

E-MELD: ELECTRONIC METASTRUCTURE FOR ENDANGERED LANGUAGES DATA

1. INTRODUCTION

Language data is central to the research of a large social sciences community, including not only linguists, but also anthropologists, archeologists, historians, sociologists, and political scientists interested in the culture of indigenous peoples. Members of this research community are currently faced with 2 urgent situations: the number of languages in the world is rapidly diminishing while the number of initiatives to create digital archives of language data is rapidly multiplying as a result of the increasing availability and sophistication of web technology. The latter might seem to be an unalloyed good in the face of the former, but there are 2 ways things may go wrong without adequate collaboration among archivists, linguists, and language engineers. First, a common standard for the digitization of linguistic data may never be agreed upon. And the resulting variation in archiving practices and language representation would seriously inhibit data access, searching, and scientific investigation. Second, standards may be implemented without guidance from the people who best know the range of structural possibilities in human language—descriptive linguists who have documented hundreds of little-known languages. Guidelines which are designed on the basis of well-known western languages will not be adequate to the urgent task of archiving as much linguistic data as possible in the face of widespread language attrition and loss.

If digital archives of language data and documentation are to offer the widest possible access and to provide information in a maximally useful form, consensus must be reached about certain aspects of archive infrastructure. As the largest linguistic organization in the world and the discipline's central electronic publication, The LINGUIST List <<http://www.linguistlist.org>> is organizing a collaborative project with a dual objective: (1) to preserve EL data and documentation and (2) to aid in the development of infrastructure for linguistic archives. One outcome of the project will be a LINGUIST List digital archive housing data from 10 endangered languages (ELs). But the focus on infrastructure will produce other, equally important results. In the first place, The LINGUIST archive will function, not only as a repository, but also as a "showroom of best practice." The archive will offer EL data marked up and catalogued according to community consensus about best practice; it will also disseminate reference material delineating best practice and software tools supporting it. A second outcome will be the establishment on the LINGUIST List site of a central metadata server for the discipline; this server will eventually organize information on all the language-related resources residing at distributed sites, not just information on EL data alone. And a third outcome—perhaps the most important—will be the involvement of a large segment of the linguistics community in the various enterprises underlying the archive and server. Capitalizing upon LINGUIST's high profile within the discipline, we will launch summer workshops and "digital institutes" to formulate and review recommendations of best practice and to train a substantial core of linguists and language archivists in their implementation. At the same time we will use the many avenues of electronic communication open to us to publicize the results of these meetings and solicit input from linguists across the world.

Although the data collection efforts will focus initially on endangered languages, the metadata server, the recommendations for best practice, and the distribution of supporting software will have a significant impact on all empirical research in linguistics. Thus the project will add value to all the other language-related projects receiving public support currently planned or underway.

2. THE PROBLEM

2.1 LANGUAGE ENDANGERMENT

Grimes (1996) estimates that there are 6703 languages spoken in the world today. LaPolla (1998) has run statistical analyses based upon census and population estimate figures that show an alarming number of these, perhaps as many as 50%, are in real danger of extinction. Fifty-two per cent of the world's languages are spoken regularly by less than 10,000 people, 28% are spoken by less than 1,000 and 10% by less than 100 (LaPolla 1998). The editors of *Ethnologue* <<http://www.sil.org/ethnologue>> estimate that 52 languages have only 1 native speaker left, and 426 are "nearly extinct." By contrast, 49% of the world's population speaks one of 10 major languages (Mandarin, English, Spanish, Hindi, Portuguese, Bengali, Russian, Japanese, French, German) as their mother tongue. Indeed, the impending disappearance of so many languages has become so pressing a problem that it has generated notice in the popular media and considerable public concern (see, for example, *Harpers Magazine* Aug.,

2000; *Newsweek International*, June 19, 2000; *The Independent*, July 16, 1999; *Daily News*, Aug. 19, 1999; *Florida Today*, May 16, 1999).

As scientists, we have a twofold reason to be concerned about this trend in rapid language loss. First, and most importantly, the death of a language or dialect represents a significant loss in knowledge and culture. Language serves as a primary means of cross-generational cultural transmission. And the death of a language may represent a serious impediment to the survival of the community—comprising, as it often does, the loss of the community's traditional poetry, songs, stories, proverbs, laments, and religious rites. Second, the death of a language or dialect represents a serious academic loss (Hale 1996, Woodbury 1993). Studies of linguistic diversity and cross-linguistic comparisons drive much of linguistic theory. Such studies also provide valuable information about population movements, contacts, and genetic relationships; thus they figure as well in research in anthropology, archaeology, history, and ethno-biology. Many (if not most) of the endangered languages have not been well studied or documented. When such a language disappears, then, there are 2 losses: the loss of valuable scientific data, and the loss of the knowledge and worldview it represents.

2.2 DIGITIZATION EFFORTS

The topic of language endangerment has thus become important to linguists of all theoretical backgrounds and areas of specialization, to professionals in related disciplines, and to concerned citizens who value cultural diversity. It is the focus of endangered language organizations across the world, e.g., the Linguistic Society of America Committee on Endangered Languages and their Preservation (CELP), The Foundation for Endangered Languages (FEL), Terralingua (TL), the International Clearing House for Endangered Languages (ICHEL), and The Endangered Language Fund (ELF). For this reason, a number of digital archives of EL data are currently being planned or developed. Among the most prominent within the United States are:

- The Linguistic Data Consortium at the University of Pennsylvania: <http://www ldc.upenn.edu/>
- The proposed web-based Archive of Indigenous Languages of Latin America (AILLA) at the University of Texas: <http://www.talkbank.org/exploration/#AILLA>
- The Comparative Bantu OnLine Dictionary (CBOLD) at the University of California, Berkeley: <http://faust.linguistics.berkeley.edu/CBOLD/info.html>
- The Center for the Documentation of Endangered Languages (CDEL) at Indiana University: http://www.indiana.edu/~aisri/projects/projects_frames.htm
- The Project for the Documentation of the Languages of Mesoamerica (PDLMA) at SUNY Albany: <http://www.albany.edu/anthro/maldp/index.html>
- The planned electronic wing of the SIL Language and Culture Archive: <http://www.sil.org/>

Significant projects outside the US include:

- The Oxford Text Archive: <http://ota.ahds.ac.uk/>
- LACITO (Langues et Civilisations à Tradition Orale) at CNRS, Paris: <http://lacito.vjf.cnrs.fr/>
- Activities focused on documenting minority languages at the University of Lancaster and the Max Plank Institutes (Nijmegen and Leipzig)
- The extensive collection on Pacific languages at the Australian National University: <http://www.sbg.ac.at/docu/net/wais/anu-endangered-languages-l.htm>

Appended is a list of 47 web sites dedicated to ELs, not including the sites listed above or sites which are focused on culture only (see Supplementary Documents, Part 2). Not all of these sites plan to establish a fully developed archive, but they all host collections of texts, grammars, and teaching materials. Thus they are potential sources of EL data and metadata.

