information retrieval, spoken language corpora, Spanish, DARDOSIPCAT

I. Introduction and Statement of Questions

The 21st century has seen an explosion in digital scholarship in the humanities and social sciences. With regard to language sciences, digital recording/dissemination technologies have allowed researchers to largely disentangle language description and language documentation [1], leading to an evolution in our understanding of what constitutes a modern linguistic record. Traditional print outputs are no longer the gold standard [2], and expectations grow annually for language researchers to provide wider electronic access to our data. With regard to library and information sciences, Dublin Core metadata formats and The Open Archives Initiative Protocol for Metadata Harvesting have paved the way for the rise of institutional repositories (IRs) at universities worldwide.

In 2021, it is now highly desirable for language researchers to archive language documentary materials digitally, and digital language archives (DLAs) abound these days. Problems may arise, however, when researchers attempt to find an appropriate DLA in which to deposit their language documentary materials [3]. Such problems may include, among others: languages archived, type of language data archived, type of speech data available, levels of access available, fees for service, audience design, degree of archive user-friendliness, and degree of archive sustainability. When, due to such problems, researchers are unable to find a suitable, pre-existing DLA in which to deposit their language documentary materials, can university IRs (UIRs) offer a viable solution? What are the challenges of building a working DLA within a UIR? In discussing such challenges, this report focuses on best practices in language documentation in relation to DLA information architecture, organization, and retrieval in UIRs.

II. Review of Literature

Over the last two decades, language documentation theory has emerged in various key publications [1, 4-8] that outline guiding principles for the field. These works demonstrate the need for digitally archiving multipurpose records of language in the form of primary data (recordings of spoken language) and apparatus (metadata and transcriptions) to responsibly preserve/disseminate such records for future uses yet unknown. These publications agree that DLAs hold the key to contemporary language documentations. In this regard, Bird and Simons [9] analyzed seven dimensions of data portability for digital language resources, with best practice recommendations that have since become seminal.

Often under the aegis of university libraries, UIRs provide the infrastructure/tools necessary for depositing, preserving, and delivering digital language resources. Consequently, UIRs provide a compelling means of archiving digital language data. Universities worldwide increasingly use repository software, which may be free or licensed and university-integrated or commercially-hosted. The two largest IR software implementations in US universities are dSPACE (120 research institutions/departments) and bepress (419 research institutions/departments) [10]. While dSPACE [11] is an open-

This research was partially funded by Western Michigan University.

source project of LYRASIS, a non-profit organization, bepress provides for-profit, hosted services licensed to universities [12].

Rapidly expanding UIRs now include DLAs. The Virtual Linguistics Lab at Cornell [13] is an example of a dSPACE implementation that, since 2010, has housed a DLA in a UIR. Pilot initiatives like Cornell's are valuable insofar as they provide proof-of-concept that UIR infrastructure suits the information architecture, organization, and retrieval needs of modern DLAs. In this regard, pre-existing commitments to resource acquisition/preservation, online cataloging/delivery systems, and access make UIR-based DLA solutions easier than DLA solutions outside UIRs. Moreover, UIRs also demonstrate strong sustainability potentials for digital resource curation [14]. Unlike "standalone" DLAs whose sustainability may be uncertain due to dependences on grant renewals or sufficient user fees, DLA sustainability in UIRs is relatively dependable. Recent research [15] suggests a trend in sustainable support for UIRs due to their frequent absorption into regular university budgets.

III. Project Design

This section focuses on best practices in language documentation for building working DLAs within UIRs, based on the author's experience designing the UIR-based DLA known as the ***D***igital ***AR***chive to ***DO***cument ***S***panish ***I***n the ***P***aïsos ***CAT***alans, henceforth DARDOSIPCAT. DARDOSIPCAT is dedicated to archiving/disseminating language resources from the Països Catalans. Resources on deposit include audio recordings of interviews about language and society and orthographic transcriptions of these recordings.

