11

A Formosan Multimedia Dictionary Designed Via a Participatory Process

Meng-Chien Yang, Hsin-Ta Chou, Huey-Shiuan Guo, Gia-Pyng Chen

Providence University

Digital archiving is important work for an endangered language, because if an endangered language disappears, associated cultural assets will disappear altogether. Several digital archiving projects are being conducted in Taiwan. Many tribal teachers are now involved in these projects. Based on the needs of these tribal teachers, this paper presents an easy-to-use system for digitally archiving Formosan Languages. The proposed approach takes advantage of the Internet and the newly launched Web 2.0 sharing platform. This chapter gives details of the development and structure of the online dictionary system. Currently, several archiving projects in Taiwan are using this system to teach tribal teachers how to develop their own language resources and online dictionaries.

1. INTRODUCTION[1]. Developing dictionaries for endangered languages is a long and complex process. Although it is easy to collect large archival databases of endangered languages, the purpose and how to best use these archives is sometimes unclear. The most frequently asked question is how the documentation benefits native speakers of the language (Eisenlohr 2004). This question can be addressed simply by creating a shared language resource from these archives. However, implementing this shared language system is a very complicated and difficult task. The Internet is probably the best vehicle for developing shareable language resources. The well-established Lexique Pro, developed by the Summer Institute of Linguistics (SIL), can transform a digital archive into a dynamic dictionary with hyperlinks that can be published on websites. Lexique Pro is a useful tool for field linguists developing shareable language resources. Conversely, mastering Lexique Pro requires both the skills associated with field linguistics and computer technology. Native speakers have recently become increasingly devoted to saving their own languages. Although indigenous peoples often possess considerable knowledge of their own languages, very few have the necessary computer and linguistics skills. Hence, their knowledge can be lost if it is not transformed into shareable digital archives. A process that assists indigenous peoples in creating shareable Internet language resources would be valuable. This project was motivated by the enthusiasm and needs of tribal teachers. The work attempts to create a platform for Formosan tribal language teachers to create their own shareable Internet dictionaries.

Research in endangered language documentation was first recognized as a separate field by Himmelmann (1998). Many computer software tools were designed for field linguists to do the documentation work. The tools available for digital archiving were well described

[1] The authors would like to thank Providence University and ELDP, SOAS, University of London for financially supporting this research under contract no. **94_11100_B10**. We also appreciate the editorial services of Dr. Gerald Rau.

in Bird and Simons's work (Bird and Simons 2003a, 2003b). Bird and Huang proposed a platform for language sharing and exchange (Bird, Simons and Huang 2001). Several studies report their collection of endangered language documentation (Lublinskaya and Sherstinova 2002, Psutka, J., et al. 2002, Johnson 2004). The use of the digital archiving for language learning was reported in Csató and Nathan's works (Csató and Nathan 2003, Nathan 2004). In Taiwan, there are several important digital archiving projects (Zeitoun et al. 2003, Zeitoun and Yu 2005). Recently, the research team doing Yami documentation has been developing an e-learning platform and model for Internet e-learning (Rau and Dong 2006, Rau and Yang 2006, Rau, Yang and Dong 2006, Yang and Rau 2005).

This chapter presents a novel web-based system that allows Formosan tribal teachers to create shareable dictionaries of their languages. Formosan tribal teachers are typically the elders, who have little experience using computers. Some tribal teachers are school-teachers or clergy residing in a community. As the government is promoting the use of the Internet and web applications, many tribal teachers may have participated in seminars introducing the Internet and have experience using the Internet. This chapter presents a simple and useful web-based system that allows tribal teachers to create their own web-publishable dictionaries.

This project capitalizes on the openness and freedom of the Internet to design an environment in which the tribal teachers can create and share their languages. The system is based on a participatory process associated with new generation web applications, such as Web 2.0 (Treese 2006). The design combines field experience and IT technologies to create an online environment for developing Formosan dictionaries. This collection of dictionaries can be utilized as a resource for creating CALL systems for teaching Formosan languages (Fujii et. al. 2000, Ward and van Genabith 2003).

The remainder of this chapter is organized as follows. Section 2 includes two case studies from the Yami documentation project. These two cases illustrate some difficulties the indigenous tribal teachers faced in building their own language documentation projects. Section 3 describes the design of the language resource use and editing environment to develop a Formosan dictionary system with user-friendly interface and simplified annotation tools. Section 4 gives a detailed description of this system. Section 5 presents the conclusions.

