E-MELD AND THE SCHOOL OF BEST PRACTICES: AN ONGOING COMMUNITY EFFORT

Jessica Boynton
Eastern Michigan University

Steven Moran
University of Washington

Anthony Aristar
Helen Aristar-Dry
Eastern Michigan University

The E-MELD Project

Many languages are being lost as smaller populations disappear in the face of the encroaching mega-cultures of our time, and numerous documentation projects have recently been initiated in an attempt to preserve as much linguistic and cultural information as possible. This response to the threat of language attrition can only be applauded, but it has also drawn attention to the need for more information about optimal formats for digital documentation. Irreplaceable language documentation is often being stored in digital formats vulnerable to hardware and software obsolescence. Moreover, the heterogeneity of formats currently in use limits the accessibility and repurposing of the data.

In order to address these problems, it is necessary for linguists, archivists, and language engineers to reach consensus about best practices in digital language documentation and for these recommendations to be promulgated within the linguistics community. The E-MELD Project (Electronic Metastructure for Endangered Languages Data),[1] organised by The LINGUIST List and funded by The National Science Foundation of the USA (NSF),[2] was conceived partly to promote such a consensus. Among its objectives were:

- The formulation and promulgation of best practice in linguistic mark-up and metadata creation;

- The development of a 'showroom of best practices'—that is, a demonstration site with data digitised according to recommended practices;
- Organised communication within the research community about developing standards and recommendations.

One of the ways these goals were pursued was through the development of the E-MELD School of Best Practices[3] (henceforth 'School'), a website built to publicise digital standards specific to linguistic material. Although the School was only a small part of the five-year E-MELD project, its coverage and potential usefulness have made it one of the most visible; hence it is the focus of the following paper.

The School showcases the recommendations formulated by experts in linguistics, archiving and language technology during the course of five annual E-MELD workshops. The first two workshops were dedicated to reaching a consensus about digital best practices for lexicons and texts. The third focused on the presentation of these recommendations in the School. The fourth was devoted to community review of the General Ontology for Linguistic Description (GOLD), a proposed standard for promoting interoperability of linguistic annotation. The fifth reviewed digital tools supporting best practices, again emphasizing the need for interoperability. Regular participants at these workshops included representatives from initiatives as OLAC,[4] DELAMAN,[5] DOBES,[6] HRELP,[7] AILLA,[8] and PARADISEC.[9] And the input of these organisations has been vital to the development of the standards, models, and suggestions that comprise the content of the School.

The Evolution of the School of Best Practices

The School began as 'the showroom of best practice', originally intended to demonstrate the results of following best practices by showcasing data from ten endangered languages. The idea of the Showroom was to present texts and lexicons from 10 typologically diverse endangered languages, offer software tools, and list resources that explained best-practice standards. Finally, the School was meant to host a Query Room, where users could ask language questions of native speakers.

These goals have changed as the needs of the user community have become clearer. At this point, near the end of the E-MELD project, the School of Best Practices has evolved into an informational resource that describes digital best practices and offers practical guidance in

implementing these. Its intended audience includes, not only linguists, but also speech community members and archivists.

The need for an instructional resource

The School was originally designed primarily to showcase data. However, as dialogue with the community progressed, it became clear that the standards and processes reflected in the data needed to be better explained and made more accessible to the users of the site, so that the information could be effectively applied.

When the School site first appeared in 2003, it contained links to tutorials about digital practices, a list of related reading materials, descriptions of software useful to linguists, areas that invited users to create metadata and mark-up, and an area for display of language data. This content was presented as a series of *rooms*: the Classroom,[10] the Reading Room,[11] the Tool Room,[12] the Work Room,[13] and the Exhibit Hall.[14] This incarnation provided an infrastructure for the School but did not provide enough guidance in implementing the technologies it recommended or an adequate rationale for the recommendations.

