A website is a website is a website: Why trusted repositories are needed more than ever

Vera Ferreira
Endangered Languages
Documentation Programme
Berlin-Brandenburgische Akademie
der Wissenschaften
Berlin Germany
vera.ferreira@bbaw.de

Leonore Lukschy
Endangered Languages
Documentation Programme
Berlin-Brandenburgische Akademie
der Wissenschaften
Berlin Germany
lukschy@bbaw.de

Buachut Watyam
Research Institute for Languages and
Cultures of Asia
Mahidol University
Thailand
buachut.bcw@gmail.com

Siripen Ungsitipoonporn
Research Institute for Languages and Cultures of Asia
Mahidol University
Thailand
siripen.ung@mahidol.edu

Mandana Seyfeddinipur
Endangered Languages Documentation Programme
Berlin-Brandenburgische Akademie der Wissenschaften
Berlin Germany
mandana.seyfeddinipur@bbaw.de

ABSTRACT

Over the last two decades there has been a surge in activists, linguists, anthropologists, documenters digitally recording endangered language use. These unique records often are uploaded to corporate social media sites or to privately run websites. Despite popular belief, uploading these materials to a server does not mean they are archived and preserved for future generations. In this paper we discuss the differences between professional archiving systems and content management system (CMS) based approaches to making language materials accessible. Looking at the example of the *Archive of Languages and Cultures of Ethnic Groups of Thailand* we discuss the benefits of a Mukurtu based community website, and how linking it to a professional archive can ensure long-term preservation of precious and unique language materials.

CCS CONCEPTS

•Information systems ~ Information systems applications ~ Digital libraries and archives •Information systems~Information storage systems~Storage management~Information lifecycle management •Information systems~World Wide Web~Web interfaces •Information systems~Information retrieval~Users and interactive retrieval~Search interfaces

KEYWORDS

Digital archiving, Community archives, CMS, Archiving systems, Data preservation

1 Introduction

Of the 7000-7500 languages spoken today less and less are learned by children who instead learn majority languages. Once children do not learn the language of their heritage, the fate of the language is sealed.

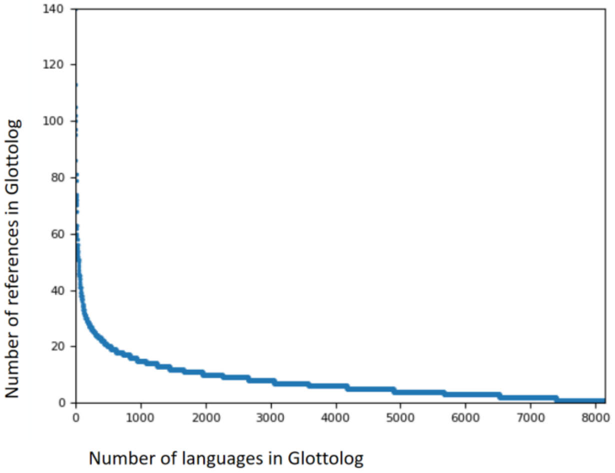

Figure 1: References in Glottolog (data extracted from [3])

Glottolog's bibliographic collection of linguistic descriptive works indicates that for about 35% of the languages there is a full grammar, for 25% there is a sketch grammar, of the remaining 40%, 12% have received some attention, meaning there is a dictionary, a translation of the New Testament or an in-depth discussion of a specific linguistic feature. For 28% there is only a wordlist or similar (see figure 1). Now, while there are some linguistic publications about languages of the world, what about recordings of language use, of people talking, chanting, praying, discussing, negotiating, narrating in situ?

OLAC data aggregator harvests metadata from around 60 language archives and it paints a grim picture with a lot of audio and video recordings and texts for some languages and nothing or very little for the majority of the languages of the world. This tells us something about the current situation of primary language materials available in archives.

Since the digital revolution many people have become active documenting languages and traditions tied to them, making recordings on their phones, on audio recorders and video cameras. If lucky, these materials do not end up on harddrives, laptops or CDs in private possession but the creators aim to make them available on the web to preserve them for posterity. Materials are uploaded to a variety of platforms, in some cases websites created for this specific purpose, in others commercial platforms such as Youtube, Vimeo or Facebook are used to publish recordings. Websites specifically created for the dissemination of language recordings need to be maintained and funded. If the person or group in charge of hosting and maintaining a website no longer has time, the interest, or runs out of funds, the website and the materials on it may be taken offline. Commercial platforms are problematic because it is at the discretion of a private company whether or not the materials stay online. Neither individual websites nor social media platforms have standardised metadata which means that even if these materials are online, they are not necessarily discoverable. And even if they are discoverable, there is no long-term preservation infrastructure. If digital materials are not migrated to newer formats, they will not be accessible in the future, which makes digital files extremely volatile (for more information on issues related to digital preservation in general see [1]; for issues in preserving language documentation data see [2]).

