

E-MELD

Electronic Metastructure for Endangered Languages Documentation

- 5 year NSF-sponsored project
- *Goal:* To aid in the preservation of endangered languages data, and the development of infrastructure for electronic archives
- *Participants:* LINGUIST List (Eastern Michigan U., Wayne State U.) U. of Arizona), Linguistic Data Consortium, Endangered Languages Fund)



Preservation, Intelligibility, Interoperability

The E-MELD Vision of Digital Language Resources

Gary Simons, SIL International

Helen Aristar Dry, Eastern Michigan U.



Preservation, Intelligibility, Interoperability

- Part 1: Toward Enduring Resources (Dry)
- Part 2: Toward Interoperable Resources (Simons)
- We also provide

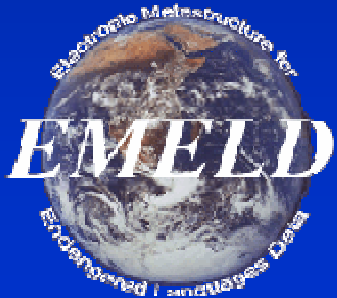
Attention AIDS:

Acronyms **I**n **D**ubious **S**hapes



Source of E-MELD Recommendations

- Working groups of language engineers and documentary linguists
- At 5 E-MELD workshops:
 - 2001: The Need for Standards
 - 2002: Lexicons
 - 2003: Texts
 - 2004: Databases
 - 2005: Ontologies
 - **2006: Tools and Standards**



E-MELD Vision of Digital Language Resources

- **Preservable:** formats are not vulnerable to physical decay or obsolescence of hardware & software
- **Intelligible:** content is easily understood by future scholars
 - “We don’t want to create another Rosetta Stone” (Whalen, 2003)
- **Accessible:** distributed resources are easily discovered and accessed
- **Interoperable:** documentation created by different scholars is easily searched, compared, and reused.



E-MELD

Recommendations of Best Practice

Practices are	if they ensure:
Good	Preservation
	Intelligibility
Better	Access
Best	Interoperability

Responsibility shared between **linguist**,
archive, & **service**.

Practices for Individual Linguists

GOOD	Preservation	Put the resource in an enduring file format
	Intelligibility	Document the content
BETTER	Access	Create an archive-ready collection and deposit it with an archive
BEST	Interoperability	Format to facilitate automatic processing

Good practice for the Linguist: Preservation of the format

An enduring file format is one that offers **LOTS**:

- **L**ossless
- **O**pen
- **T**ransparent
- **S**upported by multiple vendors

(Gary Simons, LSA 2004)



Lossless

- No content should be lost through compression
- Uncompressed file formats (lossless):
 - Audio: .wav, .aiff, .au (pcm)
 - Images: .tiff, .bmp
 - Video: .avi (depends on codec), rtv
 - Text: .txt, html, xml
- Compressed but lossless:
 - Audio: .ale (Apple Lossless Encoding)
 - Images: .gif (black & white only)
 - Video: jpeg2000 (new - 1:10 ratio)
 - Text: .zip



OPEN

- Prefer a file format whose specification is publicly available, i.e., “Open standard.”
 - Exs: html, XML, pdf, rtf
- Information in proprietary file formats will be lost when the vender ceases to support the software



Transparent

- Format requires no special knowledge or algorithm to interpret
- One-to-one correspondence between the numerical values and the information they represent, e.g.
 - **Plain text:** one-to-one correspondence between **numbers & characters**. Can be read by any program that handles text.
 - **PCM codec (.wav, .aiff, cdda):** One-to-one correspondence between **the numbers & the amplitudes of the sound wave**. Can be read by any program that handles audio.