2. INITIATIVES AND BACKGROUND OF THE PROJECT. This project is motivated by the experience and results of a project sponsored by the Endangered Language Documentation Program (ELDP) to document the Yami language (see Rau and Yang in this volume for further details). A language revitalization seminar was held on Orchid Island in 2006. Several tribal teachers were recruited to collect the Yami corpus. To help these teachers document the collected Yami corpus, a seminar was held on how to use Toolbox, a software package for language documentation developed by SIL. The teachers showed great enthusiasm for documenting the Yami language; however the teachers had difficulties using Toolbox. The next section describes the Toolbox training seminar and the difficulties encountered.

The Yami consultants on the research team contributed many corpora in various formats such as audio tapes and video tapes. To document these corpora, the research team spent considerable time digitizing and annotating these analog data. These experiences are described in Section 2.2.

2.1 THE TRAINING SEMINAR IN USING TOOLBOX ON ORCHID ISLAND. A Toolbox training seminar was held on April 1 and 2, 2006 at Lanyu High School, Orchid Island. The seminar topics were lexical annotation, glossary creation, word and phrase compilation using Toolbox and producing a digital dictionary. Eight tribal teachers participated in this seminar. Some teachers were teaching the Yami language at local schools. Therefore, the aim of the workshop was to teach the Yami tribal teachers how to construct dictionaries. A sample dictionary previously designed by the research team was provided to participants.

Toolbox operations were introduced at the beginning of the class. Research team members created a set of snapshot steps so that these Yami tribal teachers could quickly create a simple version of the Yami dictionary (Figure 1). However, the research team found that the class members were confused by the complex settings in Toolbox and its English interface. Class progress was far behind schedule at the end of the seminar.

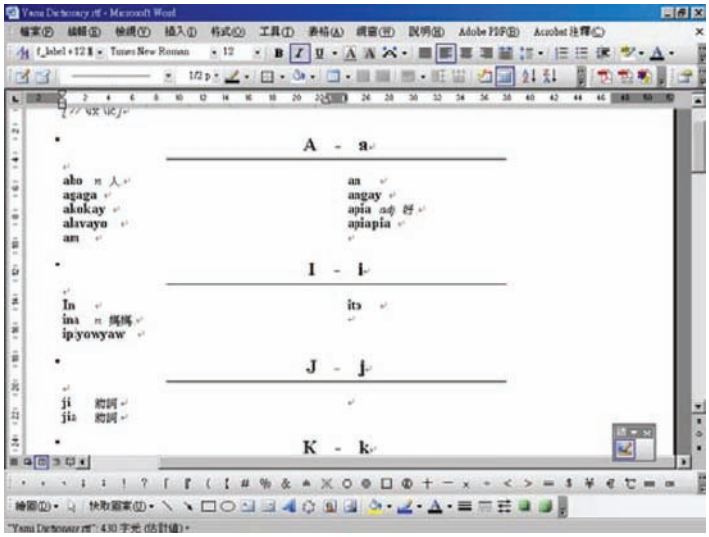

FIGURE 1. Toolbox export dictionary file

To prepare for the introductory seminar, the project team created a Chinese manual with the following content:

1. How to download Toolbox software
2. Basic operations of Toolbox
3. How to build the corpus using Toolbox

Tribal teachers had three major difficulties using Toolbox after the seminar. First, all processes, explanations and interfaces are in English, which clearly presented difficulties for these teachers. Second, Toolbox functions are numerous and diverse. Toolbox users need to define many items when a corpus is created. Finally, the tags and marks of Toolbox are user-oriented and cannot be shared with other users.

Similarly, problems were encountered when instructing tribal teachers how to use Shoebox and Lexique Pro. Lexique Pro software was specially designed for creating shareable language resources. Although Chinese translations of the manuals of these software programs were created and placed on the project web site, http://yamiproject.cs.pu.edu.tw/yami/yami_ch/link.htm, these difficulties remained. Tribal teachers could not use these manuals to help them develop their own language resources. To assist these tribal teachers in developing their own language corpora, problems were analyzed to find a solution based on the local culture and the abilities of the tribal teachers.

2.2 CONTRIBUTIONS FROM THE TRIBAL TEACHERS . The Yami documentation project invited the local Yami consultants to contribute their own language resources. One of our consultants had 101 audiotapes, recorded over three decades. These audiotapes contain many recordings of folk songs, ceremonies, special activities, and teaching from elders. However digitizing these tapes without losing contents was a challenging task, as the quality of the tapes had deteriorated. Some tapes had bad tracks that could not be digitized. Therefore, one staff member manually transformed each track of these tapes into digital data. Transforming legacy language resources into new digital data is a very common practice. We speculate that many such audiotapes exist. These audiotapes must be preserved, organized and transformed into digitally archived data.

