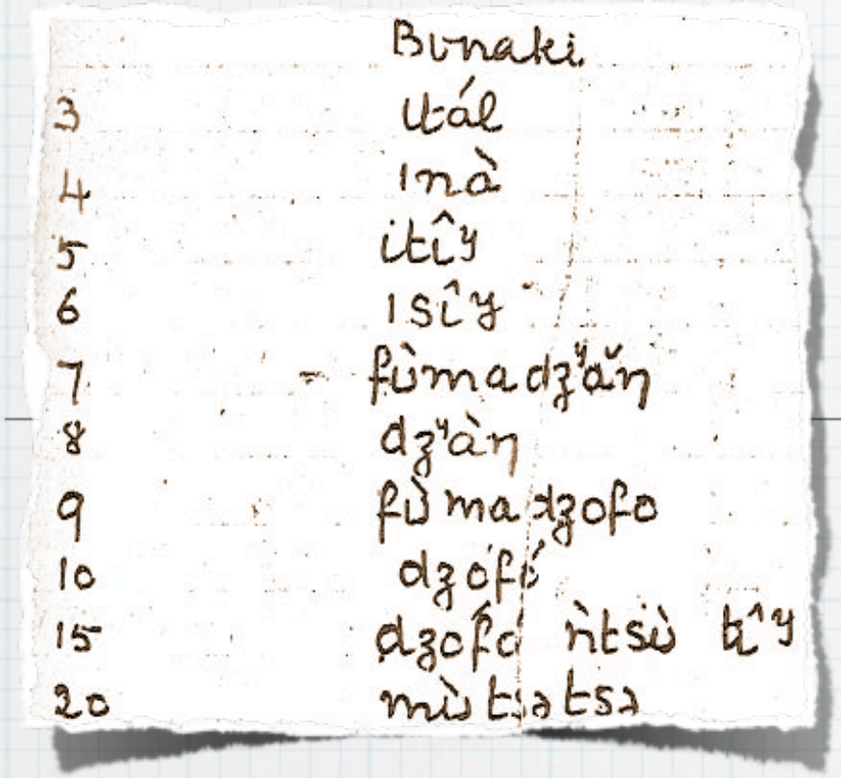


LEGO, RELISH, and related initiatives

Jeff Good
University at Buffalo
jcgood@buffalo.edu

Grants background

- LEGO: NSF-funded project to build a "datanet" of interoperable lexical resources
  - Based at Eastern Michigan University
  - Satellite work at University at Buffalo
- RELISH: DFG–NEH project to support trans-Atlantic standards harmonization
  - DFG efforts at Nijmegen and Frankfurt
  - NEH efforts at Eastern Michigan

Intellectual context

- Lexicons seem like good candidates for exploring data interoperability
- There is lots of variation in their structure, but most show a lot of overlap
  - Organized around word entries
  - Entries have a *form* part, a *grammar* part, and a *meaning* part
- How do we migrate legacy materials?
- What should new resources look like?

Lexicon formats

- There are an enormous number of encoding schemes for lexical data
- None of these has asserted itself as a general standard

TEI Example

```
<entry>
 <form>
  <orth>competitor</orth>
  <hyph>com|peti|tor</hyph>
  <pron>k@m"petit@(r)</pron>
 </form>
 <gramGrp>
  <pos>n</pos>
 </gramGrp>
 <def>person who competes.</def>
</entry>
```

Shoebox example

```
\lx srapa1
\ps vt
\ge slap
\de slap with open hand
\dt 27/Aug/91
```

