The E-MELD Project:

School of Best Practices in Digital Language Documentation

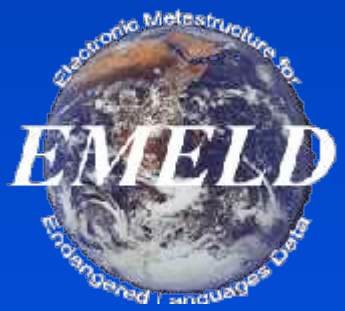
Helen Aristar Dry
The LINGUIST List
Eastern Michigan University

Jan 7, 2005

Linguistic Society of America
2005 Annual Meeting, Oakland, CA

The E-MELD Project

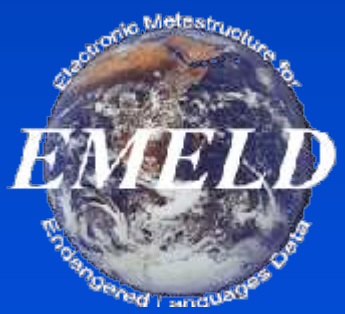
- **Electronic Metastructure for Endangered Languages Documentation**
- 5-year, NSF-sponsored project, begun Sept 2001
- Original Participants:
  - The LINGUIST List
    - Eastern Michigan U
    - Wayne State U
    - U of Arizona
  - Linguistic Data Consortium (UPenn)
  - Endangered Languages Fund



E-MELD Objectives:

To aid in …

- …the preservation of endangered languages documentation
- …fostering community consensus about best practices in the digitization of language data

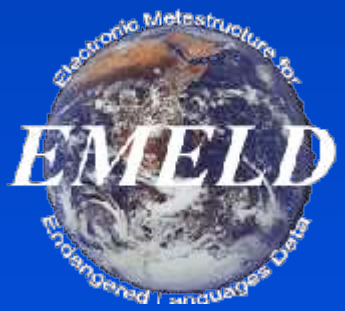


What are Best Practices?

Practices designed to insure that digital language resources :

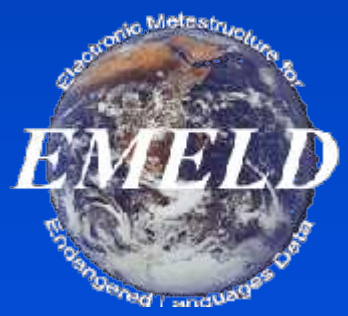
- endure through time.
- can be reused by others, both now and in the future.

-Bird & Simons 2003



Why Best Practices?

The impending "Digital Dark Age"

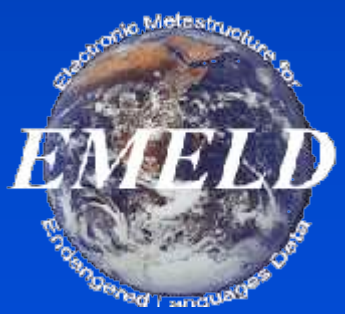


Jan 7, 2005

Linguistic Society of America
2005 Annual Meeting, Oakland, CA

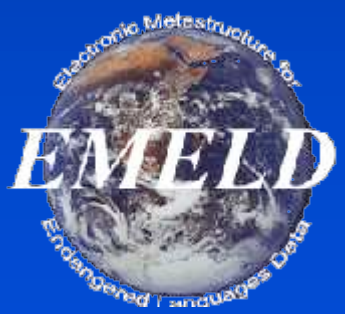
An impending "Digital Dark Age"

Future historians may see our present age as another Dark Ages since so much information documenting our current civilization is recorded digitally and may have vanished.



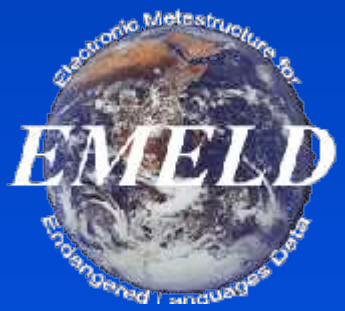
A paradox of writing history
(fr. Gary Simons, LSA 2004)

- The more advanced the writing technology, the less durable the written product.
- From most durable to least durable:
  - Clay tablets and stone
  - Velum
  - Papyrus
  - Paper
  - Digital word processing



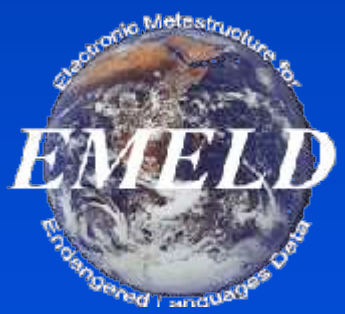
Hardware devices are ephemeral
(fr. Gary Simons, LSA 2004)

- Removable media on personal computers advance over 25 years:
  - 8-inch floppies
  - 5.25-inch floppies
  - 3.5-inch floppies
  - Zip drives
  - CD-Rs
  - DVD-Rs
  - Memory sticks?



