Principles of the GOLD Ontology & Conversion of GOLD to DCIF

Presenters: Anthony Aristar, Evelyn Richter

Outline

• The Structure of the GOLD Ontology
  – History of GOLD
  – Why an Ontology Was Chosen
  – Structure of GOLD
  – One Advantage of an Ontology

• GOLD to DCIF
  – Refining GOLD
  – Mapping the Content of GOLD to DCIF
  – Mapping the Structure of GOLD to DCIF
  – Summary & Outlook
How GOLD began

- GOLD had its origin in the very first EMELD workshop, held at Santa Barbara in 2000.
- At this stage there was no consensus on how to deal with linguistic markup.
- Most people wanted linguists to use a single set of markup categories, and try and “force” everyone to use it.

The Solution

- Since the idea that making academics do anything they didn’t want to was equivalent to trying to herd a group of Siamese cats, the workgroup on markup (which included Gary Simons & Terry Langendoen) understandably came to the conclusion that no one would use the same markup.
- So, let everyone use the markup they wanted, and build a set of concepts that everyone could link to automatically.
Background

• Behind this was the fact that both Gary and Terry had been part of the group that planned the TEI.
• Using the ideas behind the TEI, you can use automatic methods to interpret tags.
• There was a general feeling that something inspired by the TEI would be used, though TEI had been intended for format, not content.

The Beginning of GOLD

• Just then, the Suggested Upper Merged Ontology (SUMO) emerged.
• This was remarkably innovative for its time… the idea was based on the idea that with the right structure — an ontology — machine reasoning was possible.
• It was thus proposed that an ontology built on SUMO serve as “interlanguage” for translation among markups.
Such an Ontology would:

- Have a concept hierarchy
- Have an enriched taxonomy
- Act as an interlingua

And finally it would:

- Eliminate the need for a gold standard of linguistic terminology

The Structure of GOLD

- It follows the standard ontological structure you would expect:
  - It is a formal explicit description of classes or concepts in a specific domain.
  - A class can have subclasses that represent concepts that are more specific than the superclass.
  - A class can have instances, i.e. actual representatives of classes.
  - Each concept can have properties describing various features, attributes and relationships of the concept.
  - Facets: Restrictions can be instantiated on properties (sometimes called role restrictions).
In Short

• GOLD is more than a taxonomy
  – It incorporates a taxonomy
• But it can also express a part/whole relationship, and restrictions on them
• Simply put, hierarchy is taxonomic, properties are part/whole

For Example
One Advantage of an Ontological Approach: Language Profiles

• Since it has the structure it does, an ontology will also allow you to define a subset of itself which typifies a particular language.
• It will also allow you to write software that can build very basic grammars of a language automatically.

Why is this?

• The ontology “knows” where individual pieces of data fit in a grammar.
• For example, suppose through the ontology we have defined the following for a language:
  – The language has six cases that go on nouns only.
  – It knows that nouns have three numbers in that language.
Why is this?

• Then…
  – It can build a paradigm for the noun.
• And it knows this has nothing to do with verbs…
• It can also write statements such as this “Nouns have six cases and three numbers, singular, dual and plural”
• So, you can tell it to automatically generate a very simple grammar based on this knowledge.

Refrining GOLD

• December 2009 to July 2010
• GOLDComm team at LINGUIST List first checked every single existing definition & tracked any problems found
• Systematically solved the issues by adding/improving definitions & adding citations
Refining GOLD: Some statistics

<table>
<thead>
<tr>
<th>Issue</th>
<th>No. of Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>432</td>
</tr>
<tr>
<td>No definition</td>
<td>108</td>
</tr>
<tr>
<td>Bad definition</td>
<td>22</td>
</tr>
<tr>
<td>No source or incomplete source</td>
<td>194</td>
</tr>
<tr>
<td>Duplication</td>
<td>11</td>
</tr>
<tr>
<td>Formatting issues</td>
<td>97</td>
</tr>
<tr>
<td>Grammatical error</td>
<td>23</td>
</tr>
<tr>
<td>Delete final word</td>
<td>3</td>
</tr>
<tr>
<td>Page error</td>
<td>2</td>
</tr>
<tr>
<td>Lengthy definition</td>
<td>4</td>
</tr>
</tbody>
</table>

Refining GOLD: Results

- Every concept in the ontology has a definition now (except OWL-specific ones)
- Unclear or insufficient definitions have been superseded
- All but ~52 definitions have citations now
- All new sources used go into the GOLD RDF bibliography for later import
Mapping the Content of GOLD to DCIF

- Relatively straightforward since the GOLD standard (as opposed to the web view) only contains concepts, their unique URI, their definitions and relations between the concepts

<table>
<thead>
<tr>
<th>GOLD</th>
<th>DCIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept Label</td>
<td>Data Element Name &amp; Identifier</td>
</tr>
<tr>
<td>Concept URI</td>
<td>Source</td>
</tr>
<tr>
<td>Concept Definition</td>
<td>Definition (in Language Section)</td>
</tr>
</tbody>
</table>

Aug 4, 2010 RELISH Workshop 17

Mapping the Structure of GOLD to DCIF

- Hierarchical/taxonomic relations from GOLD (parent-child-relationship, other relations need more work before adding them to DCIF)
- DCR not really intended to include complex relations
- Hierarchical relations from GOLD could only be partially retained

Aug 4, 2010 RELISH Workshop 18
Mapping the Structure of GOLD to DCIF

- DCIF has <isA> element for simple data categories (DCs) and <value> elements in the Conceptual Domain section for complex/closed DCs
- DCs listed under the <value> element all have to be simple
- Complex/open DCs cannot have <value> element

Aug 4, 2010 RELISH Workshop 19

Mapping the Structure of GOLD to DCIF

- Decision had to be made which concepts in GOLD hierarchy should become complex/closed DCs
- All children and lower would have to be simple DCs
- All parents and higher would become complex/open DCs and therefore not be linked to their children in any way

Aug 4, 2010 RELISH Workshop 20
Mapping the Structure of GOLD to DCIF

Complex/open DCs
Complex/closed DC
Simple DCs

Child of VoiceProperty in GOLD
Child of MorphosyntacticProperty in GOLD

Aug 4, 2010 RELISH Workshop 21
Summary & Outlook

- High level concepts not linked to their children and ancestry in DCR anymore
- Only by looking at <isA> elements of lower level concepts can the lineage to complex/closed DCs be traced
- There are plans to establish Relations Registry (using RDF/OWL) which could solve these issues

Thank you.

GOLD 2010:
http://linguistics-ontology.org/gold

GOLD Community website:
http://linguistics-ontology.org