Lexicon diversity

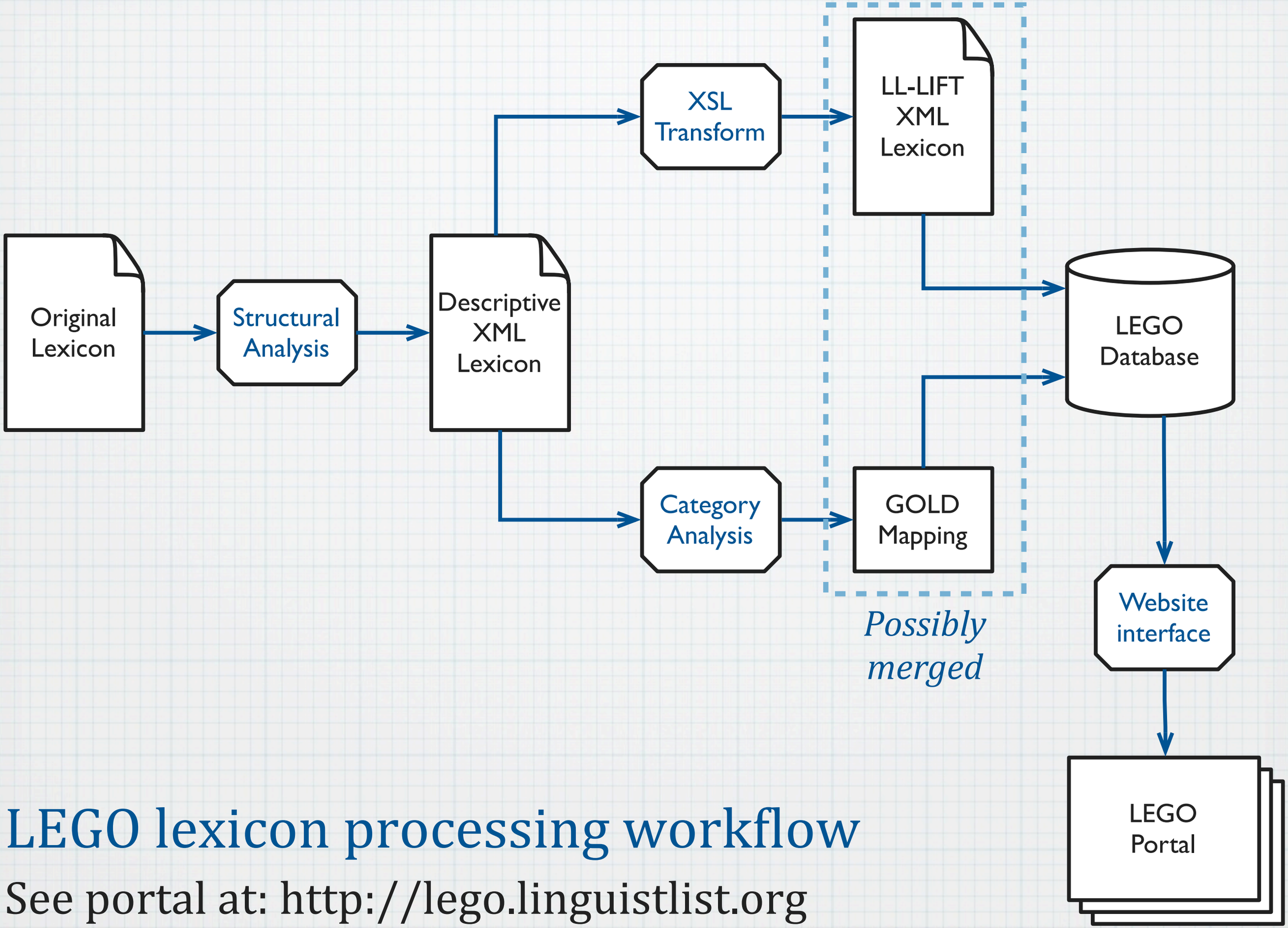
- Both LEGO and RELISH are focused on lexical resources for minority languages
- Major languages will always have extensive, dedicated support
- More generalized support is needed for languages of less economic significance
- Such languages are also less likely to have established lexicographic traditions
- Work on major languages in Europe tends to be more relevant than work in the U.S.

LEGO

- The bulk of the work of LEGO focused on legacy data conversion
- This fed into general recommendations
- Data sources
  - Around twenty dictionaries (Michigan)
  - More than 2500 wordlists (Buffalo)
- What steps were required to convert them?
- What target format would allow for tractable conversion and interoperation?