Support by multiple vendors

- Makes a file format less likely to fall victim to hardware and software obsolescence.
- Is encouraged by use of open standards:
 - If a file format is open, anyone can create programs that handle it
 - Not necessary to reverse engineer the format or purchase the specification from the developer
 - So program development is less costly



Intelligibility : Preserving the **Content**

- So longterm preservation of the file format requires **LOTS**.
- But, for longterm intelligibility, the linguist must do even **MORE:**
- **Document** the:
 - **M**arkup
 - **O**ccasion
 - **R**ubrics
 - **E**ncodings



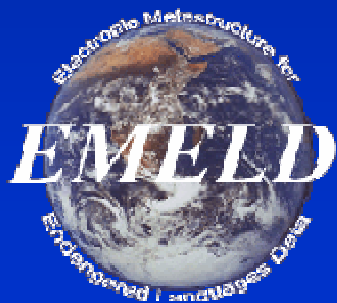
Intelligibility: Document the Markup

- Document all **markup**, whether
 - **Presentational**: make explicit the information encoded in the formatting
 - **Bolding** indicates “headword”
 - **Punctuational**:
 - “A **semi-colon** separates the different senses of a word”
 - **Descriptive**:
 - “**<pos>** stands for ‘part of speech’”



Intelligibility: Document the **Occasion**

- Record the
 - Time & place
 - Type of speech event
 - Participants
 - Language(s)
- Best practice: use OLAC or IMDI metadata standard for interoperability
- OLAC = Open Language Archives Community: 15 element metadata standard developed for language resources



Intelligibility: Document the **Rubrics**

- **Abbreviations:** list every abbreviation and what it stands for
- **Terminology:** define the concepts used in the language description
 - “Absolute” refers to an unpossessed noun in Uto-Aztecan.
- **Glossing rules:**
 - “A tilde represents reduplication”



Intelligibility:

Document the **Encoding**

■ **Encoding:**

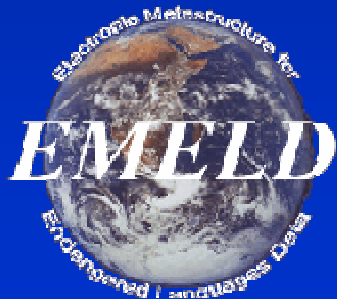
- Identify the base character set
 - Example: ISO 8859-1, CJK
- Document every non-standard character used
- Or use **Unicode** (recommended)
 - Unambiguous standard
 - Promotes interoperability
- With Unicode, document every character placed in the Private Use Area.



Better Practice:

Promote Discovery & Access

- Deposit the resource in an archive
- A file with **LOTS MORE** should be stored in an archive that offers **MUCH:**
 - **M**igration
 - **U**ser access
 - **C**ataloging
 - **H**arboring



Archive Recommendations:

Offer **MUCH**:

- **M**igration to new storage media and formats as technologies change
- **U**ser access within the bounds of IPR. Digital archives should provide more than local access (e.g., URLs) even if not interoperable with other archives.
- **C**ataloging: resources organized, metadata made available
- **H**arboring: resources conserved in a safe environment



Scale of Practices for Archives

GOOD	Preservation	If needed, transfer to a format with LOTS Migration to new media & file formats as technology changes
	Intelligibility	Retention of metadata & creation of metadata if missing
BETTER	Access	Public availability of metadata IPR agreements with time limits URL's for resources (also enables shallow interoperability)
BEST	Interoperability	On to Garys presentation....

E-MELD Vision of Digital Language Documentation

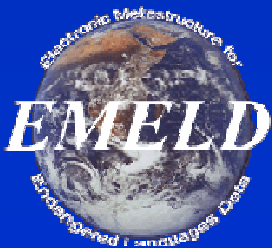
Part 2: Interoperability

Gary F. Simons
SIL International



E-MELD End Vision

- The digital products of the linguistics community's efforts to document endangered languages:
 - Will endure far into the future
 - Will be found and used by any who have an interest in the documented languages
 - Will enable our knowledge about the world's languages to be combined and searched to an unprecedented degree



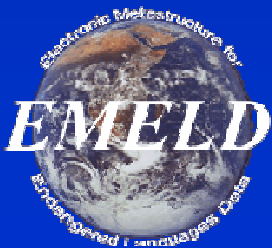
The interoperation problem

- Once the resources that linguists create are being preserved for the future in a host of archives:
 - How can potential users ever find the resources they are interested in?
 - How can users search the combined work of different linguists, especially when they have used different markup or terminology?
- Solutions require archives and resources to interoperate.