The establishment of multiple archives is to be welcomed, since the magnitude of the task requires distributed effort. No one institution can archive all the important data on all the currently endangered languages—certainly not within the time limits imposed by impending language attrition and by the ongoing deterioration of the existing documentation. Paper, audiotapes, videotapes, and computer diskettes are all prone to degradation and destruction. Moreover, most field notes and grammars currently reside on individual computers, vulnerable to disk crashes as well as file corruption. Some older notes and grammars still exist only in the form of notebooks and file cards. Because language data is difficult to publish commercially, it may be stored negligently or even abandoned once the research based on it has been completed. Even when such material is deposited in

conventional libraries, data preservation is not certain, because many libraries can not offer optimum storage conditions.¹

Digital archiving at distributed sites offers the best hope for preserving this valuable linguistic material. But developing all the infrastructure necessary for a digital archive of language data (including delivery mechanism, formatting guidelines, and supporting software) is a huge task that is beyond the capacity of any single institution to accomplish on its own (Simons, 2000b: 1). And once multiple institutions have set up online archives, resorting to different strategies for designing infrastructure, it will be more difficult to implement any general solution.

Without such a common infrastructure, the individual linguist will find it very difficult to identify all the resources pertinent to a given language. To posit an extreme case: the language in question may be classified, or even named, differently in different archives (e.g., *Waikurean* vs. *Guaicuruan*, *Lappish* vs. *Sami*). The language data may be marked up using different sets of structural tags (e.g., *possessive* vs. *genitive*). The texts may have different organizations (e.g., chronological organization vs. frequency organization of the meanings in a dictionary entry). And the files may have different formats because they have been created with incompatible software tools. In this situation, even a linguist with access to resources might not be able to compare them well enough to make reliable linguistic judgments. But—what is perhaps even more disturbing—locating all the relevant material in the first place will be a formidable task. It is unlikely that all the sound and video recordings, texts, grammars, dictionaries, and cultural information pertinent to a given language will ever reside on a single site. And if various archives develop different ways of describing and indexing their resources, no central meta-index can easily be developed. The amount of data will defeat a human librarian, and the different formats will defeat a machine.

2.3 THE SCOPE OF THE PROBLEM

It should be emphasized that all of the problems enumerated above arise in the context of archiving *any* electronic language data, not EL data alone. It is the impending disappearance of so many endangered languages that leads us to focus first on this aspect of the more general language data problem. However, this focus has a distinct—although paradoxical—benefit: the challenging nature of the data set. Many, if not most, ELs have structures which diverge so widely from each other and from those of western European languages that metadata and markup guidelines adequate for these languages will almost certainly be adequate for other language data as well. Thus an attempt to define standards for the digitization of ELs is, in fact, also an attempt to define standards for the digitization of languages in general. And all the facilities developed to provide access to ELs can—and will—be extended to provide access to other linguistic data as well.

3. TOWARD A SOLUTION: E-MELD

Any attempt to address the language archiving problem must have at least 3 components.

1) **Community Involvement.** All the different stakeholders in the EL archiving enterprise must be kept fully informed and continually consulted. In particular, we must foster communication between computational linguists and field linguists; since a computational solution developed without the input of descriptive linguists will never become widely accepted. To the extent possible, we must also involve indigenous communities: native speakers of ELs should take part in markup formulation and community leaders in archive design.

2) **Flexibility.** Any proposed solution must (a) have the capacity to handle legacy data in various formats and (b) allow for some continuing variation in individual practice. Not only will different languages and theories always call for different analytical categories, but different research questions will always call for different types of data manipulation and display.

3) **Collaboration.** Organizations must pool their resources in light of: (a) the volume of work and the range of expertise needed for a unified solution and (b) the danger that partial, uncoordinated “solutions” will only exacerbate the problem (see 2.2 above).

¹ Surveying the counties of Waterloo and Wellington in southwestern Ontario, Ritchie (1995) reports that he visited over 350 archival repositories but “identified only 4 with storage facilities for audio-visual materials that even came near the environmental conditions required for ... [black and white] film.”

The E-MELD project has been structured with these 3 requirements in mind. It implements part of a distributed solution proposed in Simons (2000b), which recommends a coordination of effort among the Linguistic Data Consortium (LDC), the Summer Institute of Linguistics (SIL), and The LINGUIST List: The Linguistic Data Consortium will function as a central repository of standards and software (which may be developed elsewhere); the SIL Ethnologue will constitute the standard reference for language classification; and The LINGUIST List will serve as a central repository of metadata, as well as an institutionalized conduit of information between language engineering projects and the linguistics community.² LINGUIST has already taken several steps toward assuming its suggested roles (see *Project Preliminaries*, 3.7.1 below). The E-MELD project, which involves The Endangered Languages Fund and The University of Arizona, as well as the 3 institutions named above, represents significant, unified progress toward this collaborative goal.

3.1 PROJECT COMPONENTS

In its general outlines the E-MELD project involves:

- Formulation and promulgation of best practice in:
 - Linguistic markup of texts and lexicons
 - The creation of metadata for language resources
- Establishment of a metadata server on the LINGUIST List site:
 - Creation of user-friendly web interfaces for input, query, and display
 - Collection of metadata on existing language resources (not just EL resources)
 - Conversion of metadata in foreign formats into the best practice format
 - Provision (in addition to metadata about language resources) of:
 - Typological information collected via questionnaire
 - Genetic and ethnographic information provided via an interface to the Ethnologue (<http://www.sil.org/ethnologue>)
- Software development
 - Development of markup conversion software
 - Development of software for field linguists facilitating best practice
- Establishment of a “showroom of best practice,” making available:
 - Texts and lexicons from 10 ELs converted into best-practice format
 - Software tools (described above)
 - Reference material (e.g., files & hyperlinks) delineating recommended standards
 - A Query Room, where questions may be addressed to native speakers and additional data provided upon request
- Organized communication with the research community, involving:
 - 2 Workshops: weekend sessions intended to promote communication among field linguists, archivists, and computational linguists.
 - 3 “Digital Institutes”: one-week institutes for 10-15 field linguists designed to:
 - Introduce proposed recommendations of best practice
 - Distribute the field software and provide training in its use
 - Test both the software and the proposed markup on samples of the participants’ data
 - Liaisons with professional associations, e.g. CELP, TL, FEL, ICHEL
 - Regular email bulletins distributed via The LINGUIST List
 - An E-MELD homepage on the LINGUIST List site offering detailed reports on Workshops and “Digital Institutes,” web questionnaires about proposed guidelines, and organized sets of links to related sites, e.g.,

² See attached letters of support from Brian MacWhinney of the Talkbank Project, Peter Wittenberg of the Volkswagen-funded Project at Max Planck Institute, Nijmegen, and Louanna Furbee, LSA Archivist. Additionally, Bruce Connell and Dafydd Gibbon of the EGA Project (<http://coral.lili.uni-bielefeld.de/EGA/>) will provide data to E-MELD. One PI (Bird) and one consultant (Simons) are involved in the ISLE initiative, as well as the Linguistic Exploration Project and the Linguistic Annotation initiative at the LDC. And Nicholas Ostler, Head of the European Foundation for Endangered Languages is a confirmed participant in our Summer 2001 Workshop.