LEGO: Eastern Michigan

- LEGO lexicons ranged from relatively simple to quite complex
- Basic work plan
  - Analysis of entries to find a consensus data model
  - Legacy format conversion (e.g., Word dictionaries to something structured)
  - Conversion to interoperation format (including mapping to GOLD)

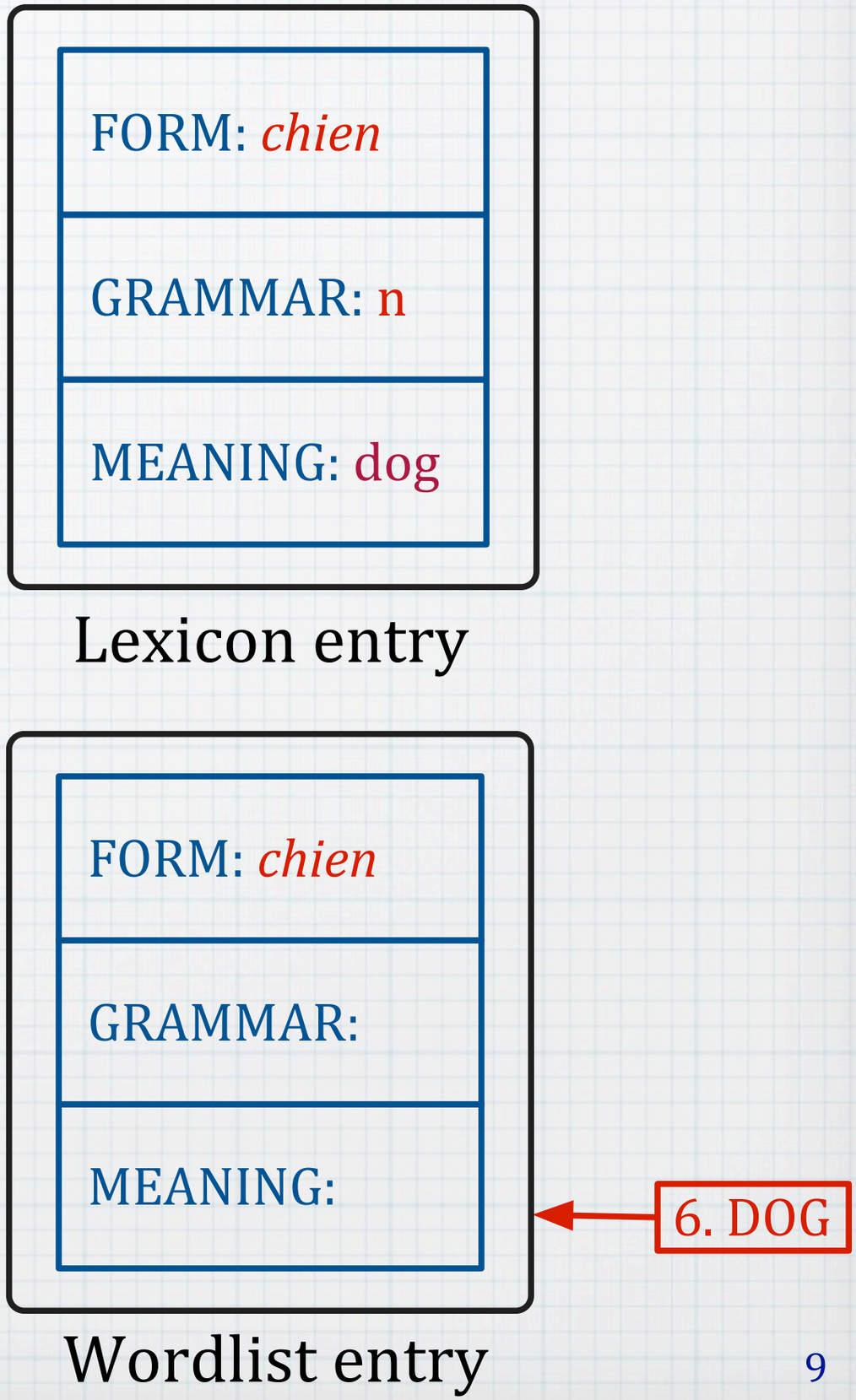


LEGO lexicon processing workflow

See portal at: http://lego.linguistlist.org

LEGO: Buffalo

- The Buffalo side of LEGO focused on the conversion of thousands of wordlists
  - Dictionary: Form → Meaning
  - Word list: Concept → Form
- Concepts are drawn from concepticons
- LEGO developed a unified concepticon



```
<skos:concept xmlns:skos="http://www.w3.org/2008/05/skos/">
   <lego:conceptId>1</lego:conceptId>
   <dc:description xsi:type="lego:default" lego:source="LWT"
       lego:sourceID="1.1" lego:label="the world" />
   <dcterms:references>
       http://wold.livingsources.org/meaning/1.1
   </dcterms:references>
   <dc:description lego:source="IDS" lego:sourceID="1.1"
       lego:label="world" />
   <dc:description lego:source="UW" lego:sourceID="2"
       lego:label="world" />
</skos:concept>
```

Unifying Usher-Whitehouse, IDS, and LWT concepts

Mocoví Lexicon

Source: Grondona, Verónica. 1991. Mocoví FIELD Database.

Entry: ɟin

Author Label(s): Verb
Gold Concept(s): Verbal

Definition:
English: cheat, lie
Spanish: engañar, mentir

Example
(Mocoví) yim se-sa-ɟiɲ-itʃ" (92-00-09)
(English) 1pron neg-1S-CHEAT-2O "I don't cheat you, I don't lie to you."
(Spanish) 1pron neg-1S-ɛNGA?AR-2O "Yo no te engaño, yo no te miento."

Bibliographic Note
English translation: Roberto Ruiz, Spanish translation: Roberto Ruiz

Bibliographic Note
Source name: Roberto Ruiz, Elicited date: 1992-07-26

Semantic Note
Character, Temperament, Manner, Behavior

Paradigmatic Variant : (da mare) ya-ɟin
Author Token(s): Poss Pron Aff, 3rdSg
Gold Concept(s): Affix, Third Person

Paradigmatic Variant : se-sa-ɟiɲ-i?
Author Token(s): NegativeMood
