Helen Aristar Dry
Eastern Michigan University
hdry@linguistlist.org

Workshop on Field Resources and Tools
LREC 2002
Las Palmas, Gran Canaria, Spain
May 26, 2002

# E-MELD: Overview and Update

*http://linguistlist.org/E-MELD/*

E-MELD (Electronic Metastructure for Endangered Languages Data) is a 5-year project with a dual objective: to aid in the preservation of endangered languages data and documentation and to aid in the development of the infrastructure necessary for effective collaboration among electronic archives. It is a 5-institution collaboration organized by The LINGUIST List; the collaborators include representatives of Eastern Michigan University, Wayne State University, the University of Arizona, the Linguistic Data Consortium at the University of Pennsylvania, and The Endangered Languages Fund at Haskins Laboratories, Yale University.

The E-MELD project was designed to produce multiple outcomes: (a) recommendations of best practice with regard to metadata, markup, and language identification; (b) a metadata server facilitating access to language archives worldwide; (c) a demonstration project showcasing data from 10 endangered languages (see Figure 1); (d) a Query Room, where questions may be addressed to native speakers; and (e) a systematic attempt to involve a large segment of the linguistics community in the standards-setting enterprise.

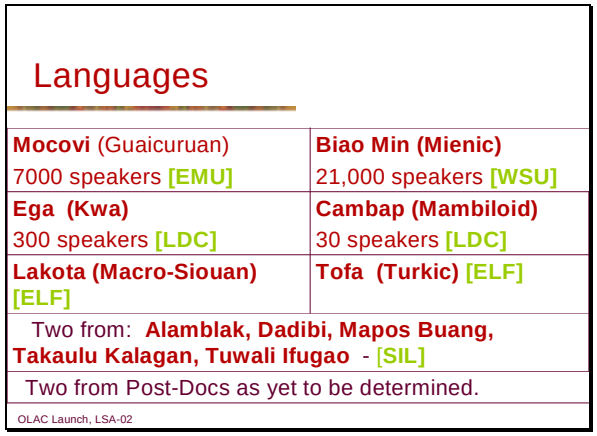


**Figure 1**

To publicize the endangered languages digitization effort, E-MELD sponsored one pre-project Workshop last summer in Santa Barbara, California. We will be holding another workshop in Ann Arbor, Michigan this summer, this one focused on lexicons. And we will be sponsoring yearly workshops for the duration of the project. The over-arching

goal of the sequence of workshops is to bring together field linguists and language engineers to explore ways to provide the widest access to, and the most reliable preservation of, electronic data on endangered languages. We also hope to use the resources of The LINGUIST List to acquaint ordinary linguists with the benefits and challenges involved in the digitization of language data.

Within the context of the burgeoning number of digital archives devoted to languages, the E-MELD project sees itself as focusing on breadth of coverage and simplicity of access. In part, this is because of the nature of the LINGUIST List. The LINGUIST email list currently has over 15,600 subscribers in 105 different countries; and we also host 97 other linguistics-related email lists on our site. In addition, LINGUIST has 4 full mirror sites in Europe, and together the 5 websites get almost 3 million hits a week. LINGUIST therefore has broad outreach into the linguistics community. However, though we have received a number of NSF grants for development of specific facilities, our day-to-day operations are funded by subscribers, through fund drives. We are not government-supported, and we do not have a permanent technical staff.

That is one reason our project emphasizes open standards and the use of existing, web-based applications. Another is that this is a way to insure wide access and rapid development of basic tools. The task of documenting endangered languages is so pressing that we feel that there is a need for breadth, as well as depth, in identification and dissemination of EL resources. If the EL documentation enterprise is to succeed, we will need both state-of-the-art technical support--of the kind of that DOBES fieldworkers get through the Max Planck Institute--and a more generalized data identification and collection effort of the kind that LINGUIST is equipped to spearhead.

E-MELD is the younger of the two projects sponsoring this workshop, having officially begun in September, 2001; but, in its first year of existence, E-MELD has made progress in metadata retrieval, language identification, and morphosyntactic markup. I will present an overview of project work in these areas; later in the Workshop, others will discuss the work in more detail.

## Metadata

In order to reach the widest possible audience, E-MELD transmuted its project goal of mounting a central metadata server into implementing an OLAC service-provider. OLAC, or the Open Language Archives Community, is a sub-community of the Open Archives Initiative, a cross-disciplinary initiative which promotes multi-archive searching by means of http protocols. OLAC is being launched at another symposium during this conference. But, briefly, it works by a very simple mechanism: participating archives or linguists, known as “data-providers,” describe their resources using the OLAC metadata set, which is based on the 15-element Dublin Core. Search facilities like The LINGUIST List, known as “service-providers,” “harvest” metadata records through periodic http requests.