At the suggestion of the participants in the annual E-MELD workshops, the site was first expanded in 2004 to include a general explanation of best practices, called the Entrance Hall.[15] Some digitisation technologies, such as audio and video data conversion,[16] metadata[17] and XSL style sheets[18] in the Classroom, were also added. In addition, a Case Studies[19] section was added to describe the digitisation process for the languages featured on the site. This version of the site provided additional guidance to the users, but still needed more usable information on the topics presented. Also, while it now gave an explanation for best practices, it still did not explain language and data endangerment.

In 2005, the community requested that more information about endangered languages and endangered data be added to the Entrance Hall in order to explain the problem the E-MELD project addresses. Extensive instructional content was also added to the Classroom, which was split into four sections: media types, documentation types, technologies, and preservation of linguistic resources. This provided clearer organisation for the growing content of the site. However, users noted that although the site explained the recommendations of best practices well, it failed to provide implementation advice that took into

account the realities faced by most projects, such as limited time and funding.

For example, the School appeared to suggest that only the very best practices had any utility. It did not indicate the value of good (versus best) practice. For example, the Case Study on Biao Min[20] presented a digitisation effort that transferred two shoeboxes of note-cards to digital format. While the Case Study focused on improving the durability of the language data by digitising it, it needed to highlight the distinction between an archival format and a presentation format, and to indicate that there are legitimate uses for the latter. The TIFF image files of the note-cards reflect best practice for archiving; however, uncompressed TIFF files are very large. For presentation, jpeg files are more practical. And, even for archiving, jpegs would have been better than nothing if there was insufficient space available for TIFFS. The School thus needed to emphasise that any data is better than no data, and any practice is better than no practice. There was a real danger in this version of the School that users would feel that since best practice was so hard to attain, it was not worth the effort even to try.

In 2006, the School took on the task of distinguishing good, better, and best practice and of clarifying the value of each. Step-by-step instructions were also added to nearly every topic covered in the Classroom. For example, the Interlinear Glossed Text section[21] of the site explained what an IGT is and why it is useful. It went on to recommend an XML schema for a best-practice IGT, then to give step-by-step instructions for creating one. By continually responding to the feedback given by the users, the School changed from simply describing best practices to making a real attempt to teach them.

Broadening the audience

As the content of the School increased, it became necessary to develop resources that were accessible to a varied audience. While the site might be best known to field linguists, it was early understood that it should also be useful to speech community members who wished to record their heritage language and to archivists who had linguistic data to preserve.

To this end, a glossary of technical terminology[22] was created in 2004 to increase the accessibility of the information to less technical users. In 2005 it was also realised that the site had become too complex. So sections of the Classroom were reorganised to make the general guidelines relevant

to a particular topic appear on a single page, thus obviating the need to spend time moving from page to page in order to track down all the details relevant to a particular process. For those users who wanted more information, the resources listed in the Reading Room were always available.

To help potential users find the sections of the School that were most beneficial for them, navigation guides were created in 2005 for linguists,[23] community members[24] and, in 2006, archivists.[25] Thus archivists, for example, were led to different parts of the School when they needed to answer one of the following questions:

- Why is it important to preserve endangered languages documentation?
- How do linguistic materials differ from other primary sources?
- How important is it to preserve the audio and video recordings?
- Are there special considerations for the digitisation of textual materials?
- Where can I find examples of digitisation projects?
- What kind of metadata is important for linguistic materials?
- Where can I find more information about endangered languages archives?

The intention was that archivists working with linguistic data for the first time could use this guide to find the information they needed.

Providing practical information

The School originally aimed to be a comprehensive guide to the tools and technologies involved in language data digitisation. However, community input made it clear that it is more important to provide practical information about implementation than it is to present a thorough explanation of the technologies involved in the data digitisation process. It quickly became clear, for example, that explaining how to encode metadata using XML was far less useful — and much more intimidating — than describing the elements of the OLAC metadata standard and providing a tool that automatically converts metadata collected via web form into XML.[26] Although the School still provides access to resources that give more technical information for advanced users, the focus of the School has shifted to offering practical guidance.