- EL sites (see Supplementary Documents, Part 2)³
- Sites involved in language engineering⁴

The most important components of the project are described in more detail below.

3.2 METADATA

Metadata—or structured data about data—can be as simple as a keyword in a META field within an HTML document. Even the simplest kind of metadata can be useful as resource description, once the document is retrieved. But resource discovery requires the use of a standardized format. In a context as vast and rapidly expanding as the modern-day Internet, data is only valuable if it is findable, and if its relevance is interpretable through computational means. Thus one of the most important parts of the E-MELD project is the initiative to collect metadata on language resources at a central site. Though we will focus initially on EL resources, the facilities created will be extended as soon as possible to catalogue linguistics-related resources of all types.

Such a catalogue will not only allow extant material to be identified and retrieved; but it will also enable distributed data to be pieced together. Given a markup standard and a metadata server, it will not matter if a dictionary of a language appears at one site and a grammar of the same language appears at another. They can be linked through their metadata, and used in conjunction with one another. But in order to establish such a central index, it will be necessary to reach consensus about best practice in the creation of metadata for language resources, as well as to collect existing metadata and convert it into this format, and to institute user-friendly systems for input and query of the information.

3.2.1 LANGUAGE RESOURCE METADATA

One possible starting point is the standard defined by the Dublin Core Metadata Initiative (<http://www.purl.org/dc/>). The Dublin Core Element Set is limited to 15 elements standardized to begin with the prefix “DC,” as in “DC.Creator.” Although some difficulties have been identified with large-scale implementations,⁵ the DC metadata standard is gaining wide, cross-disciplinary acceptance, in part because of its simplicity as compared to the full MARC standard. Resources which bear metadata in DC format are interpretable and retrievable by many existing search tools (e.g., SWISH-E, freeWAIS-sf2.0, GLIMPSE, HARVEST, ISEARCH) and others which are currently being developed (e.g., BC, described at <http://www.mpi.nl/world/tg/lapp/lapp.html>).

However, any existing metadata standard will need to be augmented by recommendations of best practice specific to language resources. Suppose, for example, that we wish to use the DC Element Set to describe an online grammar of Mocovi written in English. Part of a plausible resource description might be the HTML header lines given as (1) below:

```
(1) <meta name      = "DC.Subject"
      content     = "Mocovi" >

      <meta name      = "DC.Type"
      content     = "grammar" >

      <meta name      = "DC.Format"
      content     = "text/html" >

      <meta name      = "DC.Language"
      content     = "en" >
```

However, for the `DC.Type`, `DC.Description`, and `DC.Format` elements, “recommended best practice is to select a value from a controlled vocabulary or formal classification scheme” (Dublin Core Metadata Element Set, Version 1.1). And no vocabulary or classification scheme appropriate for linguistics yet has general acceptance. The list of suggested DC types, for example, includes *collection*, *dataset*, *event*, *image*, *interactive resource*, *model*, *party*, *physical object*, *place*, *service*, *software*, *sound*, and *text* (Guenther 1999). It does not

³ Such a web page will be a natural extension of the Endangered Languages homepage recently set up on LINGUIST, in cooperation with the Linguistic Society of America Committee on Endangered Languages and their Preservation. <http://linguistlist.org/el-page>

⁴ For example, The Linguistic Exploration Project (<http://www ldc.upenn.edu/exploration/>), The TalkBank Project (<http://www.talkbank.org/>), The ISLE Project (<http://www ldc.upenn.edu/sb/isle.html>).

⁵ Morgan (1999), for example, noted a number of problems with the implementation of Dublin Core tagging at John Wiley Publishing, most notably that the 15 DC Simple elements are all optional and repeatable, and it is not always immediately obvious what they mean.

include *grammar*. Hence, the value ‘grammar’ for `DC.Type` in (1) has little general usefulness. It should perhaps be replaced by ‘text’, with ‘grammar’ provided as a value for `DC.Description`. But ‘grammar’ is not part of a recognized classification scheme for `DC.Description` either.⁶ Lacking such discipline-specific controlled vocabulary—and lacking agreement even about the elements that should be included in linguistic metadata—language researchers will have increasing difficulty finding electronic resources.

In developing and publicizing such lists, we intend to collaborate closely with the Linguistic Data Consortium and other groups, e.g., ISLE, which are also addressing the question of metadata format for language data. For instance, we intend to support the metadata initiative at the upcoming (Dec., 2000) Linguistic Exploration Workshop at the LDC in 2 ways: (a) by submitting examples of metadata collected from the advisors to our project and (b) by gathering feedback on the Exploration Workshop outcomes at our first workshop with field linguists and archivists, scheduled for June, 2001 (see *Project Preliminaries*, 3.7.1 below).

3.2.2 TYPOLOGICAL ‘METADATA’

Typological information may include generalizing statements about (a) the set of types into which a language may fall, e.g., Subject-Verb-Object ordering, as opposed to Verb-Subject-Object and (b) the classification of a particular language according to these types. Although it is not metadata in the sense used above—i.e., it is not data about language *resources*—it may be construed as a kind of metadata about the languages themselves. And it is data of such potential usefulness to linguists that it is worth extending our concept of the metadata server to collect and provide it. At present there is no way for a linguist to find out what languages of the world are SVO or VSO. The Ethnologue does not collect such information; and though typological databases do exist in the hands of individual linguists, these are not generally accessible.

For that reason, we intend to cooperate with related projects such as WALS (*World Atlas of Language Structures*: <http://www.eva.mpg.de/~haspelmt/atlas.html>) to formulate a web-based typological questionnaire which is brief enough to be practicable and yet contains questions about the information deemed most significant by a committee representing as much theoretical diversity as possible. The Summer Institute of Linguistics has agreed to ask their numerous field linguists to answer this questionnaire with regard to the languages they have studied. The online facility will also, of course, also record and display variant answers from linguists who disagree with the original descriptions; but the initial collection procedure will immediately provide a wealth of typological data relevant to pressing questions of EL research.

3.2.3 METADATA COLLECTION

The E-MELD project will create user-friendly web interfaces for metadata input; and the PIs will contact cooperating archivists to request their metadata. In addition, LINGUIST intends to implement an innovative procedure to identify other sites on the Internet which store language data but have not yet participated in the project. This will involve using a spider to index other linguistics-related sites and configuring search software to search the index using a keyword list. In this way potential sources of metadata may be identified. The site owners will then be approached and invited to contribute to the database.

The spidering procedure will exploit LINGUIST’s comprehensive collection of relevant URLs. Almost all links related to language and linguistics are announced on The LINGUIST List, and we have had a “URL-grabber” operative on the site since 1994. This custom software copies each link that passes through LINGUIST to a file. Now that LINGUIST hosts 70 other language-related lists (see *Track Record*, 4.2 below), our ability to collect linguistic URLs has been greatly enhanced. Our collection of links is a natural domain which a spider can index, as a first step to finding additional language data.