3. PROJECT RATIONALE . This project explored possible solutions to, and techniques available for creating a shareable language resource for the Formosan languages. The advantage of the Internet was the main consideration in the design of the shared resources. In addition, this project adopted the design of the Web 2.0 platform (Millard and Ross 2006, Treese 2006). Figure 2 shows the format for the participatory process. The objective was to create a web-based online Formosan dictionary editing system. The system was designed as a shareable and easy-to-use platform.

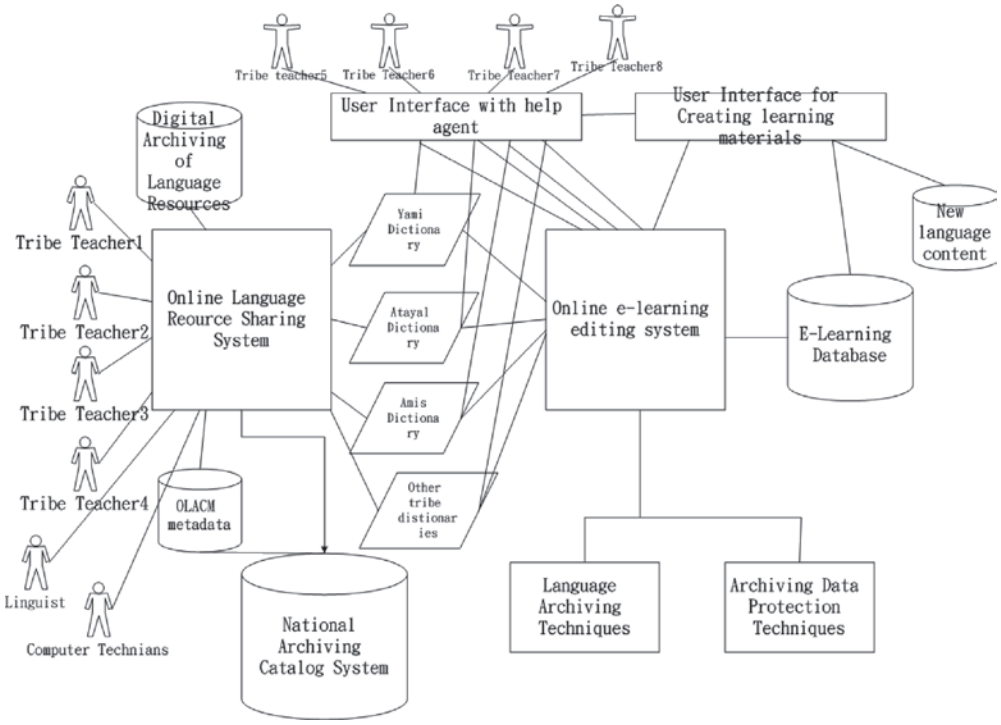

FIGURE 2. Diagram of the system based on the Participatory Process

In this system, the tribal teachers can perform three different tasks:

1. Collect language resources and annotations**:** The online language resource sharing system is designed as the entrance for tribal teachers to collect data. After the collected language resources have the OLACM metadata added, then they can be transformed directly into an online tribal dictionary (Bird, Simons and Huang 2001).

2. Create online learning materials: Tribal teachers can create and design their own e-learning materials using the language resources collected.

3. Share and searching the proper language resources: The tribal teachers can electronically search and share their language resources.

We hope the proposed system can produce an Internet environment in which all participating teachers can post their language collections and can produce their own language learning materials. This on-line environment would establish a virtual community among the tribal teachers. Based on Web 2.0, this environment should foster sharing and collaborative activities among the tribal teachers of the same Formosan language. Moreover the

system can function as an intelligent diary (de Silva et al. 2007). This on-line environment is being developed using coarse-to-fine and bottom-up strategies. Development is currently at a very early stage and focuses on the following two components.

Interface Design for the tribal language teachers and elders: Most tribal language teachers had not previously taken any computer-related training courses and tend to avoid using computers. Therefore, the application environment must be simple, easy to use and facilitate 'safe exploration' for tribal teachers. This project is not a trivial or simple project and must undergo several cycles of evaluation and refinement.

A localized and simplified version of Toolbox for creating the language resources: A simplified version of language documentation tools must be created to record the basic and important lexical items and the collected language words and phrases. This software can hopefully be extended to enable data exchange with databases created in Toolbox.

This online Formosan dictionary editing system, called "Taiwan Austronesian Language Digital Archiving System at Providence University" (TALDAS-PU) can be found at http://dicts.cs.pu.edu.tw/ada/. This system allows tribal teachers to enter their new words and eventually produce dictionary-style output.

4. SYSTEM DEVELOPMENT. A digital archiving system (TALDAS-PU) is currently being developed for the tribal teachers based on the design described in Section 3. Figure 3 shows the architecture of the system. This system includes a web server for developing server site programs and a database management tool for creating the tables for the digital archiving system.