However, there is one type of practical advice that the School has intentionally steered clear of: specific recommendations about

technologies, which change so fast that any information given rapidly becomes out of date. For example, hardware tools such as cameras are constantly evolving, and it is impossible for the School to offer up-to-date information about each camera model. However, by listing basic criteria[27] to keep in mind when selecting a camera, the School offers general guidance that should remain relevant for a longer period of time.

Review of the School project

The School has produced a resource that is gaining some recognition in the field of documentary linguistics. It has helped the E-MELD project meet some of its primary goals[28] by providing a medium through which the recommendations of best practices can be communicated to an audience of linguists, community members, and language engineers. In doing so, the School has helped raise awareness of the issue of best practices, to the point that 'best practices' is no longer a rare phrase in American documentary linguistics.

The effectiveness of the School shows in that it is listed as a resource on many sites, from that of a corpus class at University of Hawaii[29] to that of the National Science Foundation's Documenting Endangered Languages (DEL) grant initiative.[30] E-MELD is also featured on important digital resources such as the Rosetta[31] and DELAMAN[32] sites.

Representatives of these and related projects, such as OLAC, HRELP, DOBES, AILLA, and PARADISEC, have participated actively in developing and reviewing School content. Indeed, one result of the E-MELD workshops and the collaborative effort to build the School has been the creation of a community of scholars interested in digital best practices and the development of tools that support them.

The School has also been effective in the area of training. A primary goal of the LINGUIST List in all its projects, including E-MELD and the School, is to train students to meet the demands of a future career in linguistics and to continue to promote the values of the project. Since Eastern Michigan University (where LINGUIST List is situated) does not have a PhD program in Linguistics, the School's web development team has consisted of undergraduate students and MA candidates working under the supervision of the E-MELD project's principle investigators. These students have received training in language documentation and language technology, rarely taught at the MA level in America. Many of the Case Studies were developed by students, even when the data was

provided by a more established researcher. And some Case Studies, such as Sisaala[33] and Dena'ina,[34] were written by the students themselves, because they were the ones who conducted the fieldwork. Of the primary student developers of the School, four have conducted original fieldwork on endangered languages, three have gone on to PhD programs in either language documentation or language technology, and one has been awarded a Fulbright grant to study an Aboriginal language. All have presented and/or published work as they have promoted the recommendations of best practices.

The School has also served to preserve and disseminate the data for the languages featured in the Case Studies. For example, the lexicons in the School have been made searchable via the web, and all data has been stored in best-practice, non-proprietary formats.

While the visibility of the School is increasing, the website cannot be called a complete success. The user-maintained databases, including the software registry and the reading room bibliography, have not yet gained wide use, for the community does not yet fully participate in updating them.

The Future of the School

Since technologies are constantly changing, any guide to recommended practices must constantly be updated if it is to remain useful. However, grant funding tends to be focused on research and new initiatives, not on maintenance. Moreover, E-MELD funding ends in June of 2007. As part of its responsibility to the discipline, the LINGUIST List makes a commitment to provide long term hosting and technical maintenance for any infrastructure facility it develops. However, LINGUIST is funded primarily through subscriber donations; and such funding may not provide support for further content expansion. How then is the School to maintain its relevance? A partial solution was built into the School architecture, in that facilities were created to promote content update and maintenance by users.

For example, in 2004, the software registry[35] in the Tool Room was redesigned to allow users to input information about new software applications and comment on those that already appear in the database. Additional revisions and additions are being implemented in response to the recommendations given by the participants in the 2006 E-MELD

workshop,[36] but future additions to the software registry may have to come from users.

The Reading Room is similarly updateable. In 2005, the Reading Room was made into a searchable database of resources. And users can now input resources to make these available to other users on the site.

