Using the TEI framework as a possible serialization for LMF

Laurent Romary

Using the TEI framework as a possible serialization for LMF

Laurent Romary, *INRIA & HUB-IDSL*

laurent.romary@inria.fr

RELISH workshop, Aug 4-5, 2010; Nijmegen

Executive summary

- Issue: identifying an “appropriate” serialization for LMF
  - Serialization: mapping a lexical model onto a concrete (computer) representation (field based, XML, etc.)
  - Wide consensus, maintenance, flexibility, cohesion with other standardization activities

- The TEI is an ideal basis for defining standardized XML formats for lexical data
  - TEI as an infrastructure
    - Customization facilities: ODD, classes, pointing mechanisms, etc.
  - TEI as a reference vocabulary
    - Print dictionary chapter (PD)
  - TEI as an application of LMF
- Workplan proposal
  - Defining the ideal LMF compliant subset of the TEI PD chapter
  - Suggesting extensions to the PD chapter
- Convergence…
  - Contribution to making ISO and TEI work closer together

The TEI at a glance

- Started in 1987
- Organized as a consortium: 5 hosts, board, council
- Edition P5 of the guidelines: more than 500 elements covering various text genres and structures
  - Genericity: header, text structure, pointing mechanisms, paragraph level elements, surface entities
  - Precise and flexible documentation
  - Maintenance: 2 releases per year
- Wide community of users: default format for most text-based projects worldwide
  - Cf. papers from Przepiorkowski, or Erjavec at LREC
- And yes, it is XML based… e.g. preconfigured in Oxygen

Intermezzo — an XML tutorial

- XML is about awful angle brackets (serialization)

```
<gramGrp>
<gen>f</gen>
<num>p</num>
</gramGrp>
```

- XML is about beautiful trees (model)
- Issues
  - Specifying structures
  - Providing semantics

Basic concepts of the TEI technical platforms

- A specification language ODD (One Document Does it all)
  - Literate programming (Knuth)
  - Generation of both schemas and documentation
    - DTD, RelaxNG, W3C scemas
    - HTML, pdf, ePub, docx
  - Provides extended customization facilities
  - <equiv>
    - Natural link with ISOCat
- Modules
  - Each schema specification is a combination of internal or external modules
    - E.g. ISO-TEI Feature-Structure module
- Classes (sharedbehavioursorsemantics)
  - Model classes
  - Attribute classes

From ODD to documentation

<gen> (gender) identifies the morphological gender of a lexical item, as given in the dictionary. 9.3.1 Information on Written and Spoken Forms	
Module	dictionaries — 9 Dictionaries
Attributes	att.lexicographic (@expand, @norm, @split, @value, @orig, @location, @mergedIn, @opt)
Used by	model.entryPart model.morphLike

Declaration	Compact to XML format
	element gen
	{
	att.global.attributes,
	att.lexicographic.attributes,
	macro.paraContent
	}

Example	<entry> <form> <orth>pamplemousse</orth> </form> <gramGrp> <pos>noun</pos> <gen>masculine</gen> </gramGrp> </entry>

http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-gen.html

TEI and “dictionaries”

- The TEI Print Dictionary (PD) chapter
  - Initially designed by N. Ide and J. Veronis
  - Accounts for both presentational and editorial (“content”) issues
    - Cf. <entry>, <entryFree>, … and <dictScrap>
  - Based on a hierarchical abstract model (cristals)
    - <form>: for characterising the othographic or phonetic form of the word
      - <orth>, <pron>, etc.
    - <gramGrp>: grammatical features
      - May characterize an entry, a specific form or a specific sense
      - <pos>, <gen>, generic <gram> feature
    - <sense>: iterative and recursive
      - May contains definitions, examples, etymological information, translations, etc.
- Main characteristic (drawback?): +very+ flexible

Examples

<entry>
<form>一乘顯性教</form>
<sense>One of the five divisions made by 圭峰 Guifeng of the Huayan 華嚴 or Avataṃsaka School; v. 五教.</sense>
</entry>