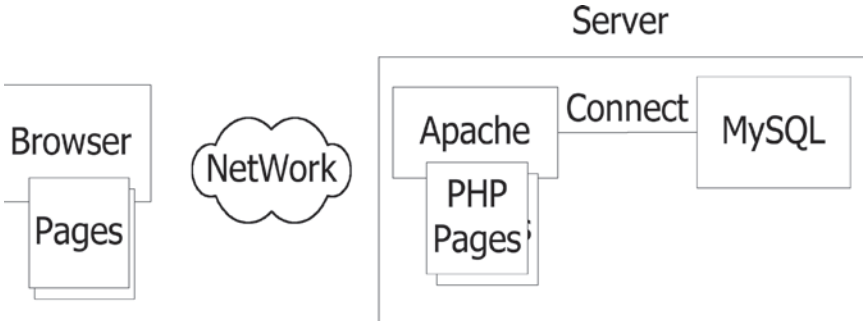

FIGURE 3. System Architecture

The whole system is described as a web-based application. Figure 4 shows the structure and modules of site web pages.

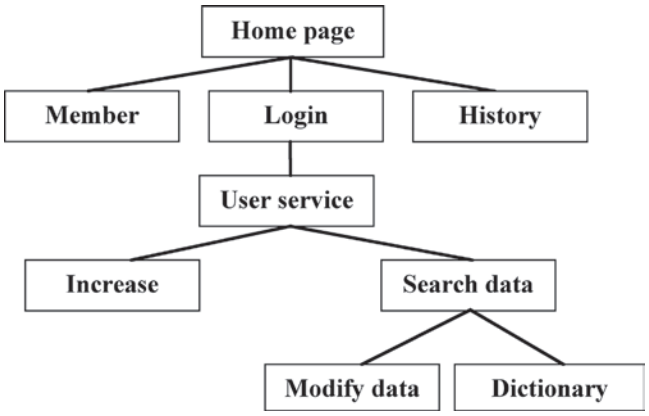

FIGURE 4. Structure of Dictionary System

Users must register when they first log in to the system. Users can then enter service modules and create their own language resources by adding new words to the database, searching the existing word entries or modifying the word entries. Additionally, the system records the action history of the users in the history module. The history data can then be analyzed to give the users proper guidance and assistance.

In each module (Figure 4), several functional sub-modules are created under the block modules:

1.	User registration	This sub-module allows users to register their personal information including name, email address, account, password, language, tribe, job and language used.
2.	Login interface:	This sub-module consists of a user interface for entering the system.
3.	Adding a new entry	This module allows users to add a new word entry.
4.	Data search module	This module provides search functions for the dictionary.
5.	Modifying existing entries	This module allows users to modify existing language data.
6.	Dictionary output module	This module allows users to output the language dictionary.
7.	History module	This module records history data when members add or modify data.

The database for TALDAS-PU consists of five relational tables: User, Language, Vocabulary, Decex and History. Figure 5 shows the data flow diagram and the program modules of the system. From the main page <Index.php>, a registered user of this system can click the [login] button to login into the web site and first-time users to click the [register] button to register as a user.

In <main.php>, users can add a new entry or search the data of the Formosan languages. When a user clicks the [increase-language-data] button, the web links to <input.php> and the user can create a new language entry. If the user clicks the [search-language-data] button, the page <search3.php> opens. If a user selects the search option and clicks the [search] button, then <result2.php> is displayed and the results of the search will be shown on the page. A user can preview the dictionary by clicking the [Dictionary-Preview] button in <result2.php>. The dictionary is shown on the page <re1.php>. If a user wishes to modify the existing data, clicking the [modify-data] hyperlink opens a new page <modify.php> that contains the data that the user wants to modify. After the user modifies the data and clicks the [modify] button, <modifydata.php> is called to check the correctness of the modified data.

