E-MELD Workshop on Morphosyntactic Annotation and Terminology:
Linguistic Ontologies and Data Categories for Linguistic Resources
Cambridge, MA, 1-3 July 2005

Beyond the Brink: Realizing Interoperation through an RDF Database

Gary F. Simons, SIL International

Introduction

The participants of the inaugural EMELD workshop (Linguist List 2001) easily reached consensus on three points:

- XML 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.

But herein lies a fundamental problem: How do we interoperate across resources when those resources use different markup schemas and the linguists have used different terminology in their analysis and description? At the heart of EMELD's solution to this problem lies GOLD—General Ontology for Linguistic Description (Lewis and others 2001, Farrar and others 2002a,b)—which EMELD hopes will one day be accepted as embodying a community-wide consensus on a shared ontology of linguistic concepts. This ontology provides the basis for interoperation across disparate markup and terminologies.

The preceding paper in this session, "Taking Resources to the Brink of Interoperation: Profiles, Termsets, and Best Practice Markup" by Will Lewis, describes how language resources are brought to the brink of interoperation by transforming them to XML markup and using termsets and language profiles to map the linguist's terminology onto the shared concepts of GOLD. This paper picks up the thread and demonstrates how those resources are taken over the brink to enable queries over once disparate resources.

RDF and the Semantic Web

One of the early decisions made by the developers of GOLD was to ground it in the Semantic Web (Farrar and Langendoen 2003). The Semantic Web is "an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation" (Berners-Lee and others 2001). It is being developed through a formally-organized Semantic Web Activity of the World Wide Web Consortium (W3C) [http://www.w3.org/2001/sw/]. It relies on the Resource Description Framework (RDF), a W3C recommendation developed within the Semantic Web Activity [http://www.w3.org/RDF/], as the mechanism for formally representing meaning (Beckett 2004, Manola and Miller 2004). RDF in turn relies on other W3C recommendations: the Extensible Markup Language (XML)

[http://www.w3.org/XML/] to provide a syntax for the interchange of semantic representations and Uniform Resource Identifiers (URI) [http://www.w3.org/Addressing/] to provide a syntax for the creation of globally unique names for resources (including concepts).

The RDF approach to semantic representation can be summarized as follows. Information is expressed as a set of statements. Each statement is a triple consisting of a subject, a predicate, and an object. The subject is a *resource*, either an anonymous resource or one that is named by a URI. The predicate is a URI that names a *property*. The object may be another resource or it may be a *literal* value. A set of statements forms a directed graph, in which the resources and literals are nodes and the properties are directed arcs from subject to object. The fact that the RDF graphs describing individual web resources can be merged into a single, large graph forms the basis for interoperation among web resources within the RDF approach.

Interoperation among web resources can occur when their RDF representations are in terms of the same types of resources and properties. Two other W3C recommendations, RDF Schema (Brickley and Guha 2004) and the OWL Web Ontology Language (McGuiness and van Harmelen 2004), are used to give a formal definition to the concepts (in particular, the resource classes and properties) used in the semantic representations of an RDF application. The most basic concepts available in RDF Schema may be summarized as follows:

- *rdfs:Class* is a built-in resource that represents the concept of a class of resources.
- *rdf:Property* is a built-in resource that represents the concept of a property.
- *rdf:type* is a property whose object identifies the class of resource of which its subject is an instance.
- *rdfs:domain* is a property whose object constrains the class of resource that may occur as the subject of the property which is its subject.
- *rdfs:range* is a property whose object constrains the class of resource that may occur as the object of the property which is its subject.
- *rdfs:subClassOf* and *rdfs:subPropertyOf* are properties that define is-kind-of hierarchies among classes and properties.

OWL supports greater machine interpretability of content by adding many more concepts for describing classes and properties. These include relations between classes (such as disjointness versus overlap), cardinality of properties (such as optional versus mandatory versus repeatable), characteristics of properties (such as symmetry or transitivity), equivalence of classes or properties, and richer typing of properties.

A simple example: interoperating across three disparate lexicons

The RDF approach to interoperation is here demonstrated by transforming sample entries from three different lexica into a common semantic representation. The three source lexica follow the EMELD best practice recommendation of using descriptive XML markup. Some of the detail in the original entries (such as morphological analysis, inflected forms, or idioms) has been omitted to reduce each entry to comparable content. All three entries are for a word meaning ‘father’.