Gold Concept(s): Negative Polarity

cf
Headword: ɟiɲi?

cf
Headword: yaɟiin

Entry: ɟin

Author Label(s): Verb
Gold Concept(s): Verbal

Definition:

English: cheat, lie
Spanish: engañar, mentir

Example

(Mocoví) yim se-sa-ɟiɲ-itʃ' (92-00-09)
(English) 1pron neg-1S-CHEAT-2O "I don't cheat you, I don't lie to you."
(Spanish) 1pron neg-1S-ɛNGA?AR-2O "Yo no te engaño, yo no te miento."

Bibliographic Note

English translation: Roberto Ruiz, Spanish translation: Roberto Ruiz

Bibliographic Note
Source name: Roberto Ruiz, Elicited date: 1992-07-26

Semantic Note
Character, Temperament, Manner, Behavior

Paradigmatic Variant : (da mare) ya-ɟin
Author Token(s): Poss Pron Aff, 3rdSg
Gold Concept(s): Affix, Third Person

Paradigmatic Variant : se-sa-ɟiɲ-iʔ
Author Token(s): NegativeMood
Gold Concept(s): Negative Polarity

cf

Headword: ɟiɲiʔ

cf

Headword: yaɟin

```
<entry id="_8590">
  <trait name="original-id" value="8590"/>
  <lexical-unit>
    <form lang="moc-Latn">
      <text>ɟin</text>
    </form>
  </lexical-unit>
  <sense>
    <grammatical-info value="Verb">
      <trait name="GOLDConcept" value="Verbal"/>
    </grammatical-info>
    <definition>
      <form lang="eng">
        <text>cheat, lie</text>
      </form>
      <form lang="spa">
        <text>engañar, mentir</text>
      </form>
    </definition>
    <example>
      <form lang="moc-Latn">
        <text>yim se-sa-ɟiɲ-iʧ' (92-00-09)</text>
      </form>
      <translation>
```

```
    </form>
    <form lang="spa">
      <text>engañar, mentir</text>
    </form>
  </definition>
  <example>
    <form lang="moc-Latn">
      <text>yim se-sa-ɟiɲ-iʧ' (92-00-09)</text>
    </form>
    <translation>
      <form lang="eng-Latn">
        <text>1pron neg-1S-CHEAT-2O ''I don't cheat you...''</text>
      </form>
    </translation>
    ...
  </example>
  ...
  <note type="semantic">
    <form lang="eng-Latn">
      <text>Character, Temperament, Manner, Behavior</text>
    </form>
  </note>
  <relation type="cf" ref="_7421"/>
</entry>
```