Essentially, best practices for building working DLAs within UIRs stem from epistemological responsibilities with regard to the disposition of language data. These responsibilities include documenting data digitally, archiving language documentary materials, and incorporating language documentation theory into information architecture, organization, and retrieval. These responsibilities relate to both the making of DARDOSIPCAT (how it came to be) and the makings of DARDOSIPCAT (the nuts and bolts that hold it together).

A. The Making of DARDOSIPCAT

My early experiences in digital language documentation trace to the Electronic Metadata for Endangered Languages Data [16] language digitization project conferences of the early 2000s, where the seeds of DARDOSIPCAT were planted. These years coincided with the rise of language documentation theory [1, 7, 17]. From these works, it became clear that good documentations should be diverse, large, ongoing, distributed, opportunistic, transparent, preservable, portable, and ethical. Accordingly, mission-critical elements of working DLAs within UIRs would have to focus on creating an architecture for collecting/transcribing/archiving primary data digitally, developing transparent corpus-level and resource-level metadata to create lasting, multipurpose records of observable community linguistic behavior, and promoting maximum accessibility. These best practices would guide the making of DARDOSIPCAT.

A grant awarded in 1998 funded transcribing a corpus of spoken language that would become DARSOSIPCAT's first deposit. Sabbatical leave in 2002-2003 enabled digitizing 20+ hours of audiotaped conversations and editing 500+ pages of transcriptions. Once the UIR at my university, a bepress implementation called ScholarWorks at WMU, became fully operational under university library administration, I collaborated extensively with the UIR's director (a university librarian) and bepress representatives to create architecture within the UIR for the DLA that would become DARDOSIPCAT. This collaboration involved (a) organizing respective DLA series for audio recordings, video recordings, and orthographic transcriptions; (b) strategically mapping Dublin Core elements to customized metadata fields for optimal harvesting/subsequent information retrieval; and (c) designing/testing extensive resource deposit forms leveraging the customized metadata.

DARDOSIPCAT's first recordings were deposited in 2013. Additional resources were added in 2015, 2018, and 2020 as successive grants were obtained in support of various RA collaborations. These collaborations involved, inter alia, digitizing additional recordings, anonymizing recordings with "bleep" tones, creating associated transcriptions, and developing resource pages with catalog metadata and hyperlinks to related resources.

B. The Makings of DARDOSIPCAT

ScholarWorks is a hosted UIR that runs on Digital Commons software. Advantages of being hosted include a proven track record, an "off-the-shelf" product for quick startup, customizable options, strong tech support, and built-in human resources at bepress. Disadvantages include two layers of UIR-related administration and "one size fits all" software limitations that in practice lead to inevitable design compromises.

Currently, DARDOSIPCAT features two (longitudinally-related) corpora on deposit, with resources distributed among three active series: two primary data series (audio recordings and video recordings) and one transcriptions series. Together the collections comprise 253 distinct catalog pages with highly robust metadata and cross-referenced hyperlinks that promote resource discovery/access within the archive by connecting metadata about individual resources to metadata about all related resources throughout the DLA, longitudinally between corpora and by modality within each corpus. Present resources include 113 audio recording resource pages, 18 video recording resource pages, 113 transcription resource pages, and a 9 page contributors galley. Many audio recording resource pages have back-end WAV archival master files (56) and front-end MP3 downloads available (37). Such audio files are encrusted with Dublin Core metadata for harvest. Future work will include uploading more audio files as well as PDF transcriptions and general access resources including usage conditions, user authorization instructions, and explanations of metadata terminology/mapping.

IV. Findings

This section reports my findings regarding the ability of my UIR to accommodate a DLA in compliance with Bird and Simons' [9] best practice recommendations (henceforth BSBPRs) regarding incorporating language documentation theory into information architecture, organization, and retrieval.

BSBPR for FORMATTING involve information retrieval, organization, and architecture issues such as ***openness***, ***markup***, and ***rendering*** respectively. ***Openness*** refers to using language resources without special software. DARDOSIPCAT's access copies are served in MP3 and PDF formats. Such published, proprietary formats are preferred over secret proprietary formats, though nonproprietary formats are considered ideal. ***Markup*** refers to how (meta)data are represented. ScholarWorks's bepress engine supports XML and the Open Archives Initiative Protocol for Metadata Harvesting (V2). ***Rendering*** refers to presenting materials in conventionally formatted displays. DARDOSIPCAT serves MP3 audios that play on common media players and orthographic transcriptions that open in conventional PDF readers.