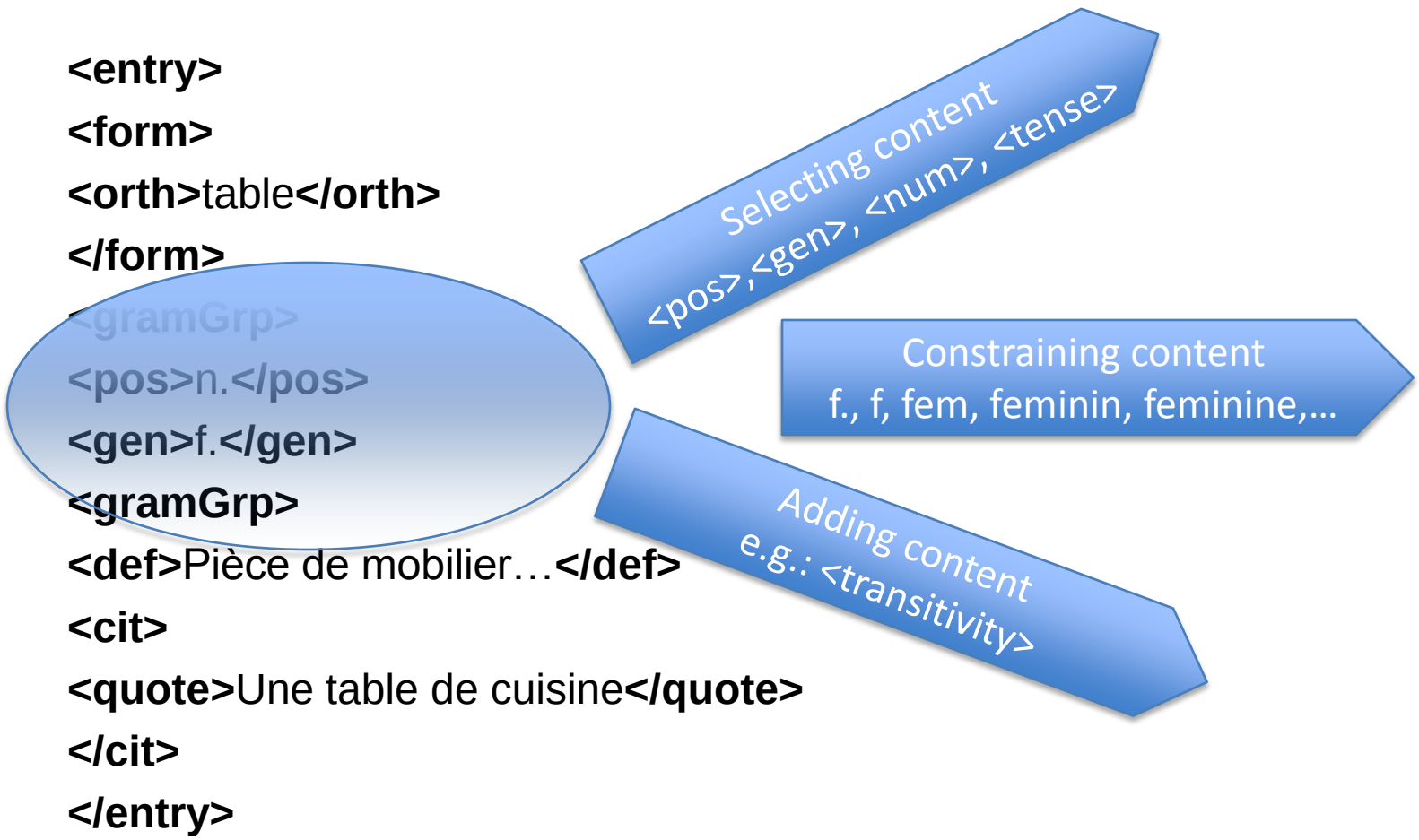
<entry>
<form>眾生不可思議</form>
<sense>
<usg type="dom">術語</usg>
<def>四事不可思議之一。見不可思議條。</def>
<xr>不可思議</xr>
</sense>
</entry>

Source: http://buddhistinformatics.ddbc.edu.tw/glossaries/; thanks to Marcus Bingenheimer

Examples – cont.

```
<entry>
<form type="lemma">
<orth>chat</orth>
</form>
<gramGrp>
<pos>noun</pos>
<gen>masculine</gen>
</gramGrp>
<form type="inflected">
<orth>chat</orth>
<gramGrp>
<number>singular</number>
</gramGrp>
</form>
<form type="inflected">
<orth>chats</orth>
<gramGrp>
<number>plural</number>
</gramGrp>
</form>
</entry>
```