Kinds of interoperation

- Various kinds of interoperation
  - Character encoding (e.g., Unicode)
  - Markup (e.g., use of XML)
  - Structural (e.g., shared entry structure)
  - Semantic (e.g., compatible categories)
  - LEGO and RELISH were mainly concerned with the last two of these

Structure: LL-LIFT

- LEGO based its data markup format on Lexicon Interchange Format (LIFT) XML
- But LIFT is very unconstrained, not allowing for data structure interoperation
- Therefore, LL-LIFT was created
  - Any LL-LIFT lexicon is also a LIFT lexicon
  - Not all LIFT lexicons conform to LL-LIFT
  - Use of LL-LIFT is what allows a new lexicon to fit into the portal interface

Structure: LMF

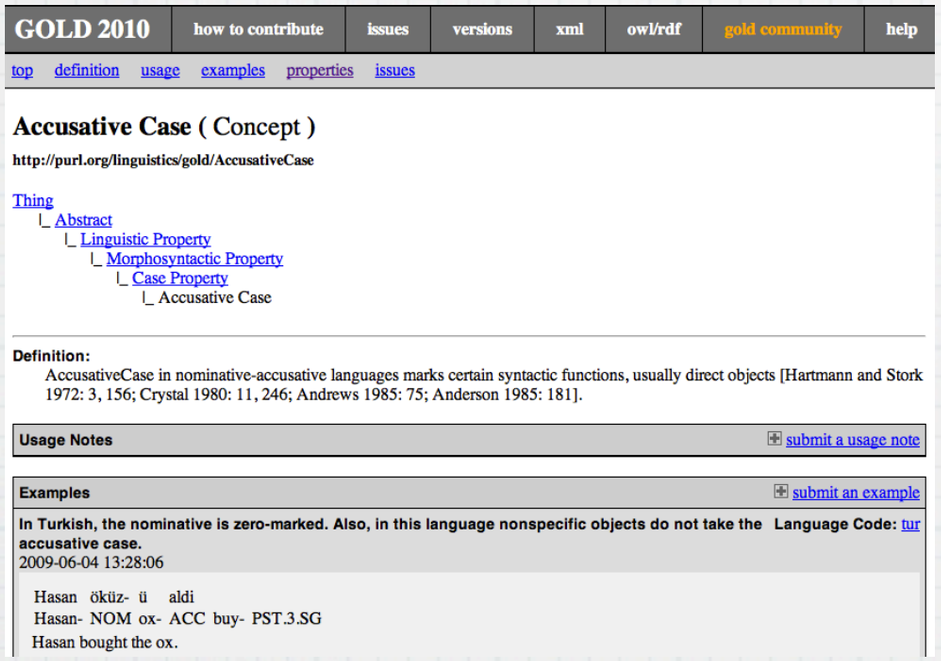
- The European side of RELISH had adopted Lexical Markup Framework
- This is a "meta-standard" for describing lexicon structures
- LEXUS, developed at MPI Nijmegen, used LMF as a framework for its lexicons
- A key activity of RELISH was devising a LEXUS-LMF↔LL-LIFT conversion strategy
- From a linguistic perspective, the differences are often trivial…
- …but conversion can be time consuming

