An e-Infrastructure for Language Documentation on the Web

Gary F. Simons, *SIL International*
William D. Lewis, *University of Washington*
Scott Farrar, *University of Arizona*
D. Terence Langendoen, *National Science Foundation*
Goals

- To provide a means by which the digital products of the linguistics community’s efforts to document all the world’s languages will:
  - Endure far into the future;
  - Be found and used by any who have an interest in those languages;
  - Be unified in such a way that knowledge about those languages can be made readily available.
The interoperation problem

- Once the resources that linguists create are being preserved for the future in a host of e-accessible archives:
  - How can users find the resources they are interested in?
  - How can users search the combined work of different researchers and projects, especially when they have used different markup or terminology?
The players

<table>
<thead>
<tr>
<th>User</th>
<th>A person who wants to use language resources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguist</td>
<td>A person who creates language resources</td>
</tr>
<tr>
<td>Archive</td>
<td>An institution that curates language resources</td>
</tr>
<tr>
<td>Service</td>
<td>An institution that enables language resource interoperation</td>
</tr>
</tbody>
</table>
A visualization

Linguist → Resources → Archive

User → Requests → Service
Shallow vs. deep 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
  - Based on 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.
Low precision and recall in shallow search

- Using Google to look for an Ega dictionary

- Ega dictionary (120,000 hits)
  - EGA is an acronym inter alia for Enhanced Graphics Adapter and Enterprise Grid Alliance.
  - Out of top 100 hits, only 2 are relevant:
    - #19: E-MELD School of Best Practice: Ega Lexicon
    - #92: Endangered Language Foundation

- 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

- The Open Language Archives Community (OLAC) uses controlled vocabulary to identify:
  - Language (ISO 639-3 three-letter codes);
  - Resource type.

- Language code=‘ega’ and Type=‘lexicon’ (6 hits)
  - All are relevant items from the University of Bielefeld Language Archive.
    - Includes typescripts, recording and transcripts of word lists
    - Also includes data files in various formats, e.g. Shoebox, XML, CSV
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

- **Closed vs. Open**
  - *Closed*: Only people inside the service know how to place new resources into the service.
  - *Open*: The specifications for entering the service are published and people outside the service can meet them.

- **Generic vs. Specific**
  - *Generic*: Supports domain-neutral shallow interoperation.
  - *Specific*: Supports domain-specific deep interoperation.

- **Examples**
  - Google: Open + Generic
  - Typical language typology projects: Closed + Specific
Further open + specific dimensions of service

- Metadata vs. Content
  - Metadata: The service operates over metadata only.
  - Content: The service operates over (aspects of) full content.

- Supplied vs. Added
  - Supplied: The depth is encoded in the form provided by archives.
  - Added: The depth is mined from shallow resources.

- Examples
  1. OLAC: Metadata + Supplied
  2. Metaschema experiments: Content + Supplied
  3. ODIN: Content + Added
Example 1. Metadata-enriched interoperation

- OLAC: Open Language Archives Community
  - An open standard for metadata and protocol for harvesting: [http://www.language-archives.org](http://www.language-archives.org)

- 34 institutions now participate by contributing to a pooled catalog of language resources.

- LINGUIST List has developed a search service over that catalog:
  - [http://linguistlist.org/olac/](http://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

Document Information

General Description:

Title: Lagos lexicon (Gbery)
Archive: U Bielefeld Language Archive
Archive URL: http://www.spectrum.uni-bielefeld.de/langdoc/
Creator(s): Gbery, Eddy Alme
Baze, Lucien
Description: Lagos lexicon in Shoebox format
Contributor(s): Lindenlaub, Juliane
Date: 2003-03
Coverage: Cote d’Ivoire
Format: shoebox
Language: French [fra]
Ega [ega]
English [eng]
German [deu]
Example 2. Content-supplied interoperation

- How do you interoperate across resources
  1. When those resources use different markup schemas?
  2. When linguists have used different terminologies in their analyses and descriptions?

- Both questions can be answered by providing a machine-readable semantics for XML syntax and (parts of) the content of resources.

- To this end, we’re developing two resources:
  - SIL (Semantic Interpretation Language)
  - GOLD (General Ontology for Linguistic Description)
    - [http://linguistics-ontology.org/](http://linguistics-ontology.org/)
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

```xml
<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>
Results to date

- Proof of concept on a small scale using Sesame, an open-source RDF database:
  1. Lexicons from 3 languages
  2. Interlinear glossed texts from 7 languages

- See papers by Simons et al. at http://emeld.org
Moving the solution out of the lab

- Analysts need to bridge the interoperability gap by creating and archiving metaschemas.
- Services can then harvest original resources + metaschemas and output interoperable resources that can be used for querying or further processing.
- Robust open RDF database technology is required.
Example 3. Content-added interoperation

- ODIN: Online Database of Interlinear Text
  - http://www.csufresno.edu/odin/

- Discussed in papers by Lewis at http://emeld.org/

- Methodology
  - Seed Google search with abbreviations used in glossing.
  - Keep URL if content has instances of text-gloss-translation.
  - Use ISO 639-3 language names to propose language identify.
  - Use GOLD to interpret selected glosses, and (English) translation to identify certain grammatical construction types (can be semi-automated).

- Service recently reported:
  - 33,713 instances of Interlinear Glossed Text examples,
  - from 701 different languages, and
  - in 2,202 different linguistic documents.
What the user sees

ODIN

The Online Database of Interlinear Text

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)

<table>
<thead>
<tr>
<th>URL</th>
<th>#</th>
<th>Verified</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://eprints.unielb.edu.au/archive/00000239/01/Musgrave.pdf">http://eprints.unielb.edu.au/archive/00000239/01/Musgrave.pdf</a></td>
<td>1</td>
<td>Highest</td>
</tr>
</tbody>
</table>
What another service sees

- `<olac:olac>
  <dc:title>Interlinear Glossed Text for Aceh</dc:title>
  <dc:creator>Lewis, William</dc:creator>
- `<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>
</dc:identifier>
</olac:olac>
Empowerment through services

- Precision
  - Through use of domain-specific standards.

- Openness
  - Anyone can implement the supporting protocol.

- Web harvesting
  - From resources on the Internet.

- Enrichment
  - Adding depth to shallow resources.

- Reach
  - Enabling search for resources from everywhere at once.