In the same year, a comments facility was created for all the content pages in the School. This facility enables users to add content to the pages as user-contributed notes. These notes will be published on the site along with the other page content, so that users can learn and benefit from each other's insights. Users are also encouraged to contribute more content or tutorials to the School, or rework existing content, in collaboration with the web development team.

Because spammers and practical jokers contribute a great deal of the material submitted to any user-maintained site, all of these user-input facilities are monitored by editors to protect the integrity of the School The School, in its current format, is a quotable and reliable resource because of this careful checking. And to ensure the long-term reliability of the resource some form of checking must continue. Thus LINGUIST List student editors will continue to monitor user input for the foreseeable future.

In making the School extensible by users, the E-MELD team has tried to produce a language digitisation resource that will serve its user community beyond the limits of the grant period. In responding to the requests of its user community, the School has grown far beyond its original design. It was originally intended merely to showcase the results of best-practice methods, list software tools, and provide a list of resources on topics related to digital language documentation. By responding to the community of users, the School has grown to become, itself, a central resource on the topic, with more than 500 web pages of instructional content. Continued community input will ensure that the E-MELD School of Best Practices functions effectively in the future, both as a teaching tool and a forum for discussion of digital best practices.

Endnotes

[1] E-MELD website: http://emeld.org (E-MELD, n.d.).

[2] The authors would like to thank the National Science Foundation for the grant (SBE-0094934) that funded the work described here. We would also like to recognise the institutions that collaborated with The LINGUIST List on the E-MELD project: the Endangered Languages Fund, the Linguistic Data Consortium, and the University of Arizona.

[3] The School of Best Practices: http://emeld.org/school (E-MELD, n.d.). Note that this link, as well as all others referenced in time-specific contexts within the paper, link to the current (2006) version of the appropriate section of the site. Because the site is continually updated, it would be impossible to provide access to the content as it existed at the time it was first created.

[4] OLAC website: http://www.language-archives.org (OLAC, 2006).

[5] DELAMAN website: http://delaman.org (DELAMAN, n.d.).

[6] DOBES website: http://www.mpi.nl/DOBES (DoBeS Archive, 2006).

[7] HRELP website: http://hrelp.org (HRELP, 2006).

[8] AILLA website: http://ailla.utexas.org (AILLA, n.d.).

[9] PARADISEC website: http://paradisec.org.au (PARADISEC, 2005)

[10] The Classroom: http://emeld.org/school/classroom (E-MELD, n.d.).

[11]The Reading Room: http://emeld.org/school/readingroom (E-MELD, n.d.).

[12] The Tool Room: http://emeld.org/school/toolroom (E-MELD, n.d.).

[13] The Work Room: http://emeld.org/school/workroom (E-MELD, n.d.).

[14] The Exhibit Hall was subsumed into the Case Studies section in 2004: http://emeld.org/school/case/ (E-MELD, n.d.).

[15] The Entrance Hall: http:/emeld.org/school (E-MELD, n.d.).

[16] The conversion section: http://emeld.org/school/classroom/conversion/ (E-MELD, n.d.).

[17] The metadata section: http://emeld.org/school/classroom/metadata (E-MELD, n.d.).

[18] The XSL stylesheets section: http://emeld.org/school/classroom/stylesheet (E-MELD, n.d.).

[19] The Case Studies: http://emeld.org/school/case (E-MELD, n.d.).

[20] The Biao Min Case Study: http://emeld.org/school/case/biao-min (E-MELD, n.d.).

[21] The IGT section: http://emeld.org/school/classroom/text/igt.html (E-MELD, n.d.).

[22] The glossary of technical terminology: http://emeld.org/school/glossary.html (E-MELD, n.d.).

[23] The linguist start page: http://emeld.org/school/lingstart.html (E-MELD, n.d.).

[24] The community start page: http://emeld.org/school/commstart.html (E-MELD, n.d.).

[25] The archivist start page: http://emeld.org/school/archstart.html (E-MELD, n.d.).