Services to the rescue

- The user can't solve these problems—there are too many archives to visit.
- An archive can't solve these problems—all the other archives have to be included.
- A service can solve the problems—
 - An automated system that supports inter-operation among all participating archives.
 - Provides a single point of entry for users.
 - Developed and maintained by an institution.

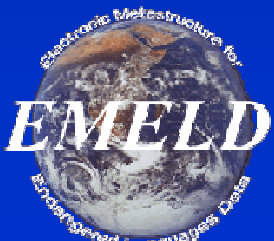
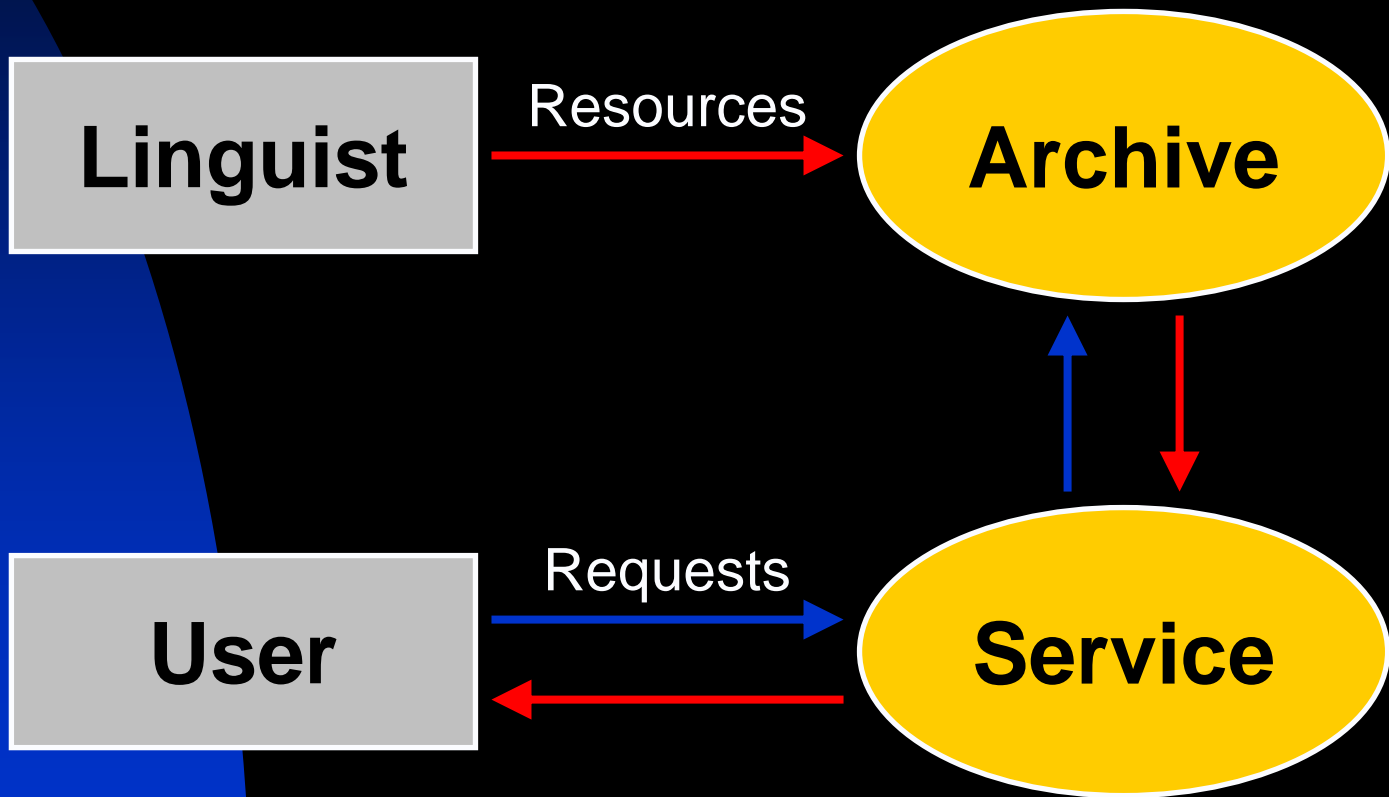