Customizing an entry



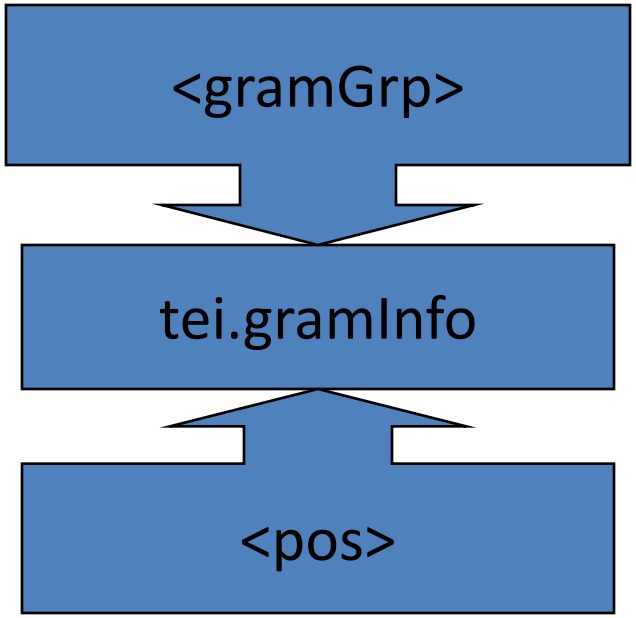
Illustrating classes: tei.gramInfo

- Grammatical information in a dictionary entry
  - E.g.:

```
<entry>
<form>
<orth>luire</orth>
</form>
<gramGrp>
<pos>verb</pos>
<subc>intransitive</subc>
</gramGrp>
</entry>
```

  - Rather homogeneous set of elements
    - <pos>, <gen>, <number>, <case>, etc.
  - May also appear in <form>

Overall picture



Declaring the class: tei.gramInfo

<classSpec xmlns="http://www.tei-c.org/ns/1.0" module="dictionaries-decl"
id="GRAMINFO" type="model" ident="**tei.gramInfo**">

<gloss>grammatical information</gloss>

<desc>groups those elements allowed within a <gi>gramGrp</gi> element in a
dictionary.</desc>

</classSpec>

<pos> belongs to tei.gramInfo

<elementSpec module="dictionaries" id="POS" ident="pos">

<gloss>part of speech</gloss>

<desc>indicates the part of speech assigned to a dictionary headword (noun, verb, adjective, etc.)</desc>