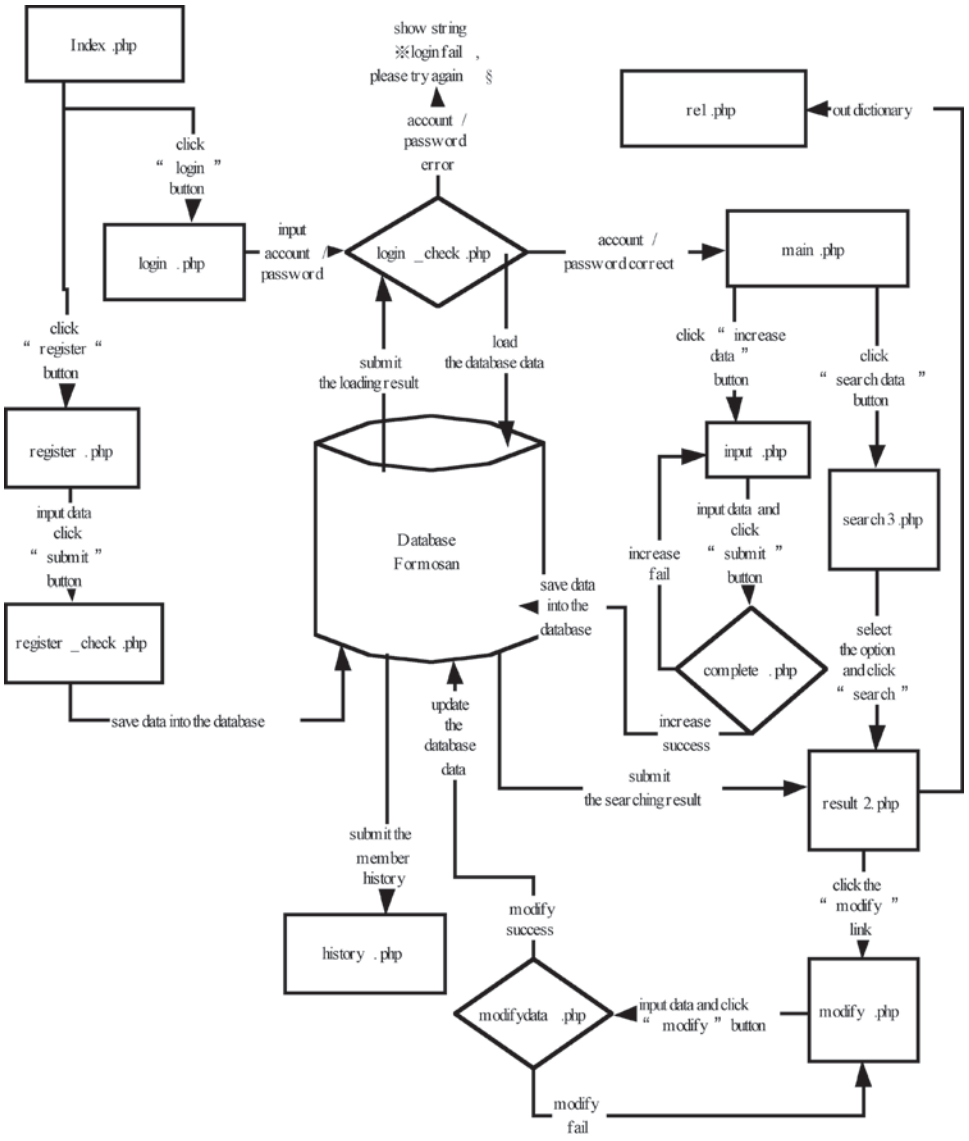

FIGURE 5. System Data Flow Diagram

4.1 INTERFACE DESCRIPTION. This section describes the interfaces and functions of the system. The system web site is located on the digital archiving project web site. Interfaces were developed based on studies of tribal teacher needs and the framework of the shareable online dictionary. The important features of the interface are described as follows.

User Registration Page: When a user clicks the [Register] button, the User Registration page is accessed (Figure 6). This page has the following eight fields. The fields with an * mark must be filled in.

a. Full Name: the user's name

b. E-mail Address: the user's e-mail address

c. Username: an identity for the user, defined by the user

d. Password: the password for entering the system

e. Language/Dialect: the main Taiwan Austronesian Languages are already registered in the system. The system currently holds 40 Formosan languages with sample words and phrases

f. Tribe/Location: the tribe or address of a user

g. Occupation: user occupation, and

h. Language Use: The situation of the language use such as day-to-day use or use only at work.

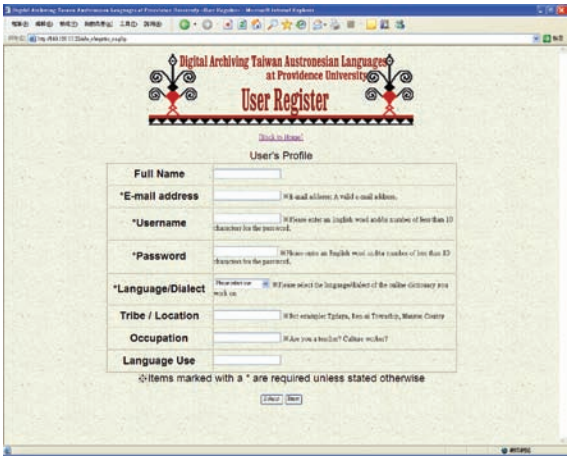

FIGURE 6. The User Registration Page

User Service Page: Once a user logs in successfully, the User Service interface appears as shown (Fig. 7). Via this page, a user can select one of the two system functions, create a new dictionary or search a specific dictionary.