The first example comes from a Hopi lexicon compiled by Ken Hill and converted to XML encoding by Will Lewis. Hopi is a Uto-Aztecan language of northeastern Arizona.

```
<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>
```

Figure 1: Sample entry from a Hopi lexicon

The second example comes from a Potawatomi lexicon compiled by Laura Buzzard-Welcher. She used EMELD's FIELD tool [http://emeld.org/school/workroom/lexicon/index.html] which automatically generated the XML. Potawatomi is an Algonquian language of the Great Lakes region.

```
<form id="9939">
   <orthographicform>n#os</orthographicform>
   <grammatical-relation relation-term="is a" pos="Noun" />
   <featurevalue type="Possessibility">InalienablyPossessed</featurevalue>
   <featurevalue type="Gender">Animate</featurevalue>
   <gloss lang="English" value="my father" />
</form>
```

Figure 2: Sample entry from a Potawatomi lexicon

The third example comes from a Sikaiana lexicon compiled by William Donner and converted to XML encoding by Gary Simons. Sikaiana is a Polynesian language of the Solomon Islands.

```
<entry id="tamana">
  <form>tamana</form>
  <sense>
    <gramGrp><pos>no</pos></gramGrp>
    <def>father; true and classificatory; first ascending generation
       lineal and collateral males on the paternal side</def>
  </sense>
</entry>
```

Figure 3: Sample entry from a Sikaiana lexicon

All three examples have comparable content: each gives a citation form, a definition, and grammatical information identifying the part of speech as noun with additional features. (In the case of Sikaiana, the given part-of-speech abbreviation means 'noun, *o* class, inalienably possessed'.) However, the XML markup conventions followed in each case are so different as to render the inherent similarities completely opaque to a machine. Even the detail that the Sikaiana

noun is inalienably possessed would be opaque to a human that did not know the full meaning of the *no* abbreviation. In order to support intelligent searching across disparate resources like these, it is necessary to make the meaning of the markup so transparent that even a machine can interpret it correctly.

This is where RDF comes into play. In an RDF Schema or OWL ontology, the vocabulary of the problem domain is formally defined by declaring standard identifiers for the concepts and specifying how they relate to each other. Then the information content of the original resources can be expressed in terms of RDF statements using that vocabulary. Figure 4 gives an extract from an RDF Schema that defines the notion "linguistic sign." It is given in a notation known as N3; it is formally equivalent to the RDF/XML notation, but much more human readable (Berners-Lee 2005).

```
@prefix gold: <http://www.linguistics-ontology.org/gold-sample#>.

gold:LinguisticSign  a rdfs:Class .

gold:form          a rdf:Property;
     rdfs:domain   gold: LinguisticSign;
     rdfs:range    gold:PhonologicalUnit .

gold:meaning       a rdf:Property;
     rdfs:domain   gold: LinguisticSign;
     rdfs:range    gold:SemanticUnit .

gold:grammar       a rdf:Property;
     rdfs:domain   gold: LinguisticSign;
     rdfs:range    gold:GrammaticalUnit .
```

Figure 4: Extract from RDF Schema defining an RDF vocabulary for *LinguisticSign*

The first line of figure 4 declares the URI prefix for the "gold" namespace. (Note that this is just a sample URI since the details of the GOLD treatment of linguistic sign are still being worked out.) Given this prefix definition, the full name of `gold:LinguisticSign`, for instance, is `http://www.linguistics-ontology.org/gold-sample#LinguisticSign`. The latter is the complete, globally unique name for the concept, while the former is a convenient abbreviation that is equivalent in the context of the namespace declaration.

In N3 notation, an RDF statement is expressed as a space-delimited sequence of subject, predicate, and object terminated by a period. The reserved word *a* is an abbreviation for `rdf:type`. Thus the first statement says that within the GOLD namespace, *LinguisticSign* is defined to be a class of resources. The next three declarations are compound statements; the semicolon indicates that the following two elements are another predicate and object related to the same subject. The three compound statements define the tripartite nature of a linguistic sign as a combination of form, meaning, and grammar. For instance, *form* is defined to be a property whose subject is a *LinguisticSign* and whose object is a *PhonologicalUnit*. This example illustrates one of the conventions followed by RDF practitioners: names of classes are capitalized, while names of properties are not.