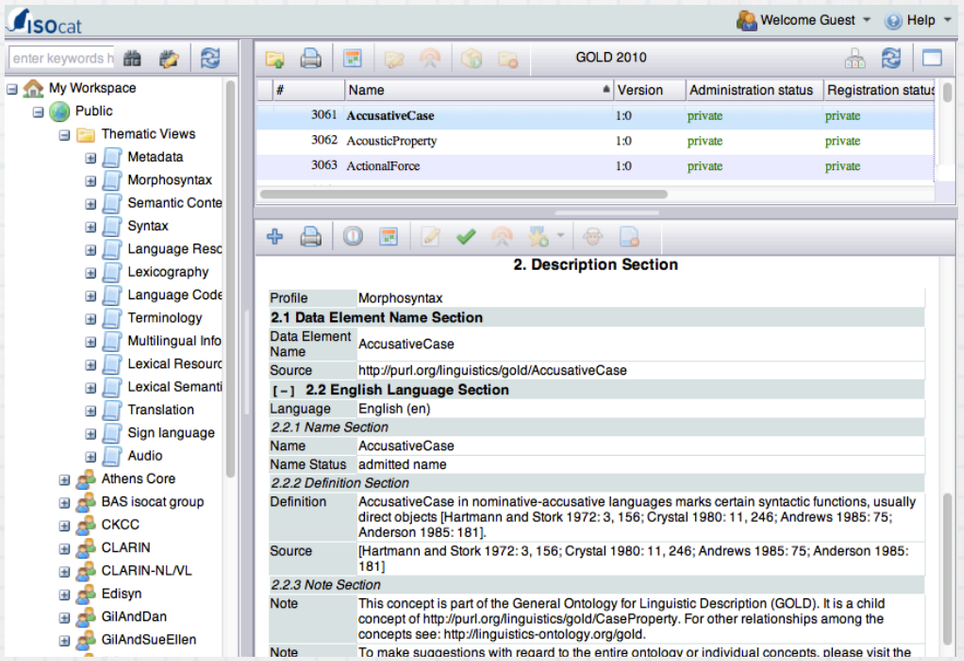
Semantic interoperation

- Semantic interoperation requires shared categories for describing data
- Somehow the fact that *noun class* and *gender* may be the "same" must be encoded
- The current standard solution
  - Allow everyone to use their own labels
  - Map the labels to a fixed list of categories
- Mapping can become hard—sometimes one finds hybrid categories, e.g., `ge` in MDF

GOLD and ISOcat

- The GOLD ontology provides one fixed category list—as well as a taxonomy
- ISOcat is a general category registry, without much additional structure
- ISOcat contains the GOLD categories, and various others, and can be easily extended





Accusative Case (Concept)

http://purl.org/linguistics/gold/AccusativeCase

Thing
  |_ Abstract
    |_ Linguistic Property
      |_ Morphosyntactic Property
        |_ Case Property
          |_ Accusative Case

Definition:

AccusativeCase in nominative-accusative languages marks certain syntactic functions, usually direct objects [Hartmann and Stork 1972: 3, 156; Crystal 1980: 11, 246; Andrews 1985: 75; Anderson 1985: 181].

Usage Notes submit a usage note

Examples submit an example

In Turkish, the nominative is zero-marked. Also, in this language nonspecific objects do not take the accusative case. **Language Code:** tur

2009-06-04 13:28:06

Hasan öküz- ü aldi
Hasan- NOM ox- ACC buy- PST.3.SG
Hasan bought the ox.

ISOcat

Welcome Guest | Help

enter keywords h

- My Workspace
  - Public
    - Thematic Views
      - Metadata
      - Morphosyntax
      - Semantic Conte
      - Syntax
      - Language Resc
      - Lexicography
      - Language Code
      - Terminology
      - Multilingual Info
      - Lexical Resourc
      - Lexical Semanti
      - Translation
      - Sign language
      - Audio
    - Athens Core
    - BAS isocat group
    - CKCC
    - CLARIN
    - CLARIN-NL/VL
    - Edisyn
    - GilAndDan
    - GilAndSueEllen

GOLD 2010

#	Name	Version	Administration status	Registration status
3061	**AccusativeCase**	1:0	private	private
3062	AcousticProperty	1:0	private	private
3063	ActionalForce	1:0	private	private