Furthermore, the index itself will be an extremely useful linguistic search tool—of a kind which, to our knowledge, no other discipline has. Since it will index only linguistics-related sites, searching this index will not return the unwieldy amount of irrelevant information that linguists inadvertently retrieve from a general web search engine. It will be a linguistics-specific Internet search facility; and it will be made freely available on the LINGUIST List site.

3.3 MARKUP

Markup is systematic annotation designed to reveal a text’s typographical and informational structure. Linguistic markup—a particularly challenging sub-variety—might be broadly described as annotation

⁶ Indeed, `DC.Description` has been called a “bucket” category (Morgan 1999), suggesting that adequate controlled vocabularies or classification schemes do not yet exist for other kinds of data either.

representing: (a) the grammatical structure of text couched in the focus language and (b) the structure of documents presenting a linguistic description or analysis of such text. Linguistic markup is required in the digitization of such language documentation as paradigms, word lists, dictionary entries, and glossed text. And most language documentation invites both types of annotation (although, of course, a distinction may be maintained, e.g. in the use of parsing vs. stylesheet software).

3.3.1 GLOSSED TEXT

Markup for interlinearized text, for example, must represent both the phonological, morphological and syntactic structure of the text and enough additional information to allow reconstruction of the conventionalized formatting which makes the information intelligible. This is exemplified in the fragment of a Mocovi text (Grondona 1998) given as (2) below:

- (2) Glossed Mocovi text fragment
- | | | | | | | | |
|----|---|------|---------------|--------------|------------------|-----|------------|
| a. | kaʔmaq | yale | yowitoʔ | ka | lawoʔ | ka | ʔna:koʔ: |
| b. | ka-ʔmaq | yale | i+owir+oʔ | ka | l+awo-r | kaʔ | Ø+ʔna:k+oʔ |
| c. | CLASS(absn)-ʔ | man | 3AC+arrive+EV | CLASS(absnt) | 3POSS+family+PCL | and | 3AC+say+EV |
| d. | And the man arrived [to] his relatives and (he) said: | | | | | | |
| e. | And the man arrived where his relatives were and he said: | | | | | | |

Following the format that is conventional in printed grammars, the Mocovi text and its analysis are laid out in 5 lines: (a) phonemic transcription in the source language, (b) the underlying morpheme sequence, (c) a morpheme-by-morpheme gloss into a partly technical vocabulary, (d) a literal English translation, and (e) a free translation. To replicate this in electronic text requires that the markup represent:

- The technical terms and abbreviations used in the glossing. Such terms lie at the heart of any linguistic markup; they must be assigned tags and attributes in a way that can support validation and searching.
- Correspondence information that allows alignment of a particular morpheme (line b) with its gloss (line c), and a phoneme sequence (line a)—e.g., information that the *ʔ* in *lawoʔ* corresponds to the *-r* in *l+awo-r* and is glossed as *PCL*.⁷
- Distinctions among morphemic boundaries—e.g., information that the ‘-’ delimiter indicates ordinary affixation; the ‘+’ delimiter cliticization.

3.3.2 LEXICAL ENTRIES

Lexical entries also require such markup distinctions, in addition to others specific to dictionary formatting. Consider, for example, (3) below (from Molina et al., 1999).

- (3) *rukte* *iv.* move, stir (*rukti-*, *rukrukte*); *Ime ne ~k.* I moved over here.
aavo ~ iv. approach, get close

This 3-line example includes 2 variant forms of the stem: *rukti-*, which is used with verbal suffixes except perfective *-k* (e.g., *ruktime* ‘will move’); and *rukrukte*, the reduplicated form. It also includes an example sentence, its gloss, and a sub-entry (*aavo*) with the subentry’s form, part of speech, and definition. Since this and similar grammatical, definitional, and illustrative information appears routinely in language documentation, it should be represented consistently and concisely by a markup scheme.

3.3.3 THE NEED FOR GUIDELINES

Guidelines for best practice in linguistic markup are crucial for the understanding of language data for 2 reasons (both of which are especially relevant to ELs because of their complex and unfamiliar structures). First, without compatible markup, no two bodies of data are comparable. The linguistic similarities and differences will be difficult to see even by human inspection. Computationally they are essentially undiscoverable, since no search-engine can be expected to “know” that differently named entities are equivalent. Second, a lack of standardization makes data difficult to interpret in and of itself, because a linguist must first learn the nature of the data markup before he or she is able to understand a new body of data.

⁷ Some ideas for how to do this are presented in TEI guidelines (Sperberg-McQueen & Burnard 1994; henceforth TEI P3) Ch 14 “Correspondence and alignment.”.

A start was made on the difficult issue of linguistic markup as part of the Text Encoding Initiative (TEI) (<http://www.uic.edu/orgs/tei/>), an international project designed to develop guidelines for the preparation and interchange of electronic texts for scholarly research. Two of the main participants in this part of the TEI were Terry Langendoen, of the University of Arizona, and Gary Simons of the Summer Institute of Linguistics, both of whom will be part of the E-MELD project.

Unfortunately, the work on linguistic markup within the TEI was incomplete when the project ended. Moreover, many of the markup structures recommended by TEI are too unwieldy to have gained much popularity. For example, Chapter 12 “Print dictionaries” of the TEI guidelines (Sperberg-McQueen & Burnard 1994; henceforth *TEI P3*) suggests the encoding given as (5) below for the basic structure of a simple dictionary entry such as (4) from Molina et al. (1999).⁸

(4) Lo’i *adj.* crippled, lame

```
(5) <entry language=EN id=loi>
    <form language=Yoeme>
      <orth>lo’i</orth>
    <gramGrp>
      <pos>adj</pos>
    </gramGrp>
    <def id="crippled">crippled</def>
    <def id="lame">lame</def>
  </entry>
```

In (5) the `<gramGrp>` ‘grammatical group’ tag has 2 salient design flaws. First, the tagset that is valid within it, namely `<pos>` ‘part of speech’, `<gen>` ‘gender’, `<number>`, `<case>`, and `<itype>` ‘inflectional class’, is too specifically tailored to the major European languages to be of much use for natural languages generally. Second, the values for these tags are ordinary text (technically #PCDATA in SGML). The possible values for the `<pos>` tag in this case are listed in the table of abbreviations section in the print dictionary from which (4) was taken. Reference should therefore be made to this list, and ultimately to master lists of grammatical tags maintained by the electronic archive. *TEI P3* Ch. 15 “Feature structures” and Ch. 26 “Feature system declarations” provide one mechanism for referencing a short list of predefined feature value pairs, which enables designers to design and maintain their own inventories of grammatical structures.⁹ However, this complex mechanism has never gained acceptance in the linguistic community.