Given these definitions (and further definitions describing phonological, semantic, and grammatical units) it is now possible to transform the disparate data sources illustrated in figures 1 through 3 into RDF graphs that are mutually compatible. These are shown in figure 5 through 7. In the graphs, ovals represent resources, rectangles represent literal strings, and the directed

arcs are predicates pointing from subject to object. Note that the `id` attribute on the entries in figures 1 through 3 are used to create a global identifier for the resource. The empty ovals represent "anonymous nodes" that serves to hold a set of relationships but have no external identifier.

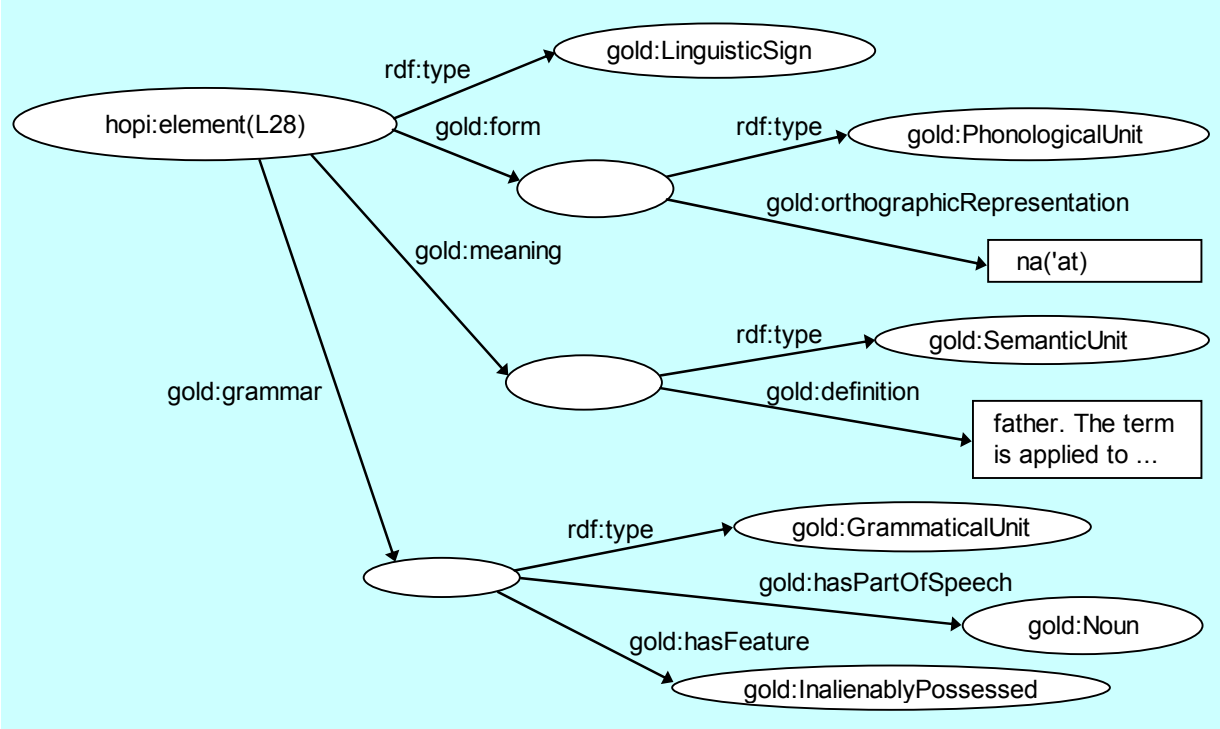

Figure 5: RDF graph for sample entry from a Hopi lexicon

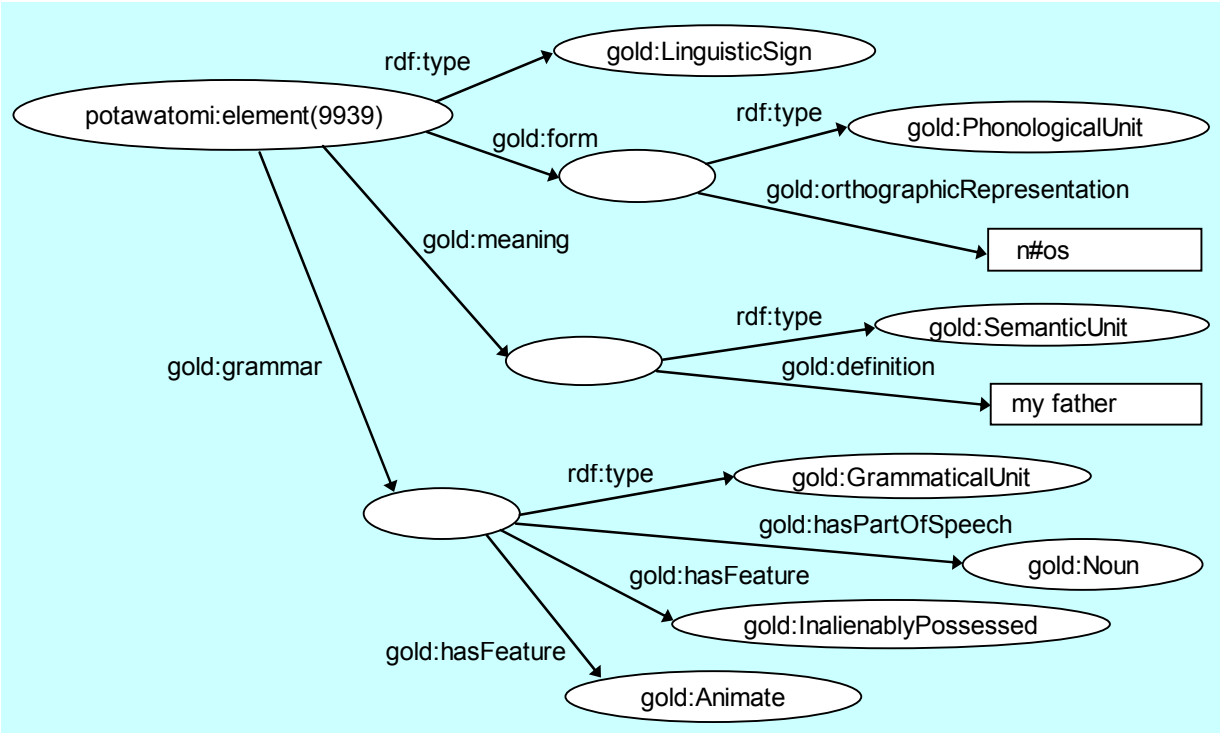

Figure 6: RDF graph for sample entry from a Potawatomi lexicon