<classes>

<memberOf key="tei.dictionaryParts"/>

<memberOf key="tei.gramInfo"/>

<memberOf key="tei.dictionaries"/>

</classes>

<content> … </content>

<exemplum> … </exemplum>

</elementSpec>

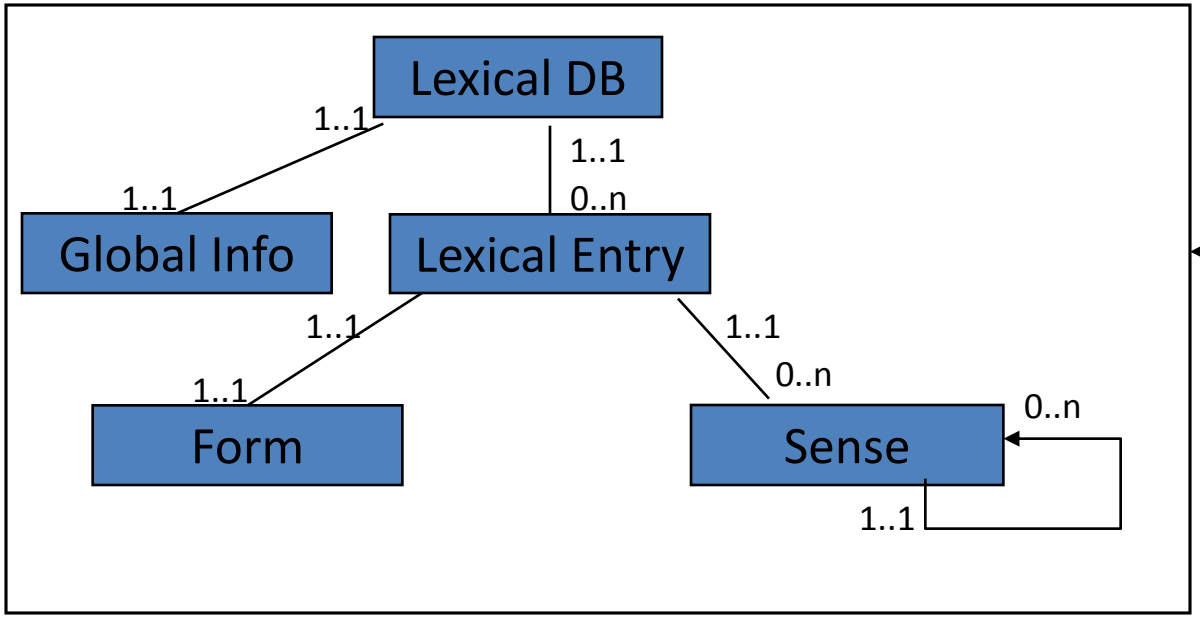
Content model for <gramGrp>

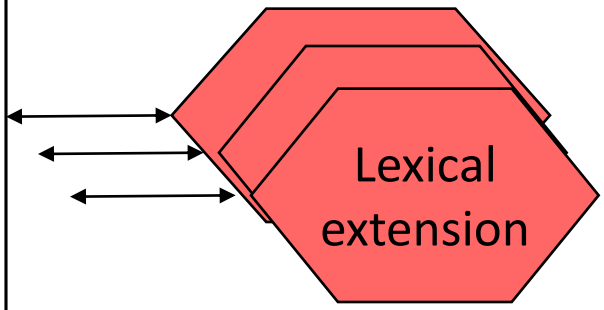
```
<elementSpec module="dictionaries" id="GRAMGRP" ident="gramGrp">
<gloss>grammatical information group</gloss>
<content>
<rng:zeroOrMore
                xmlns:rng="http://relaxng.org/ns/structure/1.0">
<rng:choice>
<rng:text/>
<rng:ref name="tei.phrase"/>
<rng:ref name="tei.inter"/>
<rng:ref name="tei.gramInfo"/>
<rng:ref name="tei.Incl"/>
</rng:choice>
</rng:zeroOrMore>
</content>
    …
</elementSpec>
```

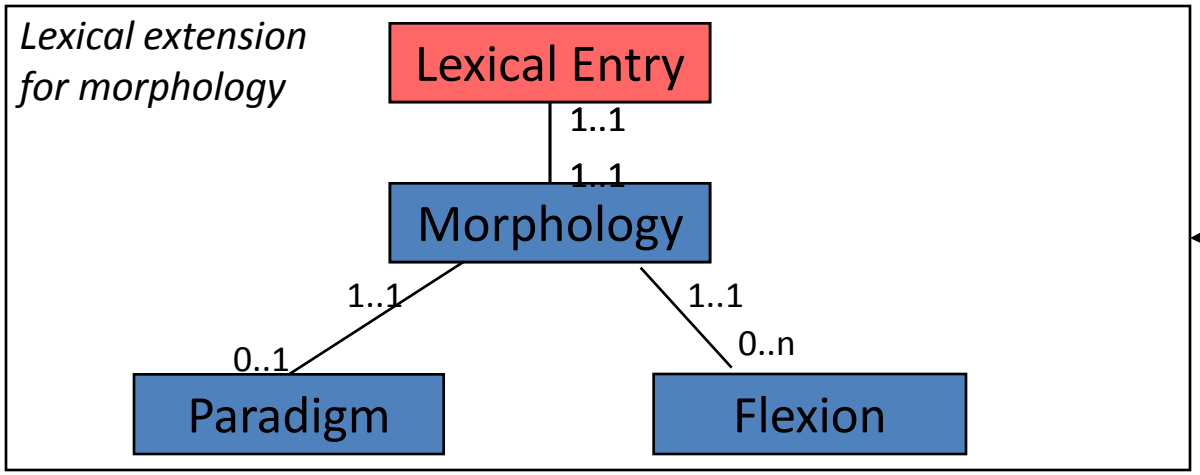
LMF at a glance

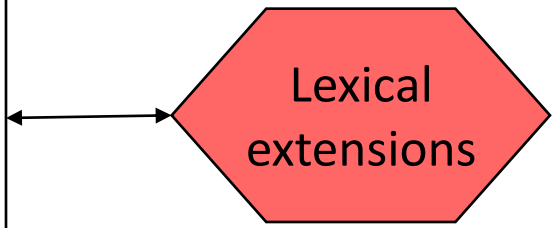
- LMF – Lexical Markup Framework
  - ISO standard 24613 (published Oct. 2008)
  - Edited within ISO committee TC 37/SC 4
- Technical content
  - Focus on provided a core meta-model with extensions
  - Potentially agnostic with regards serialisation
    - Isomorphism => interoperability
  - Default syntax to exemplify its possible use, room for improvement…
- Can the TEI be seen as a conformant implementation of LMF?