This is worrisome because many of these recordings are invaluable and may be the only recording of a ritual, of an elder, the holder of special knowledge, the shaman, or the singer of songs no one else remembers. Without these materials being professionally archived and preserved long-term, humanity’s intangible heritage is at stake of being lost.

Another issue with individually created platforms is that they rarely rely on long-term funding. This is partly due to the academic funding cycle which is usually only for three years.

Language documentation should result in a multipurpose record, serving speakers or signers of the language documented, linguists and researchers from other disciplines, as well as the general public (see [4] and [5]). These different stakeholders need different ways of accessing materials, which is why websites geared towards specific groups can be very helpful, but it is vital to keep in mind that these websites can only offer a way of showcasing materials, and do not offer actual preservation. The same holds true for social media platforms, which might be valuable for presenting and disseminating materials, but cannot guarantee that these materials will be safeguarded in the long-term.

The fact that recordings are being uploaded to social media sites and privately run websites indicates that there is a clear need to increase the number of local archives to support local efforts in safeguarding documentary records created by a multitude of stakeholders. However, the implementation of sustainable archival infrastructures requires long-term financial and institutional commitment as well as technical expertise. In the meantime, a bottom-up approach whereby local scholars and activists set up a basic content management system and create and collections is one way to secure invaluable existing data, even though it must be clear from the very beginning that a website is not an archive, as reiterated before.

In a discussion on the differences between a website and an archive, and the need to keep the materials archived sustainably to guarantee their safeguarding, in the next sections we will present a bottom-up participatory approach for archive creation which we followed in the project *Archive of Languages and Cultures of Ethnic Groups of Thailand* supported by the Newton Fund. The major goal of the project, which was carried out in a collaboration between the Endangered Languages Archive (ELAR) and researchers from the Research Institute for Languages and Cultures of Asia at Mahidol University, was the implementation of a pilot small-case digital infrastructure for preservation and dissemination of indigenous linguistic materials and cultural heritage in Thailand.

2 CMS vs Archiving

Content management systems (CMS) available at the majority of web hosting servers have made the creation of websites available to a wide variety of users with different levels of technical know-how and are therefore well suited for crowd-sourcing materials collected by a number of individuals. However, using a CMS for adding recordings of a language to a website is not to be confused with archiving and preserving these materials. While a digital archive has both a preservation layer entailing the data conservation and maintenance workflows (like automated format migration, integrity checks, version control, etc.) illustrated in figure 2, as well as a presentation layer for displaying the data, a CMS lacks the preservation layer, focusing solely on displaying materials. A preservation layer is necessary as digital formats change rapidly, and it is key to migrate archived materials to the most up-to-date formats to guarantee their accessibility. In a professional archiving system this kind of migration can be, and normally is, automated. In a CMS, the migration of formats and their conversion needs to be done manually, which is error prone and time intensive.

Figure 2 illustrates the workflow connected to an archiving system, whereas Figure 3 highlights the components that are missing in standard CMS systems (or websites in general and social media platforms).