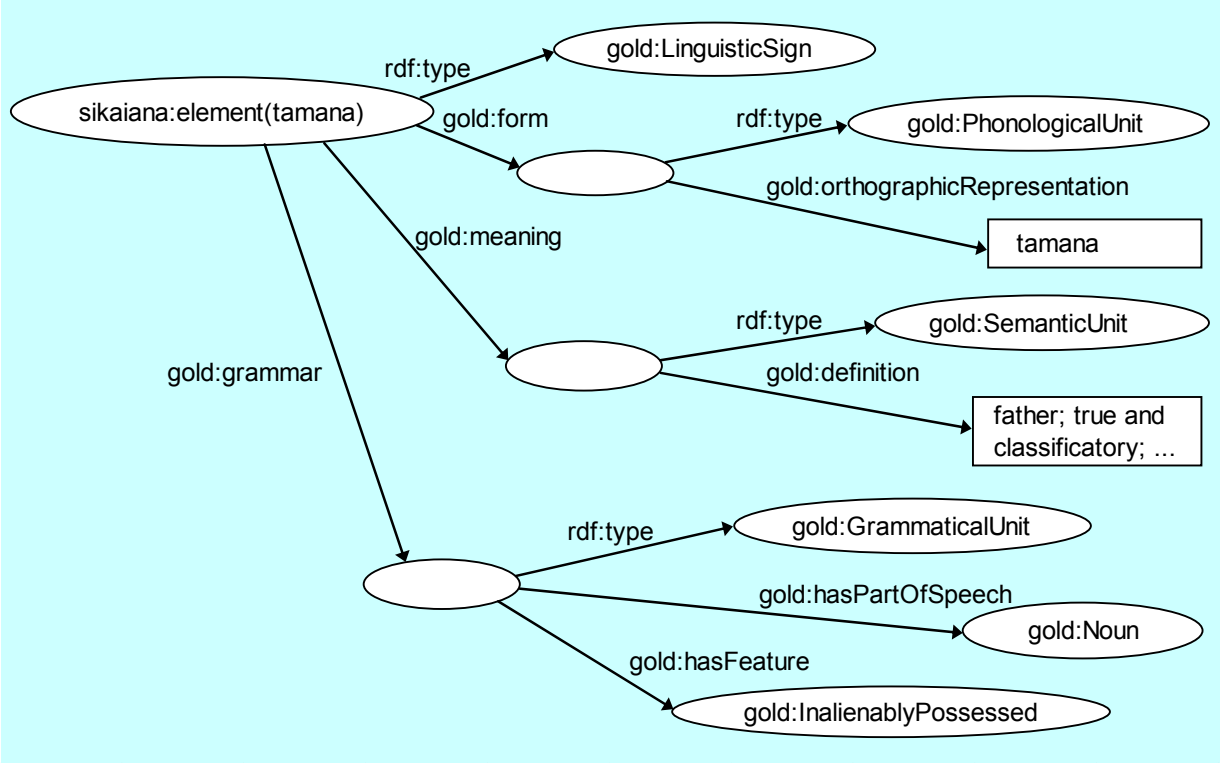

Figure 7: RDF graph for sample entry from a Sikaiana lexicon

A comparison of figures 5 through 7 shows that our once disparate data sources are now represented in forms that are compatible. It is a fundamental characteristic of RDF that any set of valid RDF statements can be combined with any other set of valid RDF statements to produce a larger set of RDF statements. It those statements have no resources in common, then the original graphs will remain disjoint. However, if they share resources, the result is a single contiguous graph. This is illustrated in figure 8 which shows the result of merging the graphs in figures 5 through 7. On the left are the three instance of LinguisticSign. In the middle is the class of LinguisticSign (defined in GOLD) with three `rdf:type` links coming into it. On the far right are the other GOLD concepts that are used by all three signs.

The ability to combine multiple RDF graphs into a single graph is the secret to interoperation. Once the information from the originally disparate resources has been converted to that form it is possible to pose meaningful queries across the combined set of resources. By contrast, it would be meaningless to combine the original XML resources in figures 1 through 3 into a single XML database.

Using a metaschema to convert from markup to meaning

The crucial step in achieving interoperation is thus the process of converting the original language resources in best-practice XML markup into a common RDF representation in terms of the concepts of the GOLD ontology. In order to do this, it is necessary to define how the elements and attributes of the markup in the XML language resources map onto the concepts that are defined in the ontology. More precisely, it is necessary to define a mapping from the