BSBPR for DISCOVERY involve retrieval issues such as ***existence*** and ***relevance***. ***Existence*** refers to finding resources easily. DARDOSIPCAT resources are full-text indexed in major search engines like Google thanks to metadata cataloging/optimization. ***Relevance*** refers to judging appropriateness of language resources. DARDOSIPCAT's resource pages' extensive catalog metadata facilitate such judgments.

BSBPR for CITATION involve information organization and architecture issues such as ***bibliography***, ***persistence*** of electronic resource identifiers, and ***immutability*** of citable materials. ***Bibliography*** refers to making referential citations of electronic resources. ScholarWorks's bepress implementation lists complete bibliographic information for resources in catalog metadata and provides recommended citations. ***Persistence*** refers to URL breakage. DARDOSIPCAT boasts persistent URLs. ***Immutability*** refers to URL versioning; DARDOSIPCAT's required catalog metadata includes last-revision dates.

BSBPR for PRESERVATION involve information retrieval issues such as ***longevity*** and ***safety***. ***Longevity*** refers to digital resource lifespans/degradation. DARDOSIPCAT stores archival resources in formats likely to endure for generations (WAV and PDF/A). ***Safety*** refers to potential catastrophic loss. DARDOSIPCAT boasts redundant off-site backup copies.

BSBPR for RIGHTS involve information retrieval issues such as ***public benefits***, which refer to user rights to fair use. Enterprise-level login/password safeguards restrict access to DARDOSIPCAT resources. Unfortunately, however, ScholarWorks's bepress implementation provides no automated way to ensure that only bona fide researchers obtain DARDOSIPCAT accounts.

V. Significance

The finding that DARDOSIPCAT has been able to meet BSBPRs as described above is significant in terms of the ability of UIRs with bepress implementations to incorporate language documentation theory into DLA information architecture, organization, and retrieval. This finding suggests that, given the right circumstances, such UIRs can indeed offer viable solutions for researchers to build their own DLAs. The challenges involved in building working DLAs within such UIRs are also significant, however. Developing such DLAs depends on the prior existence of such UIRs and the willingness of UIR directors and bepress employees to collaborate on such projects. Moreover, while the robust metadata handling in bepress UIR implementations is significant for optimal harvesting, indexing, and retrieval of DLA documentary resources, this significance is constrained by the fact that bepress services are merely licensed. One wonders what would become of DLA metadata in a bepress UIR implementation were that UIR to give up its bepress licensure. Similarly, one wonders to what degree the significance (and limitations) of the results presented here may extend to dSPACE UIR implementations as well. Further research is warranted in this regard.

VI. Conclusions

This report has explored a case study of constructing a working DLA inside a UIR to determine the viability of such a solution for researchers unable to find a suitable, pre-existing DLA in which to deposit their language documentary materials. The approaches, methods, and techniques for collection development described above lead to the inevitable conclusion that such an endeavor is a long-term proposition involving multiple, distributed collaborations with university librarians, RAs, and IR software consultants. This conclusion is consistent with prevailing wisdom in language documentation theory [18].

Informed by such theory, construction of DARDOSIPCAT followed BSBPR for language documentation inasmuch as resources allowed. In the end, the ability to satisfy best practices in relation to DLA information architecture, organization, and retrieval in DARDOSIPCAT was conditioned by the strengths/weaknesses of my institution's bepress UIR implementation. The creative vision of the design team, though realized in large part, was not fully achieved.

Since language documentation today depends so heavily on library and information science, one could conclude that the future is bright for DLAs in bepress implementations of UIRs managed by university libraries. In particular, the administration of such DLAs can build on university library strengths in resource acquisition/preservation, online cataloging/delivery systems, and resource retrieval. In this regard, future research regarding archiving language documentary materials in such DLAs may come to see language documentation theory and method as epistemologically indistinguishable from library and information science.