The key players

User	A person who wants to use language resources
Linguist	A person who creates language resources
Archive	An institution that curates language resources
Service	An institution that makes language resources interoperate



The big picture



June 20, 2006

E-MELD 2006, Lansing, MI

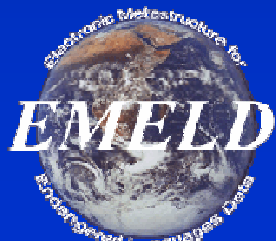
Two kinds of interoperation

- Shallow interoperation
 - Based on the surface content of plain text
 - Generic to all problem domains
 - Based on the ubiquitous HTTP infrastructure
- Deep interoperation
 - Based on underlying concepts and structures
 - Built for a specific problem domain
 - Requires a domain-specific infrastructure (e.g. protocols, markup, controlled vocabularies)



Supporting shallow interoperation

- Such services already exist: e.g., Google
- If an archive exposes its catalog as web pages, it will have shallow interoperation at the level of metadata.
- If an archive provides web links to resource content, it will have shallow interoperation at the level of data content.
- Easy for the archive to do and easy for the user to use.



So what's the problem?

- Lots of noise (= Low precision)
 - The words used to formulate the query have many irrelevant senses. E.g.
 - Ega is the name of a language
 - It is also an acronym with unrelated meaning
- Lots of drop out (= Low recall)
 - The target concept may be in the text as a word different from the one in the query. E.g.
 - Synonyms; Alternate names



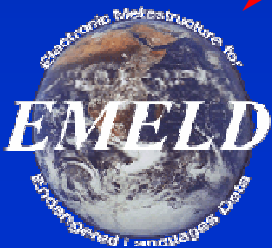
An example of shallow search

- Using Google to look for an Ega dictionary
- Try: Ega dictionary (120,000 hits)
 - Enhanced Graphics Adapter, Enterprise Grid Alliance
 - 19: E-MELD School of Best Practice: Ega Lexicon
 - 92: Endangered Language Foundation
- Try: Ega lexicon (24,500 hits)
 - 1: E-MELD School of Best Practice: Ega Lexicon
 - 2: Ega Web Archive (at Bielefeld)
 - Next 98 hits include 4 that refer to the language



An example of deep search

- Using OLAC to look for an Ega dictionary
 - Open Language Archives Community
 - Uses controlled vocabulary to identify language
 - Uses controlled vocabulary for linguistic types
- Language code='ega' and Type='lexicon' (6 hits)
 - All are relevant items from U Bielefeld Language Archive
 - Typescript, recording and transcripts of word lists
 - Data files: Shoebox, XML, CSV



Evaluation scale:

Levels of practice for archives

- Bad: Does not do MUCH
- Good: Does do MUCH
- Better: And supports shallow interoperation
 - To increase recall in generic services
- Best: And supports deep interoperation
 - To increase precision via domain services



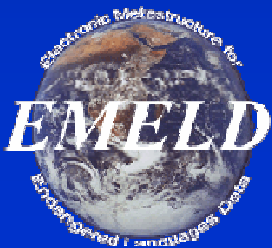
Supporting deep interoperation

- An archive supports deep interoperation if:
 - Its resources use XML markup so that machines may interpret their contents
 - The XML encoding uses domain-specific controlled vocabularies
 - It implements the protocol of a domain-specific service so that the service can access its deep resources