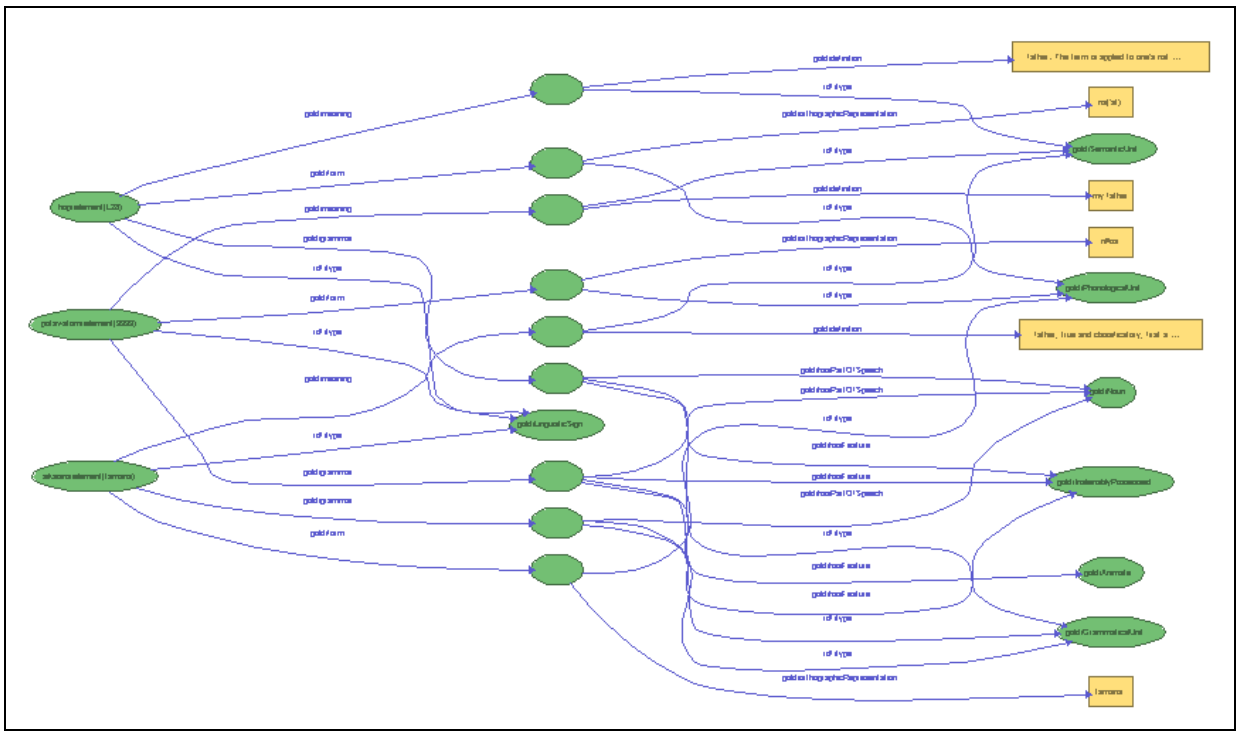

Figure 8: Combined RDF graph for the three sample lexical entries

markup schema to the *semantic schema*. This mapping is called a *metaschema*. These three key terms are defined as follows:

markup schema

A formal definition (as with XML DTD or XML Schema) of the permitted 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 to be interpreted in terms of the concepts of a semantic schema.

The notion of metaschema was developed within the framework of the EMELD project (Simons 2002).The term has been coined in view of two senses of the *meta-* prefix. First, the metaschema transcends the markup schema in that it maps it into a higher, or more abstract, level of representation. But it is not transcendent with respect to the semantic schema. When both schemas are in view, a second meaning of the *meta-* prefix is evoked, namely 'change' (which it bears, for instance, in a word like *metamorphosis*). In other words, a metaschema also embodies a transformation from one schema to another.

The Semantic Interpretation Language (SIL) was developed as a means of formally expressing how markup is to be interpreted in terms of RDF concepts (Simons 2004). An SIL metaschema is an XML document built from metaschema directives; each directive is essentially a processing instruction expressed as an XML element. The directives `resource`, `property`, and

`literal` generate RDF resources, properties, and literals, respectively. Each of these uses a `concept` attribute to name the ontological concept of which the generated element is to be an instance. The `interpret` directive matches specific markup elements of the input resource and indicates how they are to be interpreted semantically.

Figure 9 shows an extract from the metaschema that interprets the markup used in figure 1. The first directive in figure 9 declares that all occurrences of the `<Lexeme>` element are to be interpreted as instances of the GOLD concept *LinguisticSign*. The next directive instructs that `<Head>` elements are to be interpreted as the *form* predicate and the *PhonologicalUnit* which is its object. The empty directive for `<HeadWord>` indicates that it generates no output for the mapping to RDF. Finally, the directive for `<OrthographicForm>` indicates that it generates a literal string related by the *orthographicRepresentation* predicate.

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

Figure 9: Fragment of metaschema for interpreting some of the markup in figure 1

Figure 10 shows another fragment of the same metaschema. It illustrates the use of a termset to interpret the part-of-speech abbreviations originally used by the linguist. The first directive instructs that `<POS>` elements be interpreted as the *grammar* predicate and the *GrammaticalUnit* which is its object. The next directive then instructs that the content of a `<Feature name="cat">` element is to be translated from a string into an RDF reference as the object of the *hasPartOfSpeech* predicate. The translation is to be done by using the termset in the file `Hopi_pos_mapping.xml` (which maps from abbreviations to the corresponding GOLD concepts).

```
<interpret markup="POS"/>
  <property concept="gold:grammar">
    <resource concept="gold:GrammaticalUnit"/>
  </property>
</interpret>
<interpret markup="Feature[@name='cat']">
  <translate concept="gold:hasPartOfSpeech"
mapping="Hopi_pos_mapping.xml"/>
</interpret>
```

Figure 10: Fragment of metaschema that invokes a termset

The implementation of SIL includes a processor for applying the mappings defined in a metaschema to an XML language resource. The result of applying the above metaschema to the XML fragment in figure 1 is shown in figure 11.

```
<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>
```

Figure 11: RDF/XML interpretation of figure 1 in terms of GOLD concepts

The RDF information shown in figure 11 is equivalent to that shown graphically in figure 5. The format shown in figure 11 is the XML serialization format known as RDF/XML. In this notation the relationship of subject to predicate is indicated by embedding the element for the predicate within the element for the subject. An object may further be embedded in the predicate, or the object may be referenced by its URI using the `rdf:resource` attribute.