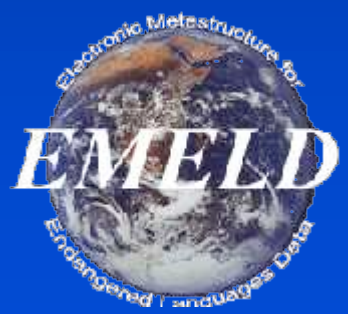
Software formats are ephemeral
(fr. Gary Simons, LSA 2004)

- Software vendors change file formats and functionality with each version.
- When we use a proprietary single vendor format, we lose access to the data when the software is obsolete.
- For instance,
  - Microsoft Word files from the 1980s cannot be read by current versions of Word



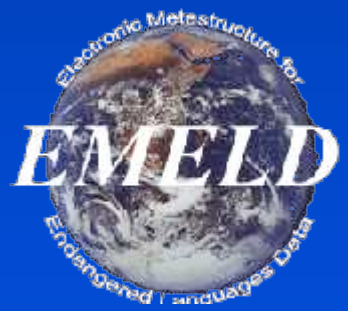
Goal: School of Best Practices

- To encourage linguists to think of themselves as creating archive-ready documentation for the benefit of future generations
- To facilitate this undertaking by providing information, models, tools and support



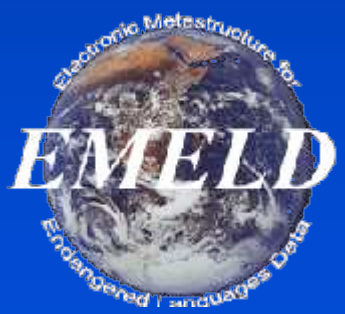
Some Best Practices

- Distinguish between
  - archival form: The form in which information is stored for access long into the future.
  - working form: The form in which information is stored as it is created and edited
  - presentation form: The form in which information is presented to the public.
- Recommendations primarily concern archival form.



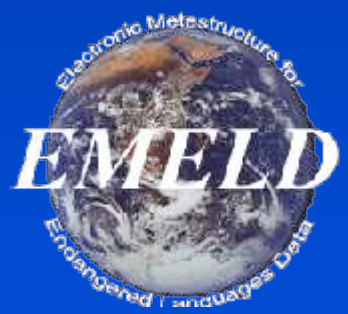
Some Best Practices

- Employ file formats that offer LOTS;
  - L = Lossless
  - O = Open (standards and formats)
  - T = Transparent (or at least well-documented)
  - S = Supported by multiple vendors
- Ex: For text files: plain text with XML markup



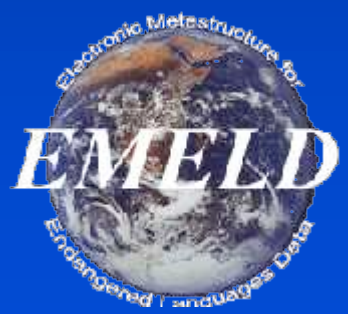
Some Best Practices (cont.)

- For character encoding, use Unicode
- For language identification, use Ethnologue / OLAC language codes
- Create metadata in a standard format (e.g., OLAC or IMDI) and make it available to a search engine
- Deposit archival copies in an established archive



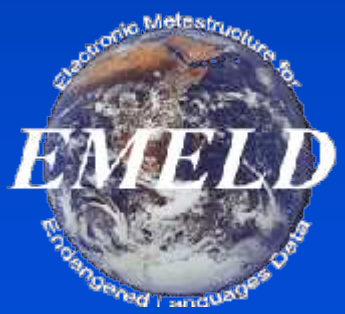
Organization of the School

- Entrance Hall: orientation
- Classroom: lessons & tutorials
- Reading Room: bibliography
- Work Room: online work
- Tool Room: links to tools
- Help (incl. Ask an Expert)
- Case Studies: documentation of 10 ELs digitized according to best practices



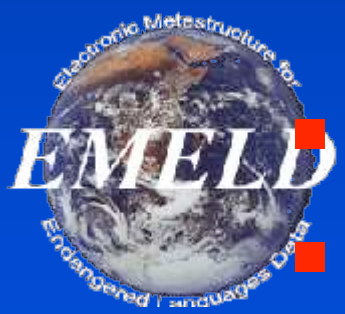
Case Studies (to date):

- Documentation from 8 ELs:
  - Mocovi
  - Monguor
  - Tofa
  - Saliba
  - Biao Mien
  - Kayardild
  - Potawatomi
  - Ega
- Also W. Sissala, Chorote, Nivacle



Developed by:

- E-MELD Project Participants
- The LINGUIST List Crews (2001-4)
  - Team Leader: Steve Moran
- E-MELD Data Providers: Harrison, Buszard-Welcher, Solnit, Grondona, Dwyer . . .
- Consultants: Simons, Hughes, etc.
- 2001-2004 Workshop Participants

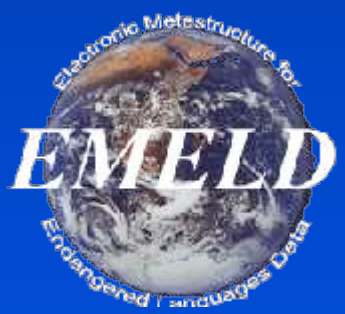


E-MELD Workshops

- 2001, Santa Barbara, CA: The Need for Standards
- E-MELD 2002, Ann Arbor, MI: Digitizing Lexical Information
- E-MELD 2003, Lansing, MI: Digitizing Texts
- E-MELD 2004, Detroit, MI: Databases and Best Practice
- E-MELD 2005, Ann Arbor, MI: Linguistic Annotation: Ontologies & Terminology

E-MELD School of
Best Practices in Digital Language Documentation

http://emeld.org/school/



Jan 7, 2005

Linguistic Society of America
2005 Annual Meeting, Oakland, CA