More streamlined recommendations need to be developed. But the simpler markup standards currently in existence—such as those developed by the Expert Advisory Group on Language Engineering (EAGLES) (<http://www.ilc.pi.cnr.it/EAGLES/annotate/annotate.html>)—are designed to handle only western European languages.¹⁰ As a result there still exists no set of guidelines adequate for the markup of EL data.

The E-MELD project will modify and extend existing specifications for linguistic markup in light of the particular needs of the EL community. Although we will tackle the full range of problems that arise, we will focus initially on the markup of dictionaries and of glossed text, since successful development of recommendations in these 2 areas will deal with 3 of the most pressing problems currently faced by the linguistic community. One is lack of standardization in electronic texts representing language data,¹¹ and a corresponding lack of interoperability in corpus-handling software. Electronic EL texts, in particular, have been developed using a wide variety of different standards, and, while they are usually consistent in themselves, little thought is typically given

⁸ The guidelines are for encoding printed dictionaries in SGML, the markup language of choice at that time, but the recommendations could be followed for marking up existing print dictionaries of ELs in XML (Extensible Markup Language), and even for the electronic preparation of new dictionaries.

⁹ See Langendoen & Simons 1995 for discussion of the design of feature structures within the TEI framework.

¹⁰ The EAGLES group is continuing its work as part of the International Standard for Language Engineering group (ISLE) (http://www.ilc.pi.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm), in conjunction with the Linguistic Annotation project at the Linguistic Data Consortium, but there has as yet been no specific linguistic markup standard developed, although general guidelines have been widely accepted. The LINGUIST project has advisors closely associated with both ISLE and the LDC, so we will not duplicate work underway in either workgroup.

¹¹ This problem is far less characteristic of electronic dictionaries, since prestigious organizations, e.g. the Oxford Text Archive and the OED, have already taken the lead in designing and exemplifying best practice in entry formatting. However, the problem of linguistic markup *within* entries still remains. Since these electronic dictionaries treat primarily western European languages, we still lack adequate markup to represent the morphological and syntactic structures of ELs.

to making them conform to a more widely acceptable format.¹² A second problem is the inability to find and compare instances of specific grammatical forms. A linguist, for example, might well need to extract from a set of glossed texts every example of a first-person plural object marker occurring in the same word as a third-person singular subject marker. Or she might wish to query a lexical database of the Yoeme material to find the base forms for every verb that has a distinct variant when the subject is plural (a matter of considerable interest for doing comparative Uto-Aztecan research). To do this simply by looking through the printed text would be prohibitively time-consuming. And a third difficulty in doing comparative linguistic research is understanding how and when the same technical term is understood differently in different texts, and conversely how different terms may be understood to denote the same thing. (A simple example: when should one use “genitive,” when “possessive,” and when doesn’t it matter?) Attempts have been made to systematize grammatical descriptions (e.g., Comrie and Smith 1987) but these still need to be translated effectively into a computational environment. Success in standardizing electronic dictionary and glossed text markup will make a start toward providing a *catalogue raisonnée* of the morphological distinctions found in the languages of the world, with guidelines for their encoding.

3.4 DATA FORMATS AND SOFTWARE TOOLS

In tandem with the linguistic markup formulation described above, a series of best-practice guidelines for the storage of lexicons, interlinear texts, and a variety of other primary data types are being developed by ongoing community initiatives (see footnote 2). All of these guidelines will be tested, refined and exemplified in the context of the pilot corpus construction described in Section 3.5.1, and software tools will be adapted or developed by the LDC and SIL to support the associated formats.

3.4.1 PILOT CORPUS CONSTRUCTION

The LDC will select suitable XML formats coming out of the community initiatives, and apply them to the primary EL data we intend to archive. This material is being collected now using existing tools (e.g., Shoebox), and will produce data in a range of legacy formats. The use of existing tools will avoid the delay caused by developing project-specific software. It will also force us to address the problems posed by legacy formats at the outset, and facilitate the reformatting of other third-party legacy data contributed to the archive.

The legacy data will be immediately converted into one or more of the new XML formats, using a shared, open source library of conversion functions that we will develop. These functions will be written in C++ and Perl, and perhaps other scripting languages as appropriate. The intention is that parsers and character translation tables for a given legacy format should only have to be developed once.

The data conversion process is expected to reveal shortcomings of the proposed XML format, and so it will be necessary to refine the format during the first year of the project. The overall outcome of this activity will be (i) best practice guidelines for data formatting; (ii) a library of conversion functions taking legacy formats to the new XML formats; and (iii) a pilot corpus exemplifying the formats.

3.4.2 ADAPTING LANGUAGE ENGINEERING SOFTWARE

One of the project PIs (Bird) is associated with 2 initiatives (e.g., ATLAS, ISLE) to produce a standard architecture, application programming interface, and interchange format for linguistic data, along with domain-specific and task-specific implementations. We will integrate our activities with this work, in order to allow maximal reuse of third-party software components for creating, storing and accessing primary data. Such integration will permit us, for example, to use (and adapt) an automatic alignment tool which will approximately align a text transcript with the corresponding audio, greatly facilitating content searches on large audio collections. Another benefit of this integration is data reuse: the data we archive will be usable without further format conversion by anyone who uses software that complies with these standards.

We will interact with the ATLAS and ISLE projects on the requirements for storage of each of the data types of interest, and will create ATLAS-compliant interfaces and/or wrappers for any software we develop, where appropriate. The overall outcome of this activity will be compliance with new industry standards for language engineering systems and formats.

¹² For example, the Australian National University, the Summer Institute for Linguistics, and the U. of California at Berkeley have developed formats for EL texts which are followed consistently within each publishing domain; but the 3 formats are not consistent with each other.

3.5 THE “BEST PRACTICE SHOWROOM”

Our work on formats and tools will be disseminated in a practical fashion, as a “Best Practice Showroom.” Although XML DTDs and coding manuals will also be provided, the focus is on a concrete demonstration of how adherence to our best practices results in the best kind of language documentation: searchable, citable and reusable primary data along with associated descriptive and analytical text, accessible in perpetuity. The ‘Best Practice Showroom’ will include data from 10 endangered languages, marked up as recommended, and user-friendly web interfaces for input, query, and flexible data display. In addition, the ‘Showroom’ will provide a Query Room, where native speakers may be consulted about language questions. Finally, by means of its metadata retrieval capabilities, the site will aid the researcher in finding other relevant linguistic resources.

3.5.1 THE DATA

Providing EL data is important, not only because it will preserve information about the languages in question, but also because it is not possible to promulgate standards in the abstract. Rather, data must be offered, both as an example and a lure. Linguists will not enter the showroom to learn about “best practice” so much as to access the data. In working with the data, however, they will become familiar with the guidelines and, we believe, will become cognizant of their benefits.