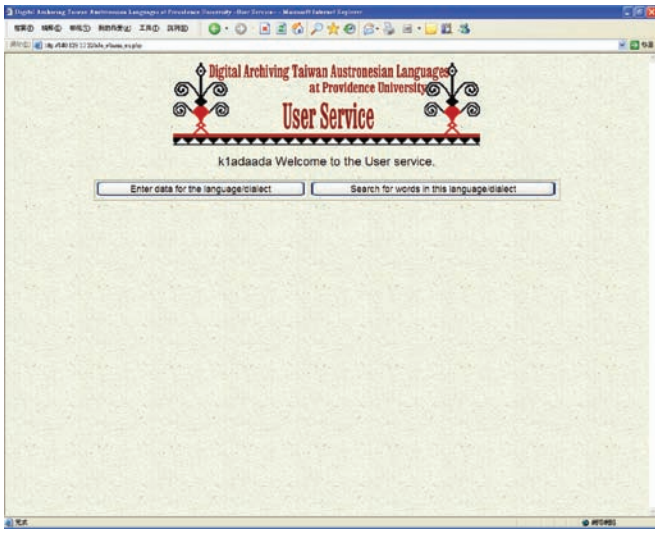

Figure 7. The User Service Page

Data Entry: The data entry page (Fig. 8) allows a user, a tribal teacher, to enter a word or a phrase following the steps shown on the page.

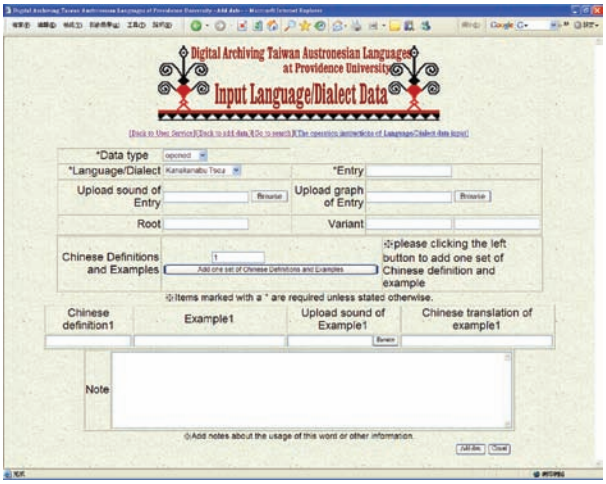

FIGURE 8. The Data Entry Page

The design of these fields is based on studies of tribal teachers' needs. This page quickly collects a large number of words with distinctive annotations. This page has fields to input language/dialect data as follows:

a. Data status: identifies whether the data have been opened. If the field value is "edited", then the data are being edited and can only be searched or modified by the editor. Other users cannot search or modify the data. If the value is "protected", then the editor and other users only can search for the data but cannot modify it. If the value is "open", then the data are open to the public, and all users can search or modify them.

b. Language/Dialect: records the language/dialect of a word a user wants to add

c. Entry: a new entry in a language/dialect

d. Upload Sound of Entry: this functionality allows a user to upload a voice file

e. Upload Graphic of Entry: this field is for uploading an image or graphics concerning an entry

f. Root: records the root form of a word

g. Variant: records word variants

h. Chinese Definitions and Examples: an extensible field for Chinese explanations and sample sentences.

Search for Language/Dialect Data: A user can search any Formosan language/dialect from a dropdown list. Figure 9 shows the search page.

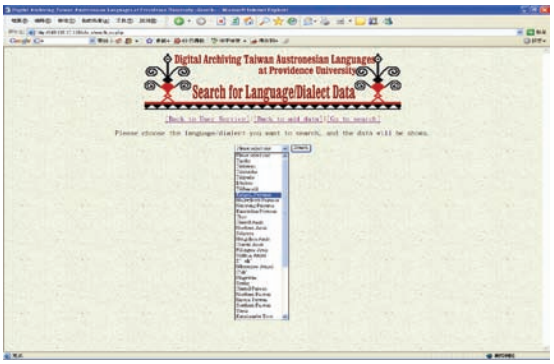

FIGURE 9. The Language Search Page

Figure 10 shows sample search results.

FIGURE 10. Results from a Dictionary Search

Online Dictionary: A user can preview the dictionary for a selected language/dialect by clicking the dictionary-preview button. The online dictionary pops up in a separate window (Fig. 11).