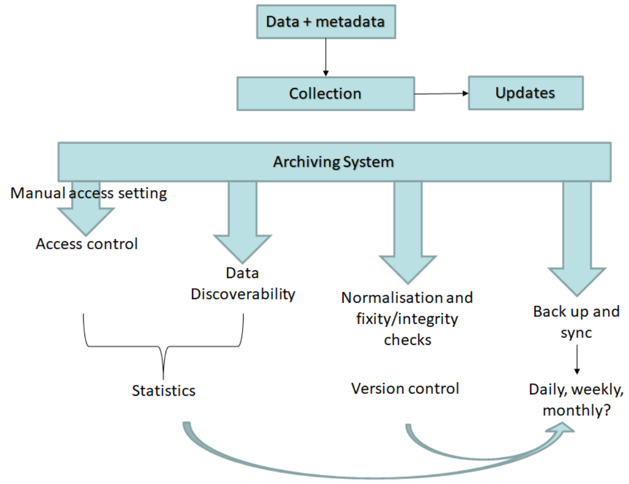

Figure 2: Workflow of an archiving system

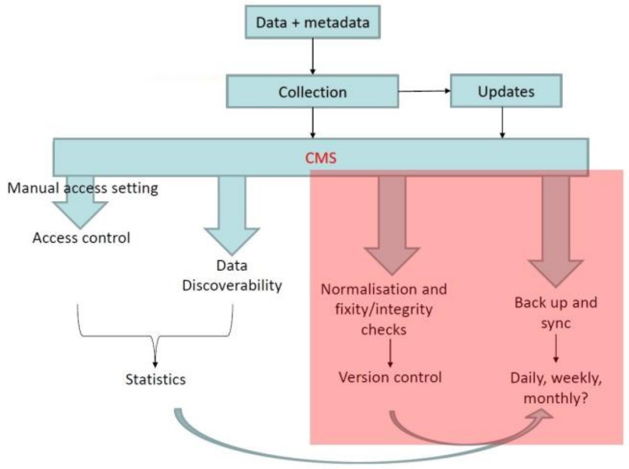

Figure 3: CMS vs archiving systems

The technical infrastructure and long-term funding necessary for archiving represent obstacles for the creation of local archives following archiving standards and best practices.

There are however intermediate solutions which combine less technical expertise and low costs with the basics of archiving, namely Mukurtu[1]. Mukurtu was developed out of the need for an easy to use out of the box system for communities to build up their own archives under their own leadership, maintaining data sovereignty. Mukurtu (meaning *dilly bag* or a safe keeping place for sacred materials in Warumungu language; see [6] and [7]) is a community-oriented CMS infrastructure based on Drupal (an open-source web content management) developed and maintained by the Center for Digital Scholarship and Curation at Washington State University. Mukurtu is a grassroots project aiming to empower local communities to manage, share, and exchange their digital heritage in culturally relevant and ethically-minded ways. It follows archiving standards by supporting and enforcing standard metadata schemas and formats; it has different levels of access, respecting data sensitivity and community wishes, in a user-friendly interface, ensuring CARE[2] and FAIR[3] data principles. It is easily customisable and localisable, allowing multilingual data presentation. Even though Mukurtu is still a CMS system without a preservation layer, it was developed based on archiving core principles and introduces its users to the basics of digital archiving.

3 Bottom-up participatory approach to archive creation

In this section we will present the project *Archive of Languages and Cultures of Ethnic Groups of Thailand*[4] as an example of an intermediate solution for digital archive creation which is based on a bottom-up participatory approach (see also [8]).

The *Archive of Languages and Cultures of Ethnic Groups of Thailand* came to fruition through a collaboration between ELAR and the Research Institute for Languages and Cultures of Asia (Mahidol University). The project was supported by the Newton Fund, with the aim to create a digital platform for the preservation and dissemination of indigenous linguistic materials and cultural heritage in Thailand. The richness of publicly unknown data collected over the years in Thailand, the activism that characterises the attitude of several language community members and scholars in the country, associated with the lack of a digital archive for language materials, led us to develop a community-oriented approach to archiving and to select Mukurtu as the digital platform. The major reason behind the selection of Mukurtu was the fact that even though Mukurtu is a CMS system, it enforces archiving best practices (like metadata consistency, file format unification, access granularity), and lays the ground for professional archiving. It is fully customisable (also in terms of language interface - the Mukurtu instance in this project was fully localised to Thai), simple to use and less academia-oriented. The resources (audio, video, pictures, texts), the languages and the speaker communities are in the foreground – which is an important feature to catch the attention of a broader audience and thus increase the usability of the archived materials.

In this particular case, Mukurtu was combined with a working and backup server, to guarantee the preservation of original primary data and the necessary format migrations.

After the digitisation of legacy materials from 15 different languages in Thailand (comprising audio, video, text, pictures), the materials were sorted according to their language and for each language and/or ethnic group a collection was created in Mukurtu. The materials that belonged together (for instance audio recordings and corresponding transcriptions) were organised in bundles and corresponding metadata was created. The metadata which followed a clearly defined structure, together with the resources, were loaded

[1] https://mukurtu.org (accessed on July 30, 2021)
[2] **C**ollective benefit, **A**uthority to control, **R**esponsibility and **E**thics
[3] **F**indable, **A**ccessible, **I**nteroperable and **R**eusable
[4] http://langarchive-th.org (accessed on July 30, 2021)

to the corresponding collections in Mukurtu and were made available for search and visualisation through the Mukurtu discovery layer. The data sets were also expanded with materials provided by researchers / community members not directly involved with the project. They were trained on data management and archiving mainly in the archiving workshops organised throughout the project. To facilitate the interaction with the archive, helpsheets on data curation and loading were created and made available through the archive website.