The LINGUIST List, the LDC, the ELF and SIL have each agreed to provide data from 2 languages and convert it into the recommended format. Data from 2 additional languages will be provided by 2 post-doctoral associates. These post-doctoral positions have been instituted with 2 aims in mind: (1) to see how effective our markup is at handling new, complex bodies of data, and (2) to bring our best-practice guidelines to the notice of younger scholars in field work—not only those who will hold the post-docs but also those who will initially compete for them. We will thus open the “Showroom Of Best Practice” with data from 10 diverse languages, including: Biao Min and Mocovi (data prepared by LINGUIST), Ega and Cambap (data prepared by the Linguistic Data Consortium), Tofa and Lakota (data prepared by the Endangered Languages Fund), and an Austronesian and a Papua New Guinean language as yet to be determined (data prepared by SIL). Brief language descriptions are provided below in order to illustrate (1) the value of the data itself to the EL preservation effort and (2) the challenging range of linguistic features represented. These will test both the completeness of the markup tagset and the utility of the data conversion software.

Biao Min is a Mienic language of southern China from the Hmong-Mien family. It is spoken by a rapidly shrinking population of approximately 21,000 speakers, most of whom are bilingual with the dominant language. It is an isolating language, and the morphology is minimal, consisting only of a dual/plural distinction in personal pronouns, complex serial verb constructions, and a set of numeral classifiers. Its phonology, however, is very complex: there are 6 distinctive tones, which show contrasts with both pitch and phonation types. The data consist of lexical lists and texts, as well as extensive annotations on cognates in other Hmong-Mien languages, and even on the Proto-Hmongic forms to which they correspond. Where a word is a loanword from Chinese, the Chinese character is also provided, thus allowing us to test the database’s multinational character display. Apart from passing mentions in Chinese, there is only 1 extant descriptive article about Biao Min, written by our data-provider, Prof. David Solnit of the Sino-Tibetan Etymological Dictionary and Thesaurus project at the University of California at Berkeley.

Mocovi is a Guaicuruan language from northern Argentina, with between 4000 and 7000 speakers. It is an active/inactive ergative language with complex person-marking, a very rich morphology and a rich deictic system. The Mocovi material is composed of texts interlinearized by hand, a dictionary of lexical items with a Spanish gloss, an English gloss, and a morpheme-by-morpheme breakdown and gloss. There are example sentences for each item, and comments by the researcher, Prof. Veronica Grondona, who has also agreed to be a data consultant on the E-MELD project. There are approximately 3000 lexical items in the database, and about 300 sentences in the collected texts. There are also 8 interlinearized texts, each taking about 5 minutes of recording time. Apart from Prof. Grondona’s unpublished dissertation, this work is the only description of Mocovi in existence.

Ega is the most western of the Kwa languages, and an isolate in the Nyo cluster. Ega is spoken by approximately 300 people in Ivory Coast, and is of particular scientific interest because its highly conservative features promise new insights into the history of the Niger-Congo languages. Funding for the documentation of Ega comes from the Volkswagen Foundation in Germany. The principal investigators, Dafydd Gibbon (Bielefeld), Bruce Connell (Oxford), and Firmin Ahoua (Cocody), have agreed to make their primary materials (including a lexicon, audio recordings, and interlinear texts) available to the LDC for conversion and archiving.

Cambap, known also as Twendi, is a Mambiloid language spoken in the Nigeria-Cameroon borderland by approximately 30 people; its youngest known speaker is about 45 years old. Cambap is the largest of 7 languages in the geographic region either on the brink of extinction or to have become extinct within the last few years.. The Cambap project, directed by Bruce Connell (Oxford), aims to document the language before it disappears. Data has been collected for a dictionary and a sketch grammar, as well as a range of sentences (based on Comrie and Smith 1987), a range of material illustrating phonemic contrasts and phonetic structures, and a number of proverbs and texts (folktales). All materials have been recorded digitally, using a total of 7 speakers.

Two of the language description projects will be performed by the Endangered Language Fund under the direction of its president, Douglas H. Whalen (of Haskins Laboratories, Yale U.). Although each language has unique structural characteristics, what is of special interest for the purposes of the present grant is the associated channels of information. The first, Tofa, will give us a test case of associated video, while the second, Lakota, will take advantage of the availability of a native speaker to acquire some physiological data that will be time-locked to the audio signal. Associating the audio signal of recorded speech with a video image is becoming increasingly common and important, both to provide more of the speaking context (e.g., manual and facial gestures) and to discriminate sounds (e.g., labial versus velar stops) that are easily confused with audio alone.

Tofa belongs to the northern (or north-eastern) branch of the Turkic family and to the Sayan areal group (Comrie 1981). It shows considerable Mongolian influence and substrate effects of an earlier, now extinct language, probably of the Yeniseyan family (Janhunen 1993). During an initial field visit in 1998, David Harrison collected new Tofa data and confirmed facts reported by Rassadin (1971, 1983), the only linguistic study of Tofa. The new material will put us on the path to an adequate grammar of Tofa, for which existing documentation is completely insufficient. Since the new documentation will include the video signal, we will be able to use this language as a test case for aligning video with the transcriptions and audio on the web-accessible database system being proposed.

Lakota is a Macro-Siouan language spoken in North Dakota primarily by the older members of the tribe. The ELF will use new utterances obtained from a speaker of Lakota to complement 8 hours of conversations already digitized. One aim will be to test the alignment of various ancillary signals with the primary audio signal.¹³ One atypical segment that Lakota has is a nasalized fricative. The ELF will examine the stability of the nasalization during changes in speaking rate by recording nasal air flow during utterances at self-selected normal, fast, and slow rates. An accelerometer, purchased with other funds, will be positioned at 1 nostril to record changes in air pressure during speech. The output of this device will be recorded simultaneously with the speech signal. This language, then, will provide us with both new data on an unusual phonetic structure and a test case for our ability to access two kinds of audio signal with our proposed database system.

The Summer Institute of Linguistics will provide data sets from 2 additional languages, as yet to be determined. SIL has access to data from an extremely wide range of languages, allowing us to choose ones with features not otherwise represented in our data. Likely candidates include:

Alamblak (1,500 speakers, Papua New Guinea, Sepik-Ramu family)

Dadibi (10,000 speakers, Papua New Guinea, Trans-New Guinea family)

Mapos Buang (7,000 speakers, Papua New Guinea, Austronesian family)

Tagakaulu Kalagan (37,000 speakers, Philippines, Austronesian family)

Tuwali Ifugao (25,000 speakers, Philippines, Austronesian family)

Each of these 5 languages has a lexicon and a text corpus that have been built using the LinguaLinks computing environment, the platform that will offer the most leverage for conversion into the markup we settle on.

3.5.2 THE QUERY ROOM

To enhance the "Best Practice Showroom," the ELF will design and implement a "Language Query Room"—essentially a chat room in which linguists can interact with native speakers and ask for utterances in a specific EL. The Query Room will not only allow practical extension of previous fieldwork, but also involve interested EL speakers in the language description enterprise. We will request such speakers to register, so that they can be sent email notification if a request relevant to their language has been submitted. However, responses will not be restricted to registered speakers, and multiple responses will be encouraged, since these will facilitate

¹³ Such signals are often the output of other recording devices, such as a nasal air pressure monitor or an electro-glottogram (EGG), but they can also be signals that are derived from the audio signal, such as fundamental frequency (F0) calculations or formant tracks.

comparison across speakers and dialects. Although most responses will be written, the submission of sound files will be encouraged. And the responses will be preserved with the aim of ultimately adding them to the data set for the language in the primary archive.