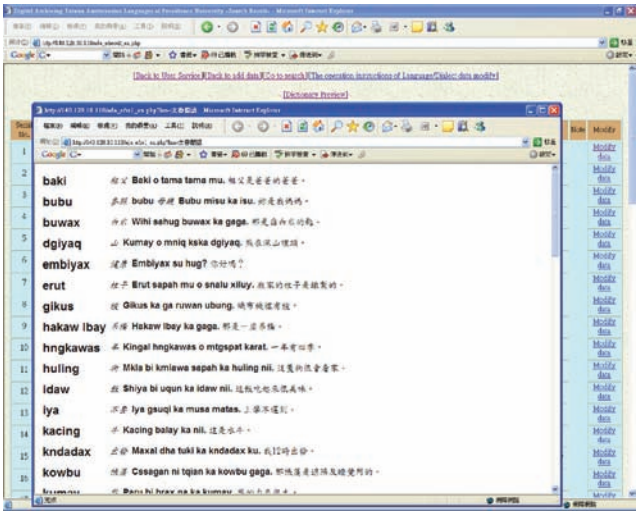

FIGURE 11. Dictionary Preview

Modify Data: To modify a word in an online dictionary, a user clicks the [Modify-Data] button to open a new "modify" page for this word. Figure 12 shows this "modify" page.

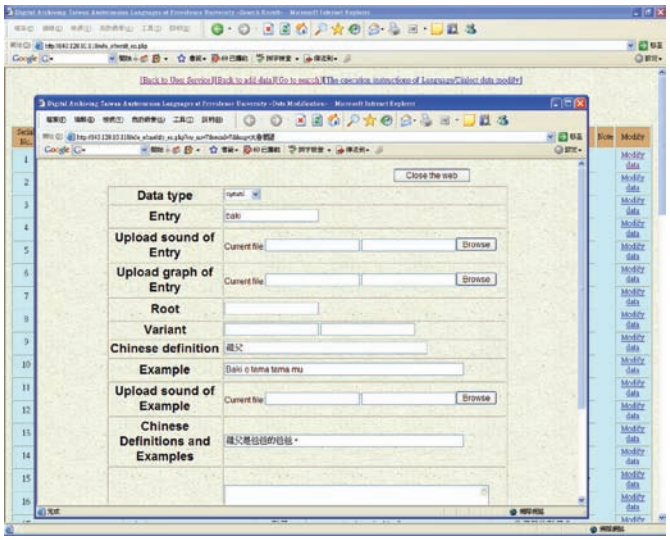

FIGURE 12. The Data Modification Page

Log of Action History: A log function records user activity. If a user adds a new word or modifies a word successfully, the system records this action in the database. Figure 13 shows a user action log. This information is used to analyze system usage.

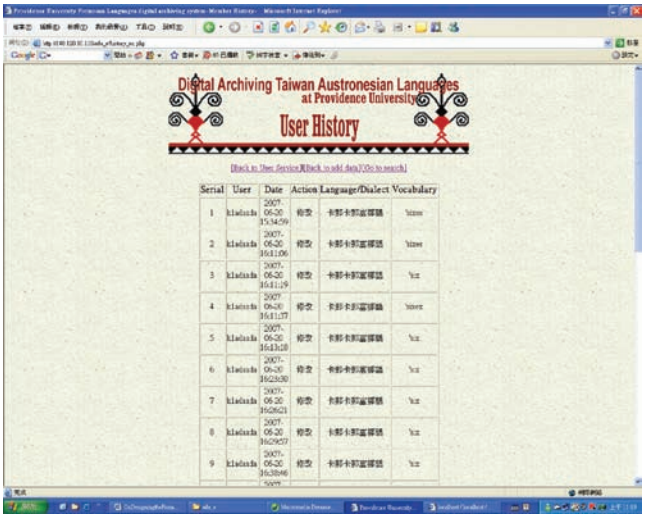

FIGURE 13. The Action Log of User History

5. CONCLUSIONS. This chapter has described the design and implementation of an attempt to create an online Formosan language resource sharing and editing system. The design idea is based on a participatory process for sharing language resources. Development of "Taiwan Austronesian Languages Digital Archiving System at Providence University" (TALDAS-PU) is also reported. The system is a web-based system that can be accessed by any web-based platform and viewed using a browser. Therefore, the system is a feasible platform for online language documentation on a very compact computer, such as an OLPC (http://laptop.org/).

Future work will develop and finish all components. In the next stage, this project will focus on how to use the Web 2.0 platform to create proper tools for sharing the language resources.

References

Bird, Steven, Gary Simons, and Chu-Ren Huang. 2001. The open language archives community and Asian language resources. *NLPRS 2001*: 31-38.

Bird, Steven and Gary Simons. 2003a. Extending Dublin Core Metadata to support description and discovery of language resources. *Computers and Humanities* 37: 378-388.