2. Description Section

Profile	Morphosyntax
2.1 Data Element Name Section	
Data Element Name	AccusativeCase
Source	http://purl.org/linguistics/gold/AccusativeCase
[–] 2.2 English Language Section	
Language	English (en)
2.2.1 Name Section	
Name	AccusativeCase
Name Status	admitted name
2.2.2 Definition Section	
Definition	AccusativeCase in nominative-accusative languages marks certain syntactic functions, usually direct objects [Hartmann and Stork 1972: 3, 156; Crystal 1980: 11, 246; Andrews 1985: 75; Anderson 1985: 181].
Source	[Hartmann and Stork 1972: 3, 156; Crystal 1980: 11, 246; Andrews 1985: 75; Anderson 1985: 181]
2.2.3 Note Section	
Note	This concept is part of the General Ontology for Linguistic Description (GOLD). It is a child concept of http://purl.org/linguistics/gold/CaseProperty. For other relationships among the concepts see: http://linguistics-ontology.org/gold.
Note	To make suggestions with regard to the entire ontology or individual concepts, please visit the

Data model vs. encoding

- Most work has focused on XML standards
- But, that turns out to be a relatively trivial part of the process
- More significant is the abstract data model
  - A lexicon built around consonantal roots is very different from one based on “words”
  - Some lexicons contain texts or rich paradigms, also complicating the model
  - A shared notion of the abstract “entry” is an important starting point

√bhw 1a [N and DTF 1.37 √bhw « couleur crème foncé »]

bæ̀hɑw- K-d R T-ka [Imprt *ì bhɑw* R T-ka, LoImpfP *-t-íbhɑw-* R T-ka]

Ω a) [intr] be smoky grey, ash-colored ✸ ê. gris, ê. de couleur cendre [K-d R T] [esp. of goats and camels] Ω b) [intr] be ugly ✸ ê. vilain (laid) [R (less common sense)].

bæhɑ́w-æn A-grm = bæhɑ́́w-æn R Ω [partpl, MaSg] smoky grey ✸ gris.

i n èrr bæ̀hɑw-æn T-ka T-md Ω [cpd nm, lit."of ashy neck"] large bustard sp. ✸ grande outarde sp. [ID: mainly *Neotis denhami* (no nape crest) but sometimes also *Ardeotis arabs* (nape crest); cf. √šɣr, √jys].

bæhɑ́w-æt A-grm = bæhɑ́́w-æt R Ω [partpl, FeSg] smoky grey ✸ grise.

bæhɑ́w-nen A-grm R Ω [partpl, Pl] smoky grey ✸ gris.

t-æbbæ̃̀hæw-t K-d R T-ka, Pl *t-æbbæ̀hæw-en* T-ka Ω [nf] smoky grey, ashy color ✸ gris, couleur cendre.

á-bhɑw T-ka, Pl *ì-bhɑw-æn* T-ka Ω [nm] grey or ashy-colored one ✸ (un) gris, (un) de couleur cendre.

t-à̃-bhɑw-t R T-ka, Pl *t-ì-bhɑw-en* R T-ka Ω [nf] grey or ashy-colored one ✸ (une) grise, (une) de couleur cendre [e.g. of goat; for use as botanical term see √bhw 1c below].

Excerpt from Jeffrey Heath's Tamashek dictionary

What is the foundation?

- The “entry” is a hybrid entity
  - It is partly a way to present data on a page
  - It is based on linguistic notions
- For a large-scale interoperable system, one needs to a relatively stable concept
- Building a platform around a notion like the sign or lexeme seems more appropriate
- Display requires extra work, but it’s probably worth it

Information “loss”

- General platforms cannot capture all the particulars of each language
- For LEGO, at least, some of the data in an entry was not properly converted
  - It wasn’t lost but relegated to a “note”
  - Many standards have more or less powerful means to encode such “extra” data

Future projects

- Construct a data model for acceptable lexical resources with
  - A reasonable baseline for publication
  - A “best practice” target
  - Built-in extensibility
  - Don’t avoid existing standards, but don’t let them stand in the way of a good product
  - Work from the linked data paradigm and think about connections to non-lexical data