LMF architecture — playing Lego

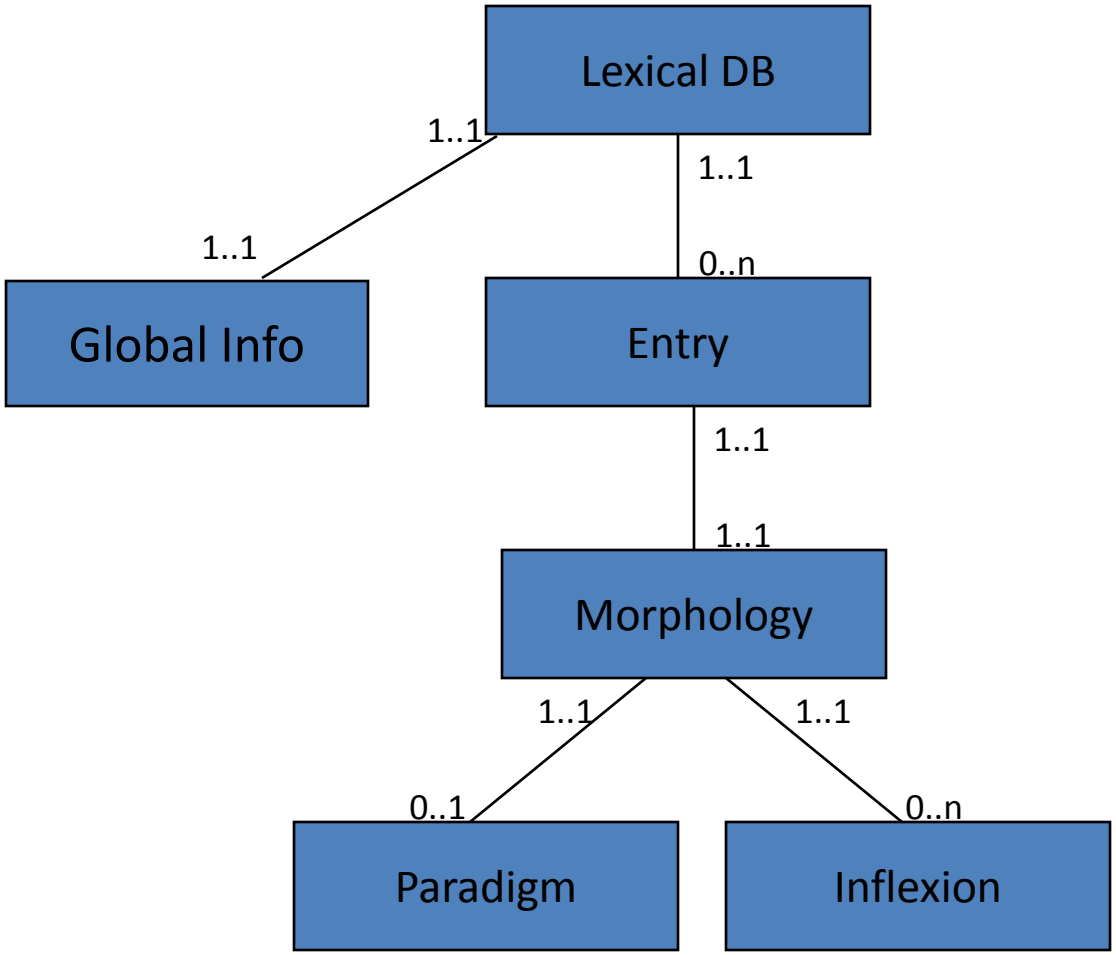




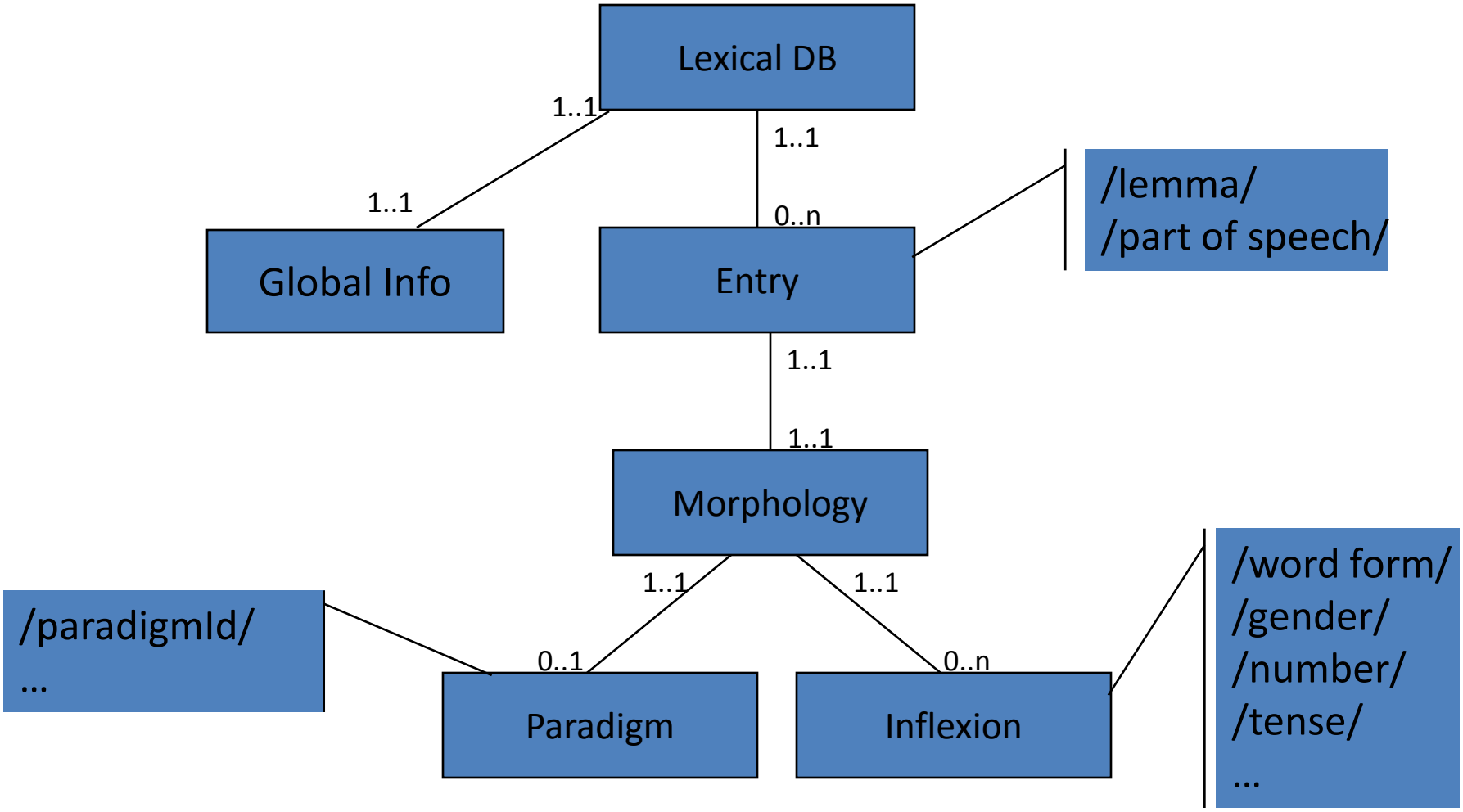




Example: designing a full-form lexicon



Decorating the model



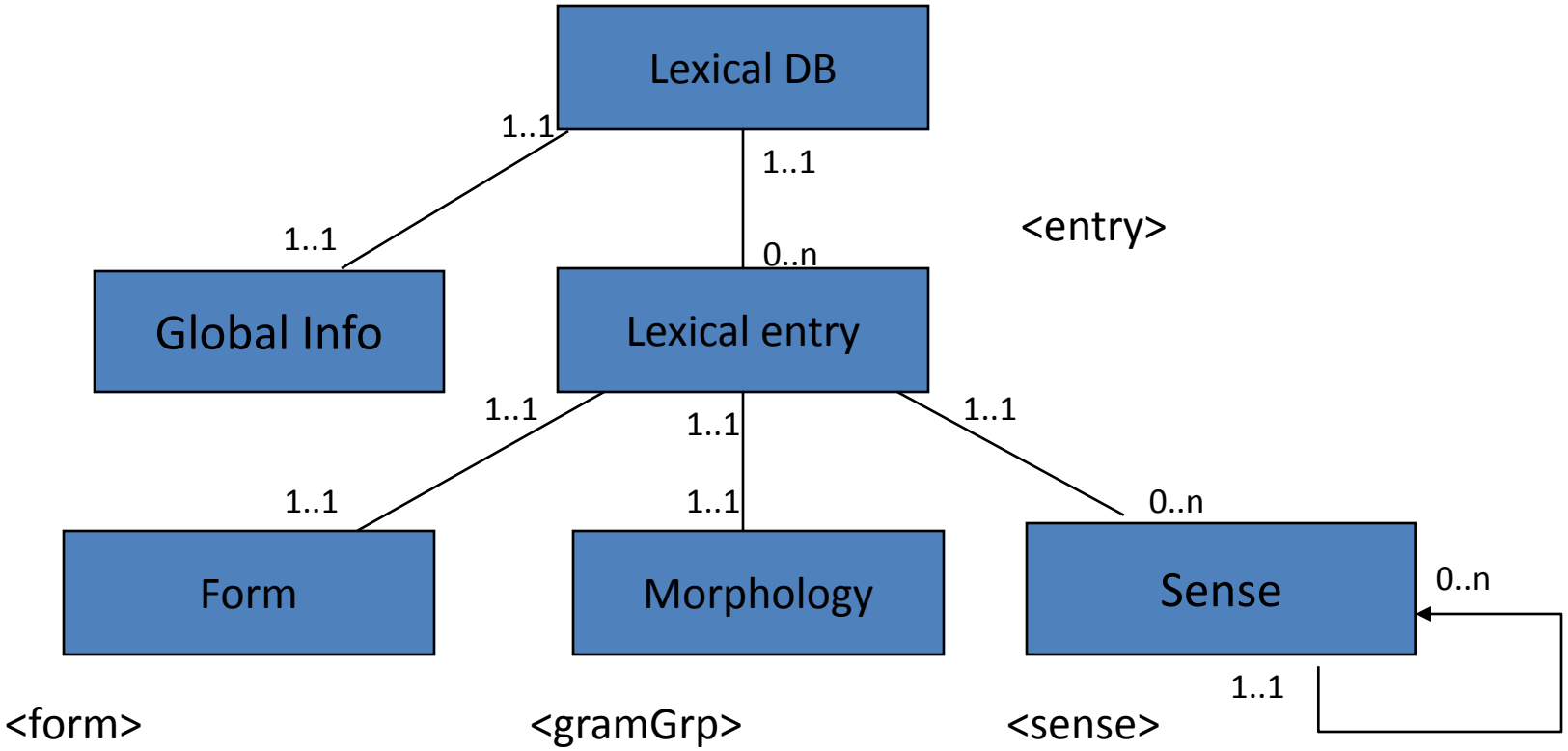
Why is the TEI a good idea for serialising LMF?

- Basic structure already defined
- Provision of additional tags
  - Surface annotation (e.g. names, dates, abbreviations, alternatives)
  - Cf<equiv> equivalences to ISOCat when needed
- Integration of lexical data in a textual macro-structure
  - Creating an edited version of a lexica
  - Grammar books, teaching material, scientific papers
- Interoperability with other lexical sources
  - Community of users: sharing a common culture of TEI tags rather than constantly worrying about mappings
  - Sharing tools: e.g. stylesheets, editors, etc. (cf. Roma)
  - Note: continuity between dictionary and lexical sources