Bird, Steven and Gary Simons. 2003b. Seven dimensions of portability for language documentation and description. *Language* 93: 557-582.

Csató, Eva and David Nathan. 2003. Multimedia and documentation of endangered languages. In Peter Austin (Ed.) *Language Documentation and Description* 1: 73-84. London: SOAS.

de Silva, Gamhewage C., Toshihiko Yamasaki and Kiyoharu Aizawa. 2007. An interactive multimedia diary for the home, *IEEE Computer* 40(5): 52-59.

Eisenlohr, Patrick. 2004. Language revitalization and new technologies: Cultures of electronic mediation and the refiguring of communities. *Annual Review of Anthropology* 33: 21-45.

Fujii, Satoru, Jun Iwata, Mayumi Hattori, Mutsumi Iijima, and Tadanori Mizuno. 2000. "Web-Call": a language learning support system using Internet. In Proceedings of the Seventh International Conference on Parallel and Distributed systems. 326-331.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36: 161-195.

Johnson, Heidi. 2004. Language documentation and archiving, or how to build a better corpus, In Peter Austin (Ed.) *Language Documentation and Description* 2: 140-154. London: SOAS.

Lublinskaya, Marina and Tatiana Sherstinova. 2002. Audio collections of endangered arctic languages in the Russian Federation. In *Text, Speech, and Dialogue: 5th International Conference Proceedings* ed. by Petr Sojka, Ivan Kopeček and Karel Pala, 347-356. Springer.

Millard, David E. and Martin Ross. 2006. Web 2.0: Hypertext by any other name? In Proceedings of the seventeenth conference on Hypertext and hypermedia: 27-30. New York:ACM.

Nathan, David. 2004. Developing multimedia documentation. In Peter Austin (Ed.) *Language Documentation and Description*. 2: 154-168. London: SOAS.

Psutka, Josef, Pavel Irching, Jesef V. Psutka, Vlasta Radová, William J. Byrne, Jan Hajic, Samuel Gustman and Bhuvana Ramabhadran. 2002. Automatic transcription of Czech language oral history in the MALACH project: Resources and initial experiments, In *Text, Speech, and Dialogue: 5th International Conference Proceedings* ed. by Petr Sojka, Ivan Kopeček and Karel Pala, 253-260. Springer.

Rau, D. Victoria and Meng-Chien Yang. 2006. Digital transmission of language and culture rethinking pedagogical models for e-learning, Paper presented at the Joint AAAL and ACLA/CAAL Conference. Montreal, Canada. 6/17-20/2006.

Rau, D. Victoria, Meng-Chien Yang and Maa-Neu Dong. 2006. Endangered language documentation and transmission. Paper presented at the 10th International Conference on Austronesian Linguistics. Palawan, the Philippines. 1/17-20/2006.

Rau, D. Victoria and Maa-Neu Dong. 2006. *Yami texts with reference grammar and dictionary*, Language and Linguistics. Special Monograph A-10. Taipei: Institute of Linguistics, Academia Sinica.

Rau, D. Victoria. 1995. Yami Vitality. NSC report (NSC84-2411-H-126-001), presented at the Symposium on Language Use and Ethnic Identity, Institute of Ethnology, Academia Sinica 5/16/1995.

Treese, Win. 2006. Web 2.0: Is it really different? *netWorker*. 10(2): 15-17.

Ward, Monica and Josef van Genabith. 2003. CALL for endangered languages: Challenges and rewards. *Computer Assisted Language Learning* 16(2-3): 233-258.

Woychowsky, Edmond. 2006. *AJAX: Creating web pages with asynchronous JavaScript and XML*. Pearson Education, Inc.

Yang, Meng-Chien and D. Victoria Rau. 2005. An integrated framework for archiving, processing, and developing learning materials for an endangered Aboriginal language in Taiwan. In *Proceedings of the 5th Workshop on Asian Language Resources*, 32-39.

Zeitoun, Elizabeth, Ching-Hua Yu, and Cui-Xia Weng. 2003. The Formosan language archive: Development of a multimedia tool to salvage the languages and oral traditions of the indigenous tribes of Taiwan, *Oceanic Linguistics*, 42(1):218-232.

Zeitoun, Elizabeth and Ching-Hua Yu. 2005. Language analysis and language processing, *Computational Linguistics and Chinese Language Processing*, 10(2): 167-200.

Meng-Chien Yang
mcyang2@pu.edu.tw

Hsin-Ta Chou
ada@hinet.net.tw

Huey-Shiuan Guo
hsguo@pu.edu.tw

Gia-Pyng Chen
png@pu.edu.tw90