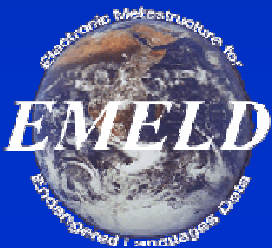
Dimensions of service

- For all services:
 - Closed vs. Open
 - Generic vs. Domain specific
- Further dimensions for domain-specific services:
 - Metadata vs. Full content
 - Precision-supplied vs. Precision-added



Good and Better in services

- The second is better than the first:
 - Closed vs. Open
 - Only people inside the service know how to place new resources into the service., vs.
 - The specifications for entering the service are published and people outside the service can meet those specs.
 - Generic vs. Domain specific
 - Supports domain-neutral shallow interoperation, vs.
 - Supports domain-specific deep interoperation.
- Examples
 - Google: Open and Generic
 - Typology projects: Closed and Domain-specific



Dimensions of the Best

- Services that are Open + Domain-specific vary in:
 - Scope
 - The service operates over metadata, vs.
 - The service operates over a focused aspect of full content.
 - Source of precision
 - The depth is encoded in the form provided by archives, vs.
 - The depth is mined from shallow resources.
- Examples
 1. OLAC: Metadata and Precision-supplied
 2. Metaschema experiments: Data and Precision-supplied
 3. ODIN: Data and Precision-added



1. Open Language Archives Community

- An open standard for metadata and protocol for harvesting: www.language-archives.org
- 34 institutions now participate by contributing to a pooled catalog of language resources
- As part of E-MELD, Linguist List has developed a search service over that catalog:

<http://www.LinguistList.org/olac/>



What the archive supplies

```
- <olac:olac xsi:schemaLocation="http://www.language-archives.org/OLAC/1.0/
http://www.language-archives.org/OLAC/1.0/olac.xsd
http://purl.org/dc/elements/1.1/
http://www.language-archives.org/OLAC/1.0/dc.xsd http://purl.org/dc/terms/
http://www.language-archives.org/OLAC/1.0/dcterms.xsd">
  <title>Ega lexicon (Gbery)</title>
  <creator>Gbery, Eddy Aime</creator>
  <creator>Baze, Lucien</creator>
  <subject xsi:type="olac:language" olac:code="ega"/>
  <description>Ega lexicon in Shoebox format</description>
  <publisher>unpublished</publisher>
  <contributor>Lindenlaub, Juliane</contributor>
  <date>2003-03</date>
  <type xsi:type="olac:linguistic-type" olac:code="lexicon"/>
  <format>shoebox</format>
  <language xsi:type="olac:language" olac:code="fra"/>
  <language xsi:type="olac:language" olac:code="ega"/>
  <language xsi:type="olac:language" olac:code="eng"/>
  <language xsi:type="olac:language" olac:code="deu"/>
  <coverage>Cote d'Ivoire</coverage>
</olac:olac>
```

What the service reports



Eastern Michigan University • Wayne State University

People & Organizations ♦ Jobs ♦ Calls & Conferences ♦ Publications ♦ Language Resources ♦ Text & Computer Tools ♦ Teaching & Learning ♦ Mailing Lists ♦ Search

Document Information

General Description:

Title: Ega lexicon (Gbery)

Archive: U Bielefeld Language Archive

Archive URL: <http://www.spectrum.uni-bielefeld.de/langdoc/>

Creator(s): Gbery, Eddy Aime
Baze, Lucien

Description: Ega lexicon in Shoebox format

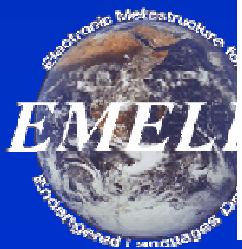
Contributor(s): Lindenlaub, Juliane

Date: 2003-03

Coverage: Cote d'Ivoire

Format: shoebox

Language: French [fra]
Ega [ega]
English [eng]



2. The metaschema experiments:

Based on E-MELD founding principles

- The inaugural EMELD workshop (2001) easily reached consensus on three points:
 - XML descriptive markup provides the best format for the interchange and archiving of endangered language data.
 - No single schema for XML markup can be imposed on all language resources.
 - Linguists need to be able to perform queries across multiple resources.