References

[1] N. P. Himmelmann, "Documentary and descriptive linguistics," Linguistics, vol. 36, no. 1, pp. 161-195, 1998, doi:10.1515/ling.1998.36.1.161.

[2] J. Good, "Valuing technology: Finding the linguist's place in a new technological universe," in Language Documentation: Practice and Values, L. A. Grenoble and N. L. Furbee, Eds. Amsterdam/Philadelphia: John Benjamins, 2010, pp. 111-131.

[3] R. E. Vann, "Frustrations of the documentary linguist: The state of the art in digital language archiving and the archive that wasn't," in Proceedings of the 2006 E-MELD Workshop on Digital Language Documentation (Tools and Standards: The State of the Art), Michigan State University, East Lansing, MI, June 20-22, 2006, [Online]. Available: http://emeld.org/workshop/2006/proceedings.html

[4] P. Austin, "Language documentation 20 years on," in Endangered Languages and Languages in Danger: Issues of Documentation, Policy, and Language Rights, L. Filipović and M. Pütz, Eds. Amsterdam: John Benjamins, 2015, pp. 147–170.

[5] L. A. Grenoble and N. L. Furbee, Eds. Language Documentation: Practice and Values. Amsterdam/Philadelphia: John Benjamins, 2010.

[6] N. P. Himmelmann, "Language documentation: What is it and what is it good for?," in Essentials of Language Documentation, J. Gippert, N. P. Himmelmann, and U. Mosel, Eds. Berlin: Mouton de Gruyter, 2006, pp. 1-30.

[7] A. C. Woodbury, "Defining documentary linguistics," in Language Documentation and Description, vol. 1, P. Austin, Ed. London: Hans Rausing Endangered Languges Project, SOAS, 2003, pp. 33-51.

[8] A. C. Woodbury, "Language documentation," in The Cambridge Handbook of Endangered Languages, P. K. Austin and J. Sallabank, Eds. Cambridge, UK: Cambridge University Press, 2011, pp. 159–186.

[9] S. Bird and G. Simons, "Seven dimensions of portability for language documentation and description," Language, vol. 79, no. 3, pp. 557-582, 2003, doi: 10.1353/lan.2003.0149.

[10] Registry of open access repositories, University of Southampton, School of Electronics and Computer Science, July 30, 2021. [Online]. Available: http://roar.eprints.org/information.html

[11] "About DSpace." About DSpace - DSpace. https://duraspace.org/dspace/about/ (accessed July 30, 2021).

[12] "Digital Commons." Digital Commons - bepress. https://bepress.com/ (accessed July 30, 2021).

[13] B. Lust, S. Flynn, M. Blume, E. Westbrooks, and T. Tobin, "Constructing adequate language documentation for multifaceted cross-linguistic data: A case study from the Virtual Center for Study of Language Acquisition," in Language Documentation: Practice and Values, L. A. Grenoble and N. L. Furbee, Eds. Amsterdam/Philadelphia: John Benjamins, 2010, pp. 89-107.

[14] J. McGann, "On creating a usable future," in Profession 2011, R. J. Feal, Ed. New York, NY: Modern Language Association of America, 2011, pp. 182-195.

[15] C. S. Burns, A. Lana, and J. M. Budd, "Institutional repositories: Exploration of costs and value," D-Lib Magazine, vol. 19, no. 1/2, Jan./Feb. 2013, doi: 10.1045/january2013-burns.

[16] "LINGUIST List Projects." E-MELD Homepage. http://emeld.org/ (accessed July 30, 2021).

[17] J. Gippert, N. P. Himmelmann, and U. Mosel, Eds. Essentials of Language Documentation. Berlin: Mouton de Gruyter, 2006.

[18] A. Dwyer, "Models of successful collaboration," in Language Documentation: Practice and Values, L. A. Grenoble and N. L. Furbee, Eds. Amsterdam/Philadelphia: John Benjamins, 2010, pp. 193-212.