There are 3 main benefits to the Language Query Room. First, returned field researchers will be able to perform follow-up work without necessarily having to go back to the field. This will greatly extend the influence of the grant money that supports such field work. Second, researchers for whom field work is impractical can research their ideas by posting queries. A typologist who becomes curious about a formation in an EL can satisfy such curiosity without making the commitment that field work requires. While such an approach certainly can not substitute for an intimate knowledge of the language, it will provide improved access to primary data and should aid in constructing accountable theories. Third, native speakers of ELs will have an opportunity to spend a few minutes, at their own convenience, increasing the documentation of their language. Although many speakers of ELs do not, of course, have access to the Internet, many others now reside in urban areas, far from their origins and from other members of their original speech communities. These have easier access to computers than do other native speakers. The Query Room will be one place where their competence in their native language will be valued and where they will have the opportunity to contribute to a permanent record.

3.6 WORKSHOPS & INSTITUTES

During the project period, we will hold summer workshops and “Digital Institutes” in order to ensure that an influential core of linguists has the opportunity to learn about and provide feedback on the metadata and markup recommendations. To solicit input on markup and metadata standards, we will hold 2 weekend workshops for archivists and field linguists (in addition to one workshop which will take place before the project period—see 3.7.1 below). We will also hold 3 ‘digital institutes’ during the project period. These will be longer meetings designed to provide field workers with software to facilitate best practice, to train them in its use, and to elicit their feedback regarding its helpfulness with their own data. Should E-MELD secure NSF support, it may be possible—and advisable—to encourage NSF-funded field workers to participate in these workshops, since considerable progress toward promulgating best practice could be made in this way.

We feel it is important that native speakers of some of the ELs be included in these meetings. The ELF, which will serve as our primary liaison to indigenous communities, has already identified several potential participants, including 4 speakers of endangered languages with advanced degrees in linguistics, and 2 others who are more directly involved in language revitalization.

3.7 PLAN OF WORK

3.7.1 PROJECT PRELIMINARIES

As emphasized above, none of these initiatives will be undertaken without extensive consultation with the linguistics community. To that end, we have already identified a group of 25 field linguists, archivists, and computational linguists who have expressed their support for the project and their willingness to act as informal advisors. They are listed in Supplementary Documents, Part 3. A first meeting of this group—in the form of a workshop on markup and metadata—will be held at the Summer 2001 LSA Institute. (Funding for this workshop has been requested from the NSF Linguistics Division.) We have also already purchased and installed database and web design software critical to the establishment of the metadata server, the EL database, and the design of the web interfaces (see Facilities, 3.7.4 below).

3.7.2 TIMELINE

Figure 1 below provides a schematic outline of the steps in the 5-year project. Since a good deal of preparation has already been done, by the end of Year 3, we expect to have formulated the recommendations of best practice, converted data from 8 ELs, and established the metadata server. Years 4 and 5 will be devoted primarily to promulgating and evaluating the recommendations.

	WSU	EMU	UofA	ELF	LDC	SIL
Year 0 2000-2001	Data input & preparation: Biao Min & Mocovi, Samples submitted to LDC Corpus Cookbook	Database Installation & configuration for LINGUIST data & Ethnologue interface			Corpus Cookbook	Fieldwork Software, Consulting on LINGUIST database
Workshop on metadata & markup (LSA 2001 Summer Institute)						
Year 1 2001-2002	Typology questionnaire	Metadata format proposal	Markup Cycle 1: lexical entries & interlinearized text	Data preparation & conversion: Tofa & Lakota lexicons	Data preparation & conversion: Ega & Cambap lexicons	Data preparation & conversion: Austronesian & New Guinean Langs
	Data conversion: Biao Min & Mocovi lexicons	Metadata server: configure database for metadata & typology info				
Workshop I: testing / feedback on metadata & markup proposals						
Year 2 2002-2003	Data conversion: Biao Min & Mocovi texts	Metadata server: web input & query interfaces	Markup Cycle 2: Revise lexical markup	Data preparation & conversion: Tofa & Lakota texts	Data preparation & conversion: Ega & Cambap texts	Data preparation & conversion.
	Typology info collection	Metadata collection			Conversion software	Software for Field Work
Workshop II: testing / feedback on metadata & markup proposals						
Year 3 2003-2004	Data preparation & conversion: 9 th language	Data preparation & conversion: 10 th language	Markup Cycle 3: Revise text markup	Query Room		
"Digital Institute I" for field linguists: training in & distribution / testing of software						
Year 4 2004-2005	Refine user-interfaces for "Showroom" to insure flexible data display		Markup Cycle 4: Add paradigms?			
	"Digital Institute II" for field linguists: training in & distribution / testing of software					
Year 5 2005-2006	"Digital Institute III" for field linguists: training in & distribution / testing of software					

WSU: Wayne State University
 EMU: Eastern Michigan University
 UofA: University of Arizona

ELF: Endangered Languages Fund (Yale/Haskins Labs)
 LDC: Linguistic Data Consortium, University of Pennsylvania
 SIL: Summer Institute of Linguistics, University of Texas at Arlington

Figure 1: Timeline

3.7.3 PARTICIPANTS

The project work will be distributed among the 6 institutions, building on work already underway at each. A strength of the project is that 5 of the 6 institutions have well-established working relationships. Only the Endangered Languages Fund is a new LINGUIST collaborator. Eastern Michigan University and Wayne State University jointly host The LINGUIST List; the U. of Arizona employs the third LINGUIST moderator, the LDC formerly housed one of the LINGUIST machines and its Associate Director has been a consultant on former projects, as has Gary Simons of the Summer Institute of Linguistics. Furthermore the project design is such as to take advantage of the various institutions' strengths.

Eastern Michigan University, which houses the LINGUIST servers and provides a sysop and technician, will oversee the establishment of the metadata server and the "showroom of best practice," configuring the databases involved and creating appropriate web interfaces to each. Wayne State University, where the project members include Anthony Aristar, a typologist, and Martha Ratliff, former Head of the LSA Committee on Endangered Languages and their Preservation, will have primary responsibility for creating the typology questionnaire and testing the markup recommendations (using Biao Min and Mocovi). The University of Arizona will be in charge of markup design. The project will be jointly led by Gary Simons and by Terry Langendoen, former Chair of the Text Encoding Initiative Analysis and Interpretation Committee and, later, of the TEI Technical Review Committee. The team will continue the work of the TEI Workgroup on Linguistic Description, which was also jointly led by Simons and Langendoen.