A fundamental problem

- How to interoperate across resources when:
 - Those resources use different markup schemas
 - The linguists have used different terminology in their analysis and description
- The EMELD solution is based on GOLD:
 - General Ontology for Linguistic Description
 - Use a shared ontology of linguistic concepts as the basis for interoperation across disparate markup and terminologies



Converting from Markup to Meaning

- markup schema
 - A formal definition (as with XML DTD or XML Schema) of the vocabulary and syntax of markup for a class of source documents.
- semantic schema
 - A formal definition (as with RDF Schema or OWL) of the concepts in a particular domain.
- metaschema
 - A formal definition of how the elements and attributes of a markup schema are interpreted in terms of the concepts of a semantic schema.



A sample Hopi lexical entry

```
<Lexeme id="L28">
  <Head><Headword>
    <OrthographicForm>na('at)</OrthographicForm>
  </Headword></Head>
  <POS>
    <Feature name="cat">n</Feature>
    <Feature name="type">poss</Feature>
  </POS>
  <Sense><Gloss>
    <OrthographicForm>father. The term is applied to
      one's natural father.</OrthographicForm>
  </Gloss></Sense>
</Lexeme>
```

A metaschema fragment

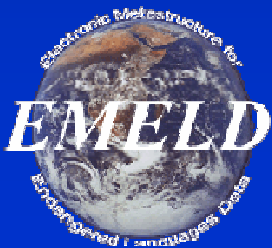
```
<interpret markup="Lexeme">
  <resource concept="gold:LinguisticSign"/>
</interpret>
<interpret markup="Head">
  <property concept="gold:form">
    <resource concept="gold:PhonologicalUnit"/>
  </property>
</interpret>
<interpret markup="OrthographicForm">
  <literal concept="gold:orthographicRepresentation"/>
</interpret>
```

The interoperable interpretation

```
<gold:LinguisticSign rdf:about="#element(L28)">
  <gold:form>
    <gold:PhonologicalUnit>
      <gold:orthographicRepresentation>na('at)</gold:orthographicRepresentation>
    </gold:PhonologicalUnit>
  </gold:form>
  <gold:meaning>
    <gold:SemanticUnit>
      <gold:definition>father. The term is applied to one's natural
        father,</gold:definition>
    </gold:SemanticUnit>
  </gold:meaning>
  <gold:grammar>
    <gold:GrammaticalUnit>
      <gold:hasPartOfSpeech rdf:resource="&gold;Noun" />
      <gold:hasFeature rdf:resource="&gold;InalienablyPossessed" />
    </gold:GrammaticalUnit>
  </gold:grammar>
</gold:LinguisticSign>
```

Best practice opens the playing field

- Linguist achieves best practice
 - Deposits resource in XML descriptive markup
- Archive achieves best practice
 - Supports access to that resource
- Service achieves best practice
 - Supports an open protocol on a focused data type
- Analyst can then bridge the interoperability gap
 - Analyst creates and archives a metaschema
 - Service harvests original resource + metaschema



Results to date

- Proof of concept on a small scale using Sesame (an open-source RDF database):
 1. Lexicons from 3 languages
 2. Interlinear texts from 7 languages
- See papers by Simons *et al.* at emeld.org
 - Project Documents
 - 2004 Workshop Proceedings
 - 2005 Workshop Proceedings



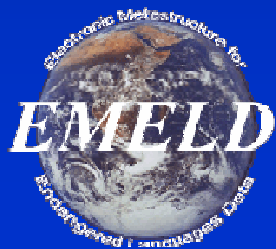
3. Mining the depths of shallow resources