[26] The OLAC Repository Editor page: http://linguistlist.org/olac/ore (E-MELD, n.d.).

[27] The hardware section: http://emeld.org/school/toolroom/hardware (E-MELD, n.d.).

[28] The goals of the E-MELD project are to formulate and disseminate a consensus about digital language documentation standards. The School is but one of many E-MELD initiatives, which include five annual workshops; numerous tutorials organised at linguistics meetings, such as the Linguistic Society of America Summer Institute; three internally developed tools; the GOLD ontology; and several language search facilities.

[29] See Keira Ballanty's *Yapese Corpora* website: http://www2.hawaii.edu/~ballanty/corpusintro.html (Ballanty, 2006).

[30] See the National Science Foundation's *Documenting Endangered Languages* website: http://www.nsf.gov/pubs/2006/nsf06577/nsf06577.htm (National Science Foundation, 2005). Other sites include Wikipedia (Wikipedia, 2006), Yale language resources for African studies (Yale University Library, 2006), the Teaching Indigenous Languages' list of organisations and projects (Northern Arizona University, 2006), and Humboldt State University's English Department website (Humboldt State University English Department, 2006).

[31] Rosetta Project links page: http://www.rosettaproject.org/about-us/links (Rosetta Project, n.d.).

[32] DELAMAN links page: http://www.delaman.org/links.html (DELAMAN, n.d.).

[33] Sisaala Case Study: http://emeld.org/school/case/sisaala (E-MELD, n.d.).

[34] Dena'ina Case Study: http://emeld.org/school/case/denaina (E-MELD, n.d.).

[35] The Software registry: http://emeld.org/school/toolroom/software (E-MELD, n.d.).

[36] View the papers presented and the working group suggestions at the workshop proceedings page: http://emeld.org/workshop/2006/proceedings.html (E-MELD, n.d.).

References

Archive of Indigenous Languages of Latin America (AILLA). (n.d.). Retrieved October 27, 2006, from http://ailla.utexas.org

Ballanty, K. (2006). *Yapese corpora*. Retrieved October 27, 2006, from http://www2.hawaii.edu/~ballanty/corpusintro.html

Digital Endangered Languages and Musics Archives Network (DELAMAN). (n.d.). Retrieved October 27, 2006, from http://delaman.org

DoBeS Archive. (2006). *Dokumentation Bedrohter Sprachen (DOBES)*. Retrieved October 26, 2006

Electronic Metastructures for Endangered Languages Data (E-MELD). (n.d.). Retrieved October 25, 2006, from http://emeld.org

Hans Rausing Endangered Languages Project (HRELP). (2006). Retrieved October 25, 2006, from http://www.hrelp.org

Humboldt State University English Department. (2006). *Other sites of interest*. Retrieved October 27, 2006, from http://www.humboldt.edu/~english/sitesofinterest.html

National Science Foundation. (2005). *Documenting endangered languages (DEL): Program solicitation*. Retrieved October 27, 2006, from http://www.nsf.gov/pubs/2006/nsf06577/nsf06577.htm

Northern Arizona University. (2006). *Teaching Indigenous Languages: Indigenous Language Links*. Retrieved October 27, 2006, from http://jan.ucc.nau.edu/~jar/links.html

Open Language Archives Community (OLAC). (2006). Retrieved October 27, 2006, from http://www.language-archives.org

Rosetta Project. (n.d.). *The Rosetta project: Building an archive of all documented human languages*. Retrieved October 27, 2006, from http://www.rosettaproject.org

Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). (2005). Retrieved October 25, 2006, from http://paradisec.org.au

Wikipedia. (n.d.). *Endangered language*. Retrieved October 27, 2006, from http://en.wikipedia.org/wiki/Endangered_languages

Yale University Library. (2005). *Selected resources for African studies*. Retrieved October 27, 2006, from http://www.library.yale.edu/african/internet.html