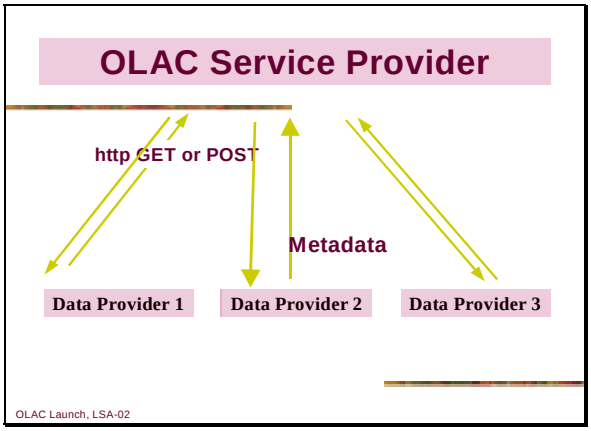


**Figure 2**

Service-providers harvest the metadata records using either the http POST or GET method and key/value pairs which typically consist of *verb* = followed by one of 6 allowable verbs, such as *ListRecords* or *GetRecord*. The response to such a request will be metadata on the various items in the archive, which the harvester will then expose to the user community via a search interface.

OLAC has refined 2 of the Dublin Core metadata elements to better meet the needs of the language community (see figure 3). First, the DC element "type" has been refined to reference "linguistic type," e.g. *narrative, lexicon*, as well as the broader DC types, such as *software, text,* and *collection*. A controlled vocabulary for linguistic types is now under discussion. Second, the DC element "subject" has been refined to reference "subject language," or the language that the resource is *about*. This was a necessary addition since the DC element "language" references the language that the resource is written in.

**OLAC Metadata . . .**

built on Dublin Core set of 15 elements:

Contributor
Coverage
Creator
Date
Description
Format
Identifier
Language
Publisher
Relation
Rights
Source
Subject
  Subject.language
Title
Type
  Type.linguistic

**Figure 3**

An archive or field linguist may become an OLAC data-provider simply by creating OLAC-compatible metadata for their resources and making this metadata harvestable in

one of 3 ways (described at http://www.language-archives.org/docs/implement.html). One method is to implement a software interface, but this is recommended primarily for archives with catalog data which is already in a database. E-MELD is particularly interested in serving individual linguists whose language data and documentation does not reside in an archive. It is quite simple to make such data harvestable by an OLAC service provider; it requires only that the resources be described using some or all of the 15 elements in the OLAC metadata set and that a web-readable XML file of the metadata be created. The metadata can then be harvested using the Virtual Data Provider (VIDA) created by the Linguistic Data Consortium; this harvester is activated simply by typing a URL into a browser.

To create the metadata, or harvestable "repository," a linguist can follow either of 2 procedures:

(1) create OLAC metadata herself using an XML editor and following the XML schema for OLAC repositories found at: http://www.language-archives.org/OLAC/0.4/oryx.xsd. (An example can be found at: http://www.language-archives.org/OLAC/0.4/oryx.xml ). Then upload the file(s) to a website and activate VIDA via a browser.

(2) use the OLAC Repository Editor (ORE) developed by the Linguistic Data Consortium. ORE creates OLAC metadata based on information entered into web forms, stores it on a website, and allows you to harvest it simply by selecting the "Activate repository" link.

The E-MELD project is particularly interested in encouraging field linguists to make information about un-archived EL resources available to the linguistics community. Consequently, The LINGUIST List has made the OLAC Repository Editor available at: http://saussure.linguistlist.org/olac/ore/

The LINGUIST List has also implemented an OLAC harvester at http://linguistlist.org/olac/ . This harvester now retrieves over 18,000 records from 13 major archives, including ELRA, LACITO, the Perseus Project, and others. The data which our OLAC harvester retrieves is written to a database on the LINGUIST site. This is a relatively simple database but one that uses some of the same codes and controlled vocabulary which is employed elsewhere on the LINGUIST List site. We intend to set up utilities which will merge the harvested metadata seamlessly with the main LINGUIST and E-MELD databases, so that searches will access all data sources on our site.

