GOLD AS A STANDARD FOR LINGUISTIC DATA INTEROPERATION: A ROAD MAP FOR DEVELOPMENT

By

Gary F. Simons and Baden Hughes

Paper presented at

2006 E-MELD Workshop on Digital Language Documentation
Lansing, MI.
June 20-22, 2006

Please cite this paper as:

GOLD as a standard for linguistic data interoperation: A road map for development

Gary F. Simons* and Baden Hughes**

* SIL International, Email: <gary_simons@sil.org>
** The University of Melbourne, Email: <badenh@csse.unimelb.edu.au>

1. Introduction

GOLD, the General Ontology for Linguistic Description [1], has somewhat unexpectedly emerged from the EMELD project. Originally conceived of as a morphosyntactic annotation inventory and label mapping scheme, GOLD has now been formalized as an ontology by which disparate data sets can be integrated through a common representation of the basic linguistic features.

In November 2004, a core group of GOLD developers met at the University of California, Fresno to consider the future of GOLD contiguous with, and beyond the lifetime of the EMELD project. The outcome of this meeting was the formation of the GOLD Community.

The overall vision of the GOLD Community is that:

“By agreeing on a shared ONTOLOGY of linguistic concepts and on a shared infrastructure for INTEROPERATION, the linguistics community will be able to produce RESOURCES that describe individual languages in a comparable way, to develop TOOLS that produce these comparable resources, and to query SERVICES that aggregate as many comparable resources as are available.” [2]

In the EMELD context, a significant amount of effort has been invested in the development of GOLD in the first dimension of this vision, namely a shared collection of linguistic concepts. Initial surveying work been completed to glean linguistic concepts and their definitions from published materials. This survey work has been complemented by web data mining activities [3] to further increase the coverage of GOLD. GOLD has been instantiated in several formal versions, and a range of proof of concept implementations have featured at previous EMELD events [4, 5, 6] and other venues [7]. At the 2005 EMELD workshop, version 0.2 of GOLD was reviewed in detail by all the workshop participants in six parallel working group sessions [7a].

However the latter four items from the GOLD Community vision (a shared infrastructure for interoperation, production of resources, development of tools, and provision of services) remain largely unaddressed, and thus there remains considerable effort to be expended in achieving the vision in its entirety. Upon reflection, we believe that there are presently three significant barriers to the widespread adoption of GOLD and subsequent realization of the interoperation goals, vis:
• The complexity of the dissemination format which in effect places the threshold for engagement with GOLD at too high a level;

• the absence of a well defined change process through which GOLD can evolve into a standard that is truly community grounded;

• the lack of compelling GOLD-enabled applications which provide traction amongst end user communities.

In this paper we discuss each of these problems in turn and offer a number of concrete suggestions as to how they can be addressed.

2. Dissemination Format

The current expression of GOLD is as an OWL/RDF document [8], a relatively complex representation grounded in a subtype of formal logic known as “description logic”. This format we have identified as the first impediment. We argue that this need not be an impediment since we believe that this particular expression is but one way that GOLD can be legitimately expressed.

At its core GOLD is a set of linguistic concepts and their corresponding definitions that have been associated with Uniform Resource Identifiers (URIs) — the World-Wide Web Consortium’s standard for constructing globally unique identifiers [8a]. For instance, the concept of “complementizer” in the sense of “a connective that marks a complement clause” is assigned the following URI in GOLD version 0.3:

http://www.linguistics-ontology.org/ns/gold/0.3/gold.owl#Complementizer

Everything up to and including the ‘#’ uniquely identifies the GOLD version 0.3 “namespace”, the string following the ‘#’ is the unique identifier for the concept within GOLD version 0.3. The OWL/RDF format for defining the above concept is as follows (assuming that the whole RDF document defines its default namespace to be: “http://www.linguistics-ontology.org/ns/gold/0.3/gold.owl#”):

<owl:Class rdf:ID="Complementizer">
  <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
    A complementizer is a connective that marks a complement clause (Crystal 1997:75).
  </rdfs:comment>
</owl:Class>

The OWL/RDF format permits many alternative ways of expressing the same information. This poses a problem for software developers who want to use the ontology in a non-OWL application. For instance, the following three representations are different ways of expressing the fact that a Complementizer is a kind of SubordinatingConnective:

<owl:Class rdf:ID="Complementizer">
  <rdfs:subClassOf rdf:resource="#SubordinatingConnective"/>
</owl:Class>

<owl:Class rdf:ID="Complementizer">
  <rdfs:subClassOf>
    <owl:Class rdf:about="#SubordinatingConnective"/>
  </owl:Class>
</owl:Class>
The GOLD source file has instances of all three kinds of statement. They all resolve to the same RDF graph, so this poses no problem for an OWL/RDF application. But for a developer who wants to use GOLD in a different environment, it means understanding the intricacies of RDF representations and programming code that recognizes all three styles of statement above and treating them as equivalent. To make matters worse, the GOLD 0.3 source file actually has the following for this case:

```xml
<SubordinatingConnective rdf:ID="Complementizer">
</SubordinatingConnective>
```

Namely, that Complementizer is an instance of a SubordinatingConnective, which isn't exactly the same as the above, but ends up being a fourth way of expressing what is meant to be the same relationship.

As stated earlier, at its core GOLD is the set of linguistic concepts and their corresponding definitions that have been associated with globally unique identifiers (URIs). It is these identifiers that are the basis of interoperability, not a particular technology (such as OWL/RDF) for expressing the association between identifiers, concepts, and definitions. Hence we argue that GOLD can be validly expressed in a variety of forms including comma or tab delimited files, relational (SQL) database tables, various XML forms recommended by the W3C (e.g. RDFS [9], SKOS [10], OWL [11]), and so on.

Another option would be to use a simplified XML form that could be trivially generated from the existing OWL/RDF by a stylesheet. Our experimentation has resulted in the following XML document structure:

```xml
<concept>
  <id>Complementizer</id>
  <parent>SubordinatingConnective</parent>
  <label>Complementizer</label>
  <definition>A complementizer is a connective that marks a complement clause (Crystal 1997:75).</definition>
</concept>
```

To express GOLD in a comma delimited format, we can simply transform the OWL/RDF to simplified XML format as explained previously, and then reduce the XML even further to CSV. In the CSV case, we envisage a data structure consisting of: id, parent_id, label, definition. Hence the snippet above would be represented thus:

"Complementizer", "SubordinatingConnective", "Complementizer", "A complementizer is a connective that marks a complement clause (Crystal 1997:75)."
(It should be noted that in the original OWL/RDF expression, the label is identical to the identifier or it is missing altogether. In every case it should probably be the standard orthographic expression of the term as lower-case words separated by spaces.)