Conclusion

Once the metaschema processor has been used to convert XML language resources to their RDF interpretation in terms of GOLD concepts, the resources can be loaded into an RDF database. As each resource is loaded, it is merged into the combined RDF graph representing all the information in the database. A query against that database is thus able to operate uniformly across all the resources that have been loaded. The proof-of-concept development work within EMELD has used an open-source RDF database named Sesame [http://www.openrdf.org/] for this purpose. Earlier project papers illustrate RDF queries against collections of lexical data (Simons and others 2004b) and interlinear text data (Simons and others 2004a).

References

Beckett, Dave. 2004. RDF/XML Syntax Specification (Revised). W3C Recommendation, 10 February 2004. Online: http://www.w3.org/TR/rdf-syntax-grammar/

Berners-Lee, Tim. 2005. Primer: Getting into RDF and Semantic Web using N3. Online: http://www.w3.org/2000/10/swap/Primer.html

Berners-Lee, Tim; James Hendler, and Ora Lassila. 2001. The Semantic Web. *Scientific American*, May 2001. Online: http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21

Brickley, Dan and R. V. Guha. 2004. RDF Vocabulary Description Language 1.0: RDF Schema. W3C Recommendation, 10 February 2004. Online: http://www.w3.org/TR/rdf-schema/

Farrar, Scott, William D. Lewis, and D. Terence Langendoen. 2002a A common ontology for linguistic concepts. *Proceedings of the Knowledge Technology Conference,* Seattle, WA, March 2002. Online: http://emeld.org/documents/KnowTech-CommonOntology.pdf

Farrar, Scott, William D. Lewis, and D. Terence Langendoen. 2002b. An ontology for linguistic annotation. *Semantic Web Meets Language Resources: Papers from the AAAI Workshop,* Technical Report WS-02-16, pp. 11-19. Menlo Park, CA: AAAI Press. Online: http://emeld.org/documents/AAAI-OntologyLinguisticAnnotation.pdf

Farrar, Scott and D. Terence Langendoen. 2003. A linguistic ontology for the semantic web. *GLOT International* 7(3), pp. 97-100. Online: http://emeld.org/documents/GLOT-LinguisticOntology.pdf

Lewis, William D., Scott Farrar, and D. Terence Langendoen. 2001. Building a knowledge base of morphosyntactic terminology. In S. Bird, P. Buneman, and M. Liberman (Eds.) *Proceedings of the IRCS Workshop on Linguistic Databases,* 11-13 December 2001, pp. 150-156. Online: http://emeld.org/documents/IRCS-BuildingKnowledgeBase.pdf

Linguist List. 2001. *Workshop on the digitization of language data: the need for standards*. Santa Barbara, California, 21 -24 June 2001. Online: http://linguistlist.org/~workshop/

Manola, Frank and Eric Miller, eds. 2004. RDF primer. W3C Recommendation, 10 February 2004. Online: http://www.w3.org/TR/rdf-primer/

McGuinness, Deborah L. and Frank van Harmelen. 2004. OWL Web Ontology Language Overview. W3C Recommendation, 10 February 2004. Online: http://www.w3.org/TR/owl-features/

Simons, Gary F. 2002. The electronic encoding of lexical resources: A roadmap to Best Practice. *EMELD Workshop on Digitizing Lexical Information*, Ypsilanti, MI. Online: http://emeld.org/documents/roadmap.htm

Simons, Gary F. 2004. A metaschema language for the semantic interpretation of XML markup in documents. Technical report, SIL International, Dallas. Online: http://www.sil.org/~simonsg/metaschema/sil.htm

Simons, Gary F., B. Fitzsimons, A. Lanham, R. Basham, D. T. Langendoen, H. Gonzalez, W. D. Lewis, and S. Farrar. 2004a. A model for interoperability: XML documents as an RDF database. *EMELD Workshop on Linguistic Databases and Best Practice,* Detroit, Michigan, 15-18 July 2004. Online: http://emeld.org/workshop/2004/langendoen-paper.html

Simons, Gary F., William D. Lewis, Scott Farrar, D. Terence Langendoen, Brian Fitzsimons and Hector Gonzalez. 2004b. The semantics of markup: Mapping legacy markup schemas to a common semantics. *Proceedings of 4th Workshop on NLP and XML: RDF/RDFS and OWL in Language Technology.* Association for Computational Linguistics, Barcelona, Spain, July 2004. Online: http://emeld.org/documents/SOMFinal1col.pdf