## Language Identification and Classification

One feature which we expect to be particularly useful to linguists is database searching by subject language. To implement this type of searching LINGUIST found it necessary to adopt an unambiguous means of referencing all languages and language families.

A computational search system must able to be find all and only those data which are relevant to any one language. But most languages are known by multiple names. And,

conversely, many common names, e.g. *Quechua*, are used to refer to more than one distinct language variety. The fact that there are 56 different language varieties which are sometimes called *Quechua* may not be a major stumbling block to their human speakers and investigators; but it can be fatal to the search process of a machine. For mechanized information retrieval, language names must be replaced by codes. Hence the International Standards Organization has developed the set of language codes described in ISO-639-1 and 639-2. However, these codes have serious limitations for use by professional linguists. In the first place, the approximately 500 ISO codes fail to cover the over 6000 languages in the world, so many different languages share the same code; for example, all Australian languages share the code AUS. Furthermore, the ISO codes are not based on a linguistically consistent definition of "language," so in a number of cases two different ISO codes are assigned to mutually intelligible varieties which linguists would consider to be the same language.

The E-MELD Santa Barbara workshop concluded that the most nearly complete and consistent system of codes for extant languages is that of The Ethnologue. (For a comparison of the Ethnologue and ISO codes, see: http://www.ethnologue.com/iso639/default.asp). The Ethnologue codes, however, are intended only to include languages currently in use. Furthermore, though the Ethnologue categorizes languages by family, language classification is not a major focus, thus it does not guarantee that the genetic subgroupings represent the most current view of the linguistics community.

The Summer Institute of Linguistics generously provided E-MELD with their database of over 6000 unique languages codes and 42,000 alternate names for languages. E-MELD then undertook to complement the Ethnologue codes with codes and descriptions for ancient, extinct, and constructed languages and to devise a coding system for language families which incorporates genetic information into its syntax. Some 230 additional codes for ancient languages were added to the database, as well as some 20 codes for constructed languages. In the process, advice was requested from experts in the different genetic groupings; for example, Deborah Anderson of the University of California at Berkeley is currently reviewing all the Indo-European ancient languages to ensure the accuracy of the descriptions and classifications.

The complete set of codes is being proposed as an OLAC standard; and we are in the process of recruiting a group of linguists whose charge will be to advise OLAC, LINGUIST List, and The Ethnologue to ensure that this language coding system is expanded and modified in accordance with evolving linguistic thought. Gary Simons will be giving more detail about the language codes later in the workshop, as will Anthony Aristar, Gayathri Sriram, and Michael Appleby, who will discuss specifically language classification. The family coding scheme and the database architecture behind it support the definition of multiple family trees for a given subgroup, each of which may represent a different scholarly view of the set of linguistic relationships. Information about the provenance of the subgroupings can also be incorporated, so that users can discover which linguists proposed or advocated the variant classifications.

The complete set of codes and a search facility can be accessed from:

http://linguistlist.org/languages/

All data on the LINGUIST site is now categorized according to these language codes. And you can search for people, linguistic programs, dissertations, books, and thousands of linguistically-relevant web links by subject language. For example, searching for publications by subject language is available at the URL:

http://saussure.linguistlist.org/pubs/

More importantly for the E-MELD project, however, the metadata server which we instituted allows you to search for resources in language-related archives by subject language. Providing this facility has involved recoding the metadata from OLAC data-providers with the full Ethnologue/E-MELD language code set. Archive search via subject language is available at:

http://saussure.linguistlist.org/cfdocs/new-website/LL-WorkingDirs/olac/olac-search1.cfm

## Linguistic Markup

The third focus of E-MELD this year was linguistic markup. The Santa Barbara workshop identified morphosyntactic markup as the markup initial focus; and, in 2001-2, the U. of Arizona E-MELD team, led by Terry Langendoen, began developing tools for encoding, searching, and querying morphosyntactic information about endangered languages on the World Wide Web.

Initially E-MELD proposed to try to discover or promote community consensus on best practice in morphosyntactic markup. However, we rapidly realized that this approach was impractical, given the many different markup systems currently used by field linguists and in the literature of language description. The Arizona team also came to the conclusion that promoting a "gold standard" in morphosyntactic markup was unnecessary. Rather than attempting to dictate a standard, the markup group decided to create a system that would allow

(a) the individual field worker to submit data in whatever markup she prefers, and

(b) the searcher to retrieve all relevant data whatever its original markup system

So, for example, one field worker may mark a particular morpheme as "possessive" while another may mark a similar morpheme as "genitive." If the two terms refer to the same concept, the system should "know" this. And a user should be able to retrieve both sets of data using a single query term, i.e., either "genitive" or "possessive."