- The service widely harvests shallow resources
 - E.g. through web crawling or Google API
 - Uses domain knowledge to add precision
- The service can serve at two levels:
 - Direct service to users who use it to access the harvested shallow resources
 - Indirect service through other services by implementing a best-practice (domain-specific) metadata provider



ODIN: Online Database of Interlinear Text

- See paper by Will Lewis at emeld.org
 - 2003 Workshop Proceedings
- Methodology
 - Seed Google search with abbreviations used in glossing
 - Keep URL if content has instances of text-gloss-translation
 - Use Ethnologue names data to propose language identify
- Service currently reports:
 - 33,713 instances of Interlinear Glossed Text examples
 - from 701 different languages
 - in 2,202 different linguistic documents



What the user sees

ODIN - The Online Database of Interlinear - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

http://www.csufresno.edu/odin/

Language name/

- [Aari \(AIZ\)](#)
- [Abkhaz \(ABK\)](#)
- [Abun \(KGR\)](#)
- [Aceh \(ATJ\)](#)
- [Adi \(ADI\)](#)
- [Adyghe \(ADY\)](#)
- [Afrikaans \(AFK\)](#)
- [Aghem \(AGQ\)](#)
- [Ainu \(AIN\)](#)
- [Akan \(TWS\)](#)
- [Akawaio \(ARB\)](#)
- [Alamblak \(AMP\)](#)
- [Albanian, Arvanit](#)
- [Albanian, Gheg \(A](#)

ODIN

The **O**nline **D**atabase of **I**nterlinear **T**ext

List of documents and pages with Interlinear examples for [Aceh \(ATJ\)](#)

(Alternate names and dialects for Aceh are Achehnese, Achinese, Atjeh, Atjehnese, Banda Aceh, Baruh, Bueng, Daja, Pase, Pedir, Pidie, Timu, and Tunong)

URL	#	Verified
http://eprints.unimelb.edu.au/archive/00000239/01/Musgrave.pdf	1	Highest
http://rspas.anu.edu.au/linguistics/iwa/Arka-Kosmas-final.pdf	2	High

Done

What another service sees

```
- <olac:olac>
  <dc:title>Interlinear Glossed Text for Aceh</dc:title>
  <dc:creator>Lewis, William</dc:creator>
  <dc:subject xsi:type="olac:language" olac:code="x-sil-ATJ">
    Aceh</dc:subject>
- <dc:description>
  A listing of Web resources containing Interlinear Glossed Text for
  the language Aceh: 2 document(s), 3 instance(s) of interlinear text.
</dc:description>
  <dc:publisher>California State University, Fresno, ODIN
  project</dc:publisher>
  <dc:date>2005-02-02</dc:date>
- <dc:identifier>
  http://www.csufresno.edu/odin/igt_urls.php?lang=ATJ
  </dc:identifier>
</olac:olac>
```



Services in a word

- Services give the linguist **POWER.**
- The best services offer:
 - **P**recision
 - **O**penness
 - **W**eb harvesting
 - **E**nrichment
 - **R**each



The elements of POWER

- **Precision**
 - Precision through domain-specific standards.
- **Openness**
 - Anyone can implement the supporting protocol.
- **Web harvesting**
 - Harvesting resources from around the Internet.
- **Enrichment**
 - Adding precision to resources born shallow.
- **Reach**
 - Searching resources from everywhere at once.



Conclusion: Toward best practice

- Digital language archiving holds the potential of unparalleled access to information, but only if:
 - **Linguists** do **LOTS MORE** to ensure that the resources they create endure far into the future.
 - **Archives** do **MUCH** to ensure the preservation of those resources.
 - **Services** give users **POWER** to retrieve everything that is relevant (and only what is relevant).
 - The linguistics community embraces the domain-specific standards that support interoperation.