In addition to providing and converting Ega and Cambap data, The Linguistic Data Consortium will design the conversion software to be featured in the Showroom. The LDC has been involved in corpus design since its inception in the early 90's; and PI Steven Bird, Associate Director of the LDC, is currently involved in 3 language engineering initiatives (ISLE, TalkBank, and the Linguistic Exploration Project) designed to result in a "Corpus Cookbook" of standards for formatting and storage of text.

The Summer Institute of Linguistics, although not a formal subcontractor on this grant, will provide a 10-year license and an interface to its Ethnologue, the primary language classification resource on the Internet. SIL, which developed Shoebox, currently the most widely-used linguistic field tool, is now in the process of developing Fieldworks. This language analysis software will be adapted to the needs of the E-MELD project by a workgroup also directed by Simons, SIL Associate Vice President and the primary technical consultant to the E-MELD project.

The Endangered Languages Fund, under the direction of its president Doug Whalen (of Yale and Haskins Laboratories) will be a liaison with indigenous communities, as well as a data provider and the designer of the Query Room facility.

Simons and Whalen, as directors of data conversion teams, will support the project both in a consulting and an administrative role. In addition, John Ockerbloom of the U. of Pennsylvania Digital Libraries Initiative will function as a consultant on metadata. As consultants on data preparation, we are fortunate to have David Harrison, who collected the Tofa data, and Veronica Grondona, whose grammar of Mocovi was a runner-up for the 1999 Mary Haas award. (See attached vitae and Budget Justification).

3.7.4 FACILITIES

LINGUIST has collected a substantial set of resources, which can be used in part for this project. Our collection of dedicated resources includes:

- One Sun Enterprise 450, with 2 Ultrasparc-II 400Mhz processors, 512 Mg Memory, and 3 9.1 GB drives. More disk space is now being added to this machine.
- One Dec Server 10, with 1 EV6 466MHz processor, 128 Mg of memory, and 3 9 GB drives.
- One Sun Ultra 1 Sparc Station, with 1 Ultrasparc-I 143MHz processor, 1 2.1 GB hard drive, and 2 9.1 GB hard drives, and a 7/14 GB 8mm tape drive.
- Two Model 440 Sun Workstations, with 440-MHz UltraSPARC-IIi processors, 1GB of memory each, 1 9.1 GB hard drives each, and 19" Monitors are also now on order, as well as 3 Windows workstations to add to the 2 we already possess.

Our machine resources are adequate for this project, and we already have, in addition to standard office software, two pieces of commercial software necessary for the E-MELD project: we have licenses to run Oracle on two machines, and we own a license for ColdFusion web application server.

4. WHY LINGUIST?

4.1 VISIBILITY

The LINGUIST List currently consists of a 12,500-member subscriber list and a website which is the primary international repository of electronic information relating to language and linguistics. The LINGUIST website is widely acknowledged as the most important source of linguistic information on the Internet (see, for example, recent articles in *Glott* and mentions in *Der Spiegel* and *The Wall Street Journal*); it organizes the Internet's largest collection of links to language-related electronic resources; and it is mirrored in its entirety by the U. of Stockholm, the U. of Edinburgh, the U. of Tübingen, and Moscow State University. LINGUIST has achieved previous recognition as an innovative application of Internet technology deployed in support of academic linguistic research, including 2 awards from the Linguistic Association of America and 4 grants from the National Science Foundation.

4.2 TRACK RECORD

LINGUIST has consistently fulfilled the commitments made in its grant applications. In 1993 LINGUIST received a \$4000 grant (NSF grant SBR-9311748) to design the custom software which allows our graduate student editors to generate and mail out LINGUIST issues. Partly as a result of this software, the LINGUIST email list has grown from 2,500 in 1993 to 12,500 in 2000; our graduate student staff now compose issues which generate approximately 67,000 email messages per day.

In 1996, LINGUIST received a grant from NSF for \$114,000 (SBR-9601352), with which we produced a greatly enhanced web site and search facilities. A partial list of the facilities added includes:

- An archive of all LINGUIST issues, searchable by topic, e.g., book announcements
- A searchable archive of dissertation abstracts in linguistics
- A searchable directory of professional information about individual linguists
- A searchable directory of information about programs in linguistics
- An email address-finder which is updated daily from our 12,500-member email list
- An Internet “Noticeboard” where subscribers can post individual announcements
- A special outreach web-site called “Ask-A-Linguist” where members of the public can ask linguistic questions of a panel of highly-qualified professional linguists

In part because of these innovations, our primary website—exclusive of our 4 mirror sites—now sustains over 300,000 page views per week.

In 1999 LINGUIST was awarded an NSF grant for \$173,000 (SBR-9975299, “The LINGUIST Multi-List Support Project”) to create a searchable archive of the postings of other linguistics-related mailing lists on the Internet. Though the project still has 12 months to go, we already have 70 other lists on our site (see: <http://linguistlist.org/list-archives.html>), and the multiple-list search facility is now being tested. By the project completion date, we expect to have approximately 100 lists archived on our site or distributed from our machines. And our access to these lists will be an inestimable advantage in our attempt to involve the whole linguistics community in the standards-setting enterprise. Similarly, our most recent NSF grant, a \$50,000 SGER grant awarded in June, 2000 (BCS-0003197) has allowed us to make a start on the E-MELD initiative by purchasing and installing database software (see 3.7.1 above).

4.3 SUPPORT & SUSTAINABILITY

LINGUIST is one of the oldest of linguistic organizations on the Internet. It has been in existence since 1989, and has developed a stable base of support within the discipline. Almost half of its funding derives from voluntary subscriber donations, including regular donations from organizations as diverse as The Linguistic Society of America, The Linguistic Society of Great Britain, The Linguistic Association of Finland, and The Linguistic Society of South Africa. The full LINGUIST site is mirrored once a day to the 4 other LINGUIST sites. Thus at least 5 full copies of our data and metadata, each less than 24 hours old, will exist on machines in widely separated parts of the world. Not only will these mirrors allow speedy, distributed access to the data, but they will help to preserve it from accidental destruction. In addition, we will make CD pressings of the primary data at 2 year intervals and make the CD’s available upon request to other organizations. (Funding for this effort will be requested from other sources.) The LINGUIST List is thus well-situated to organize the E-MELD enterprise, because of its size and number of distribution sites, its international character, and its high profile within the discipline.

5. PROJECT BENEFITS

There are numerous direct benefits to science which will derive from the creation of well-designed EL archives. For example, future scholars will have direct access to enormous numbers of fieldwork notes and recordings; they will not have to rely on data as presented in third party discussions. Cross-linguistic hypotheses can be pursued. Linguistic features can be statistically analyzed, studied in context, and plotted on maps. Perhaps most importantly, data will be preserved for whatever use future scholars may have. There are also important educational benefits. Members of small communities who are losing their ancestral language, or indeed have already lost it, can use such an archive for purposes of study or revitalization efforts.

However, to ensure that we reap these benefits from electronic archives of endangered languages, many guidelines for best practice must be developed, publicized, and adhered to. The E-MELD project will constitute a major step toward reaching disciplinary consensus about the markup, metadata, and software standards necessary to create adequate archive infrastructure.