Central to such a system is an ontology of linguistic concepts, which the Arizona team has decided to create as a domain-specific ontology within the Standard Upper Merged Ontology (SUMO) developed by Niles and Pease (2001). Later in this workshop, Terry Langendoen will tell you more about this ontology and its place in the E-MELD

architecture. But essentially the ontology will act as an interlanguage, allowing different markups to be related via the concepts represented (see figure 4 below, taken from Farrar, Lewis, and Langendoen (2002).

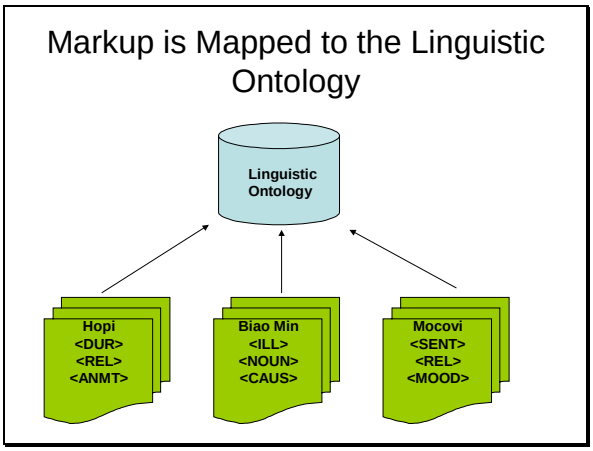


**Figure 4**

The ontology currently contains about 1000 morphosyntactic terms. Arizona has obtained permission from SIL International to use an ontology of about 500 terms prepared in 1997 to be part of their Lingua Links system, and has extended it with about 500 additional terms culled from standard sources such as David Crystal's Dictionary of Linguistics and Phonetics. They have incorporated it as a domain-specific sub-ontology within the IEEE SUMO upper ontology, and are in the process of restructuring the entries they have created to fit into SUMO. So far they have reanalyzed the set of terms for case, and are now working on those for tense, mood, and aspect.

Initially the team intended to require that researchers submit their data marked up in XML, even though it need not conform to a particular standard, like TEI. Now, however, they intend to create a tool which will translate common formats like Shoebox, Word, and RDF into XML and to create another tool which will help the linguist identify aspects of his markup with concepts in the ontology, should he wish to do so. This enriched data representation will be fed to an expert system shell (JESS) which will create a knowledge base consisting of the ontology plus the language facts. This knowledge base, in turn, can be searched by a “smart” web-based query engine, as suggested in Figure 5 below (from Farrar, Lewis, and Langendoen, 2002).

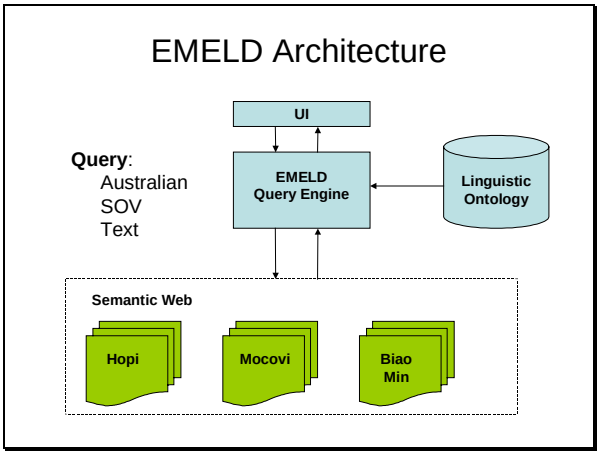


**Figure 5**

This approach is, we feel, more likely to be accepted within the linguistic community than the attempt to specify standards. Moreover, it is also more in keeping with the E-MELD project's general thrust, that of maximizing ease and accessibility, both for the linguist who supplies EL data and for the linguist who seeks to retrieve it. The ontology and accompanying database of language information will thus complement the metadata server and editor in helping us to identify as much EL data and documentation as we can and to disseminate information about these as widely as possible.

## References

Farrar, S., Lewis, W., Langendoen, T. (2002) A common ontology for linguistic concepts. In Proceedings of the Knowledge Technologies Conference, March 10-13, Seattle.

Niles, I. and Pease, A. (2001). Toward a standard upper ontology. In Proceedings of the 2nd Annual Conference on Formal Ontology in Information Systems (FOIS-2001).