Another format is HTML, which is arguably most accessible by virtue of the internal linking that is possible through named anchors and HREFs. In the example below, we have transformed the simplified XML representation into an XHTML presentation form, where each entry is a definition list headed by concept identifier, followed by the definition, and a list of the relations between concepts expressed as broader term (bt) generalizations, narrower term (nt) specializations, or just related terms (rt). The following source code:

```xml
<dl>
  <dt><a class="alpha" name="GenderFeature"></a>GenderFeature</dt>
  <dd>GenderFeature the class of all grammatical genders found in language. It may be best subsumed under a more general class for noun classification.</dd>
  <dd>bt:<a class="normallink" href="#MorphosyntacticFeature">MorphosyntacticFeature</a></dd>
  <dd>rt:<a class="normallink" href="#GenderValue">GenderValue</a></dd>
</dl>

<dl>
  <dt><a class="alpha" name="GenderValue"></a>GenderValue</dt>
  <dd>bt:<a class="normallink" href="#FeatureValue">FeatureValue</a></dd>
  <dd>nt:<a class="normallink" href="#Animate">Animate</a></dd>
  <dd>nt:<a class="normallink" href="#Feminine">Feminine</a></dd>
  <dd>nt:<a class="normallink" href="#Inanimate">Inanimate</a></dd>
  <dd>nt:<a class="normallink" href="#Masculine">Masculine</a></dd>
  <dd>rt:<a class="normallink" href="#GenderFeature">GenderFeature</a></dd>
</dl>

renders an output vis:

**GenderFeature**

GenderFeature is the class of all grammatical genders found in language. It may be best subsumed under a more general class for noun classification.

- bt: MorphosyntacticFeature
- rt: GenderValue

**GenderValue**

- bt: FeatureValue
- nt: Animate
- nt: Feminine
- nt: Inanimate
- nt: Masculine
- rt: Gender Feature

These examples, along with the transformations, are posted online at [http://www.csse.unimelb.edu.au/~badenh/papers/emeld06/examples.html](http://www.csse.unimelb.edu.au/~badenh/papers/emeld06/examples.html) for the interested reader to consider.
The key to the use of alternate expressions for GOLD is that they must be automatically derived from a core representation, and hence able to be quickly and reliably regenerated if the core GOLD changes.

Given that there are a range of options for expressing the concepts, definitions and identifiers that make up GOLD, what is the best format for dissemination? We believe that some internal division of GOLD is required, based around the idea that there are in fact three major kinds of content that are combined in GOLD (data structures, concept hierarchies, and properties).

It is likely the most widely used parts of GOLD will be the mutually exclusive concept hierarchies. We envision that these will be of use to the most applications as well as being most familiar to linguists. Of these, the most immediately useful are probably the syntactic categories (e.g., part of speech), the unit types (e.g., phrase, word, morpheme, and their subtypes), and the morphosyntactic properties (e.g., features). Applications are likely to treat these data sets as lists from which selections can be made, and so it would be convenient to at least store these subsets in a variety of formats as distinct resources. For this purpose, SKOS provides an immediately tractable specification for the information we should retain about each concept (including alternate terms, terms in other languages, and notes of various types).

The other two kinds of content in GOLD (the data structures and the properties) are unlikely to be used directly by application instances (e.g., in drop down selectors). Hence we propose that their dissemination format could be quite different to the concept hierarchies.

However, this functional modularization poses a question with regard to overall management GOLD: how do we manage the entire ontology? We suspect that a SQL database (with the advantage that the relational data model is well grounded in application development contexts) is likely to give the most flexibility as the primary repository for managing GOLD. Various dissemination formats could then be generated by automatic transformation. We recognize that this recommendation has an impact on the infrastructural requirements for ongoing development and on the associated editorial processes.

Overall, we believe that such expression in multiple forms is critical to the adoption of GOLD by software developers, since it significantly lowers the barrier to entry imposed by having only the single representation in OWL. Adoption by more software developers should in turn promote wider adoption among linguists; and hence consideration of the dissemination issues warrants considerable debate.

3. Ontology Evolution and Change

The second issue we have identified is the absence of a well-defined change process through which GOLD can evolve into a standard that is truly community grounded. In its initial stages, GOLD was the work of a small EMELD team at the University of Arizona, and then predominantly developed to its current status by one member of that team. While the content within GOLD is based on broad coverage surveys across the linguistics community and input from many linguists who have given feedback, it still ‘belongs’ to a small group. If GOLD were a finished product this approach might work, since users would be able to evaluate that product in order to decide if they would use it. However, in its current 0.3 version GOLD is far from finished and will obviously undergo many rounds of successive refinement before it reaches the status of a 1.0 version. In order to achieve a version 1.0 that has broad community
acceptance, there must be substantial community participation in the development process. Thus we are caught in a ‘catch 22’ situation — linguists and developers aren’t ready to adopt it because it isn’t yet fit for final use, but it can’t reach that state until a substantial number of linguists and developers adopt it in order to test its use in real applications. In order to break this cycle, the change management process must satisfy two key requirements:

- In order for linguists to start using an ontology they can see to be deficient, they need to understand the process by which it can be changed and have faith that their participation in that process would fix the deficiencies that affect their work.
- In order for tool developers to start using an ontology that is guaranteed to change, they need to understand the constraints that are being placed on the change process and have faith that subsequent changes to the ontology will not break their implementation and render their work useless.

The current approach to managing change in GOLD does not appear to be meeting these two requirements. We believe that absence of such a change process serving as a significant barrier to adoption by would-be users and developers.