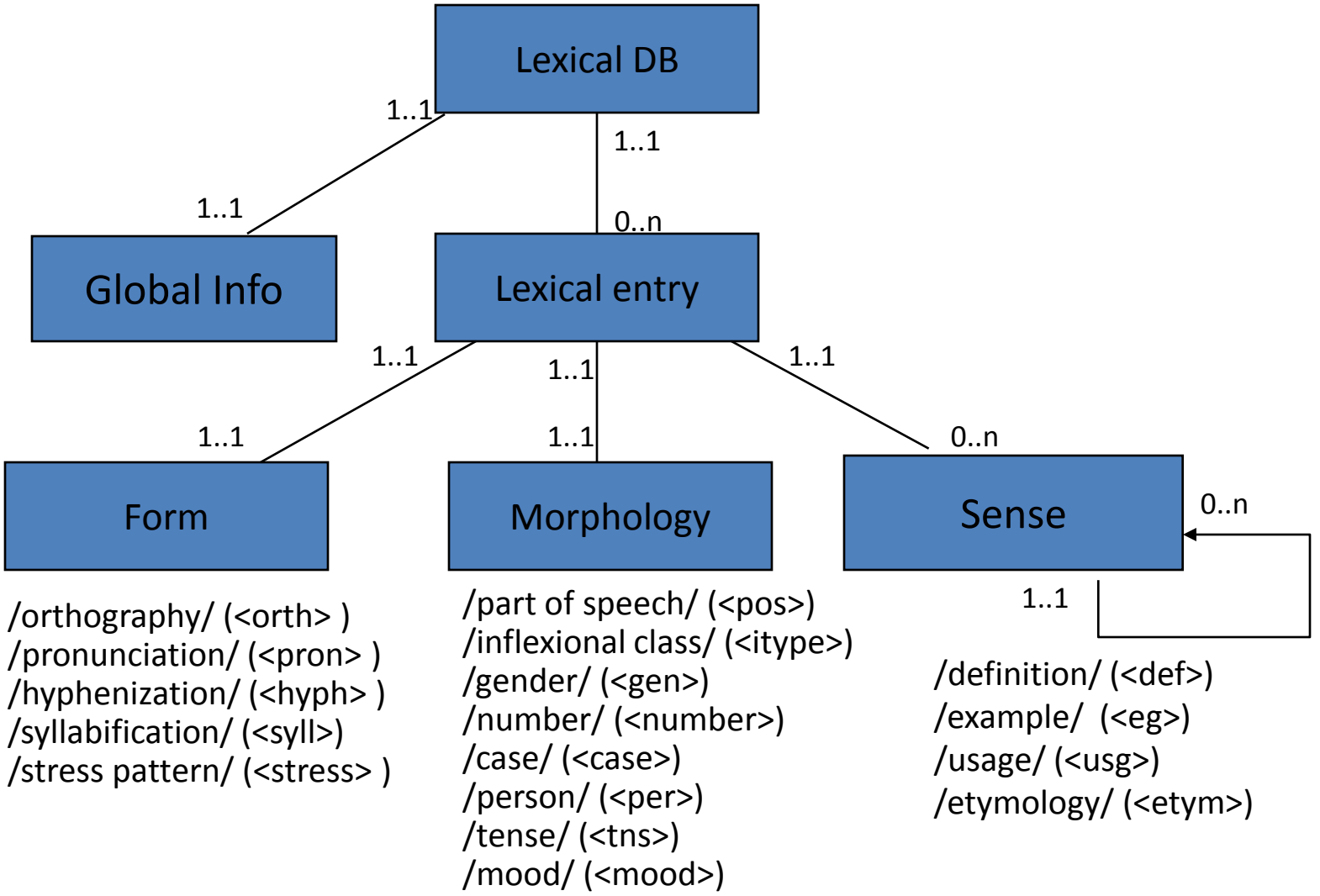
A typical entry

<entry>
<form>
<orth>**demigod**</orth>
<pron>...</pron>
</form>
<gramGrp
<pos>**n**</pos>
</gramGrp>
<sense n='1'>
<sense n='a'>
<def>**a being who is part mortal, part god**</def>
</sense>
<sense n='b'>
<def>**a lesser deity**</def>
</sense>
</sense>
<sense n='2'>
<def>**a godlike person**</def>
</sense>
</entry>

Identifying the meta-model components



Data categories



Customizing the TEI lexical model

- Constraining the TEI model
  - Sub-setting the TEI default dictionary module
  - Providing additional rules (XSLT, Schematron)
  - Constraining possible values
- Complementing the TEI model
  - Defining additional data categories
  - Defining missing LMF extensions as TEI components
    - Make use of the class mechanisms
    - A natural implementation of the LMF extension mechanisms

Towards a joint ISO-TEI activity

- Contributing to convergence, with a pragmatic prespective
- Benefiting from advantages of both sides
  - TEI reactivity and community support
  - ISO stability and international validation
- LMF serialization seen as
  - A subset of the TEI when equivalent construct exist
  - An extension of the TEI for missing constructs (e.g. syntax)
- Some concrete work on the table…
  - Aside activities: specifying LL-LIF in ODD

- Reference
  - L. Romary, “Standardization of the formal representation of lexical information for NLP”
  - http://hal.archives-ouvertes.fr/hal-00436328/fr/