It is the only digital platform in Thailand which entails primary data for different ethnic groups and their languages in a consistent and methodological way. It is the first fully localised Mukurtu instance. It includes 15 collections on 15 different ethnic groups in Thailand (Hakka, Tak Bai, Gong, Pattani Malay, Chung, Saek, Chong, Urak Lawoc, Northern Khmer, Kasong, Nyah Kur, Kuy, Moken, Akha and Bisu) with more than 110 digital heritage items (5 hours of video, 7 hours of audio, 90 text files and around 140 pictures) and detailed metadata in Thai.

Processing the legacy materials and making them available digitally following best practices on data processing and metadata creation has a huge impact not only at socio-cultural level by contributing to the promotion and preservation of language diversity in Thailand but also at academic level by fostering research both on language documentation, linguistics in general and pedagogy (several teaching materials can now be created based on the data that was made available through the archive). Moreover, training community members on data curation and archiving, so that they can expand the database created within this project was key for the necessary empowerment that allows community members to have control over their language and culture and to take part in decision making processes.

However, this is only the first step towards sustainable digital archiving. As mentioned before, while Mukurtu enforces basic archiving workflows, it is merely a CMS rendering a presentation layer. Throughout the project, the users inputting data into the Mukurtu platform became aware of the importance of rich and standardised metadata, format consistency and format migration, i.e. they became aware of the core digital archiving principles and how they differ from a simple website creation. Due to the clear workflows and basic archiving principles implemented throughout the project, the shift to or the combination of Mukurtu with a professional archiving system with an automated preservation layer will be much easier in the future.

4 Conclusion

Platforms such Mukurtu offer an opportunity to break with the tradition of an extractivist North-South relationship, where data is kept securely in western academic institutions, while the rich materials compiled by language community members and activists in the Global South are not preserved and made accessible locally. Having a platform which can be easily localised, as is the case with Mukurtu, is already an essential step to make the materials discoverable by and accessible to their own authors and creators, strengthening the relationship between archives and their users – a tendency we could observe during the Thai project.

In terms of community archiving, the ideal scenario would be the combination of the functionalities offered by Mukurtu with an automated preservation layer or with the additional storage of the materials in a professional archive that guarantees their preservation and accessibility over time. The same applies to websites dedicated to individual languages or larger scale projects. While all of these efforts are important for making materials more easily accessible to communities and the general public and can be very valuable for crowd-sourced collection of materials, they need to be linked to or integrated in a professional archive to ensure that the data is preserved long-term.

REFERENCES

[1] Digital Preservation Handbook, 2nd Edition, https://www.dpconline.org/handbook, Digital Preservation Coalition © 2015.

[2] Bird, Steven and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. LANGUAGE, 79 (3), 557-582. https://doi.org/10.1353/lan.2003.0149.

[3] Hammarström, Harald, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2021. ‘Glottolog Database 4.4’. Max Planck Institute for Evolutionary Anthropology. https://doi.org/10.5281/ZENODO.4761960.

[4] Himmelmann, Nikolaus. 1998. Documentary and descriptive linguistics. Linguistics 36:161-95.

[5] Himmelmann, Nikolaus. 2006. Language documentation: What is it and what is it good for? In Jost Gippert, Nikolaus Himmelmann and Ulrike Mosel (eds.), *Essentials of Language Documentation*. 1-30. Berlin: Mouton de Gruyter.

[6] Christen, Kimberly, Alex Merrill, and Michael Wynne. 2017. ‘A Community of Relations: Mukurtu Hubs and Spokes’. D-Lib Magazine 23 (5/6). https://doi.org/10.1045/may2017-christen.

[7] Christen, Kimberly. 2015. Tribal Archives, Traditional Knowledge, and Local Contexts: Why the “s” Matters. Journal of Western Archives*:* Vol. 6, Iss. 1, Article 3. DOI: https://doi.org/10.26077/78d5-47cf.

[8] Ungsitipoonporn, Siripen, Buachut Watyam, Vera Ferreira, and Mandana Seyfeddinipur. 2021. Community Archiving of Ethnic Groups in Thailand. Language Documentation & Conservation 15, 267-284. Handle: http://hdl.handle.net/10125/24975.