In order to address these requirements, we propose that GOLD requires a formalized process for managing change, perhaps modeled after practices in other communities [12]. It should be noted that this change process is not simply the use of a version control system to handle revisions to an inherently textual object such as an OWL document, but rather a process by which the community engages with the evolution of a standard. Such a process would specify how changes are proposed by members of the community, how such proposals are reviewed by the community at large, and ultimately how the decision to change or not to change is actually made. Supporting infrastructure such as version control systems and version specific namespaces are secondary to this overall change management process.

Without a constrained change management process, a GOLD adopter would have little confidence that an incremental version update would not break their implementation. The constraints needed by developers go beyond specification of the change process to include constraints on the nature of change as well—some kinds of changes to GOLD require only that a tool reload a revised ontology file, whereas other kinds of changes require that the tool be reprogrammed. We have seen this problem already with GOLD in the significantly different structure of versions 0.1, 0.2 and 0.3, whereby applications which use GOLD in its OWL instantiation have been required to be reengineered around the changing structure of the ontology itself. Adoption by the developer community requires that there be constraints that will keep that kind of change to a minimum.

A community-centric change management process allows for review of proposed changes by the wider community, thus fostering greater acceptance and wider adoption. In similar community-based efforts, the actual work of developing a standard is performed in working groups [13]. We propose that in the absence of dedicated funding for the expansion of GOLD's coverage, the working group model would allow interested parties to contribute to the development of GOLD in a lightweight but effective manner.

4. Applications and Data

The third impediment to adoption, the lack of compelling GOLD-enabled applications which provide traction amongst end user communities, we believe can be addressed in several ways.
Language profiles are an application for GOLD that has been introduced into the community [14]. A language profile is a formalized account of the morphosyntactic categories and features of a specific language in terms of the concepts in GOLD. Language profiles can be compared directly to compare the categories and features of languages; they have also proven foundational for interoperation of textual and lexical data. An initial schema governing the format of a language profile was introduced in [5]; this has not been refined further at the time of writing, and represents the current standard for the expression of a language profile.

Developing tools to facilitate the creation of language profiles as well as services to facilitate their comparison are obvious first steps. One such tool which includes support for building a language specific selection from GOLD is the EMELD FIELD [15] tool, which allows for an end user to select the specific features from GOLD which are applicable to a given language under analysis within the FIELD environment.

Additionally there are very few examples of linguistic data that are annotated with GOLD concepts, despite notable efforts [3]. We propose that another first step would be to start a repository for such data, and to publish the metadata for the collection as an OLAC data provider [18], thus enabling discovery of richly annotated and freely available data for interested researchers to work with.

Templates for linguistic data structures, such as the models proposed for interlinear text [19] and paradigms [20], already support the embedding of GOLD concept identifiers through the use of XML namespaces [21]. Where there are small, carefully curated collections of linguistic data based on these templates, it could be possible to provide services which automatically embed GOLD concept identifiers into these data structures, and any guidance for using these templates should include similar advice.

In addition, we require tools which fit within existing linguistic analysis workflows, yet natively support GOLD. The EMELD-developed FIELD [15] is one example of such a tool, although there are relatively few users of this framework for analysis at this time. Other available tools which support GOLD include OntoELAN [16] and the Fieldworks Morphological Glossing Assistant [17], although these too lack user uptake at the time of writing.

There are well known impediments to the adoption of new linguistic software. To counter this we need to provide examples of ‘embedded’ GOLD in commonly used linguistic analysis tools. This will allow documentary linguists to start to use GOLD without even realizing it (or at least without additional effort required to change software of choice). Initial efforts in this area have included experiments with the inclusion of GOLD-based rangesets in Shoebox/Toolbox [22], and the recently emerging support for GOLD annotated data from computational frameworks such as Alchemist [23]. Such stealth mode adoption, based on the collaboration of GOLD maintainers and application developers, will significantly advance the currency of GOLD.

5. Conclusion

It is our view that these steps are necessary to ensure both the ongoing development of GOLD and its adoption in the wider community. This agenda is strongly motivated by the goal of interoperation across a wide variety of resources and services as outlined in the
GOLD Community Vision statement. Once such interoperation is achieved in the short to medium term, individual services will be able to harvest resources and exploit the inferencing capabilities of semantically rich expressions of GOLD (like an OWL rendition) to achieve the more ambitious goal of a linguistic knowledge base in the future.

References


Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Simons, G. F.; Hughes, B.

Title:
Gold as a standard for linguistic data interoperation: a road map for development

Date:
2006

Citation:

Publication Status:
Published

Persistent Link:
http://hdl.handle.net/11343/34527

File Description:
Gold as a standard for linguistic data interoperation: a road map for development

Terms and Conditions:
Terms and Conditions: Copyright in works deposited in Minerva Access is retained by the copyright owner. The work may not be altered without permission from the copyright owner. Readers may only download, print and save electronic copies of whole works for their own personal non-commercial use. Any use that exceeds these limits requires permission from the copyright owner. Attribution is essential when quoting or paraphrasing from these works.