ISO and the TEI

When standards speak to standards

Laurent Romary, Loria-INRIA

Background

- 10 years of experiments around the TEI
  - Multilingual alignment, on-line text server, oral corpus transcription, reference annotation, dictionary mapping, putting our heads together…
  - Council member since 2001
- 4 years of ISO involvement
  - 2000: Project leader of ISO 16642 (Terminology Markup Framework)
  - 2002: ISO/TC 37/SC 4 chair
- Schizophrenia?

One voice, three caps

- TEI council member
  - Achievements of the TEI meta WG and validation on the print dictionary chapter
- ISO/TC 37/SC 4 chair
  - Recent activities related to language resources
    - LMF, DCR…
- TEI host candidate
  - a possible (joint) work plan for the TEI

The TEI council view

- The new power of ODD
  - A powerful, yet easy to learn, specification platform
  - Facilitates the design of the TEI P5 edition
  - Provides means for the user to be part of the game
    - Nicer modular view
    - Class system
    - Extension mechanisms
- Sharing this with you today…

The ISO perspective

- TEI and ISO: a long story
  - Infrastructural standards: SGML, ISO 639-x, ISO 10646
  - ISO 12200 (Martif) as a descendant of the “old” terminology chapter
- What about ISO/TC 37/SC 4?
  - Language resource management
    - Data modeling dedicated to various types of linguistic information (Morphosyntactic, syntactic, discourse level annotation, lexical structures, etc.)
- Showing that the two communities can be useful to one another…

Specifying and extending TEI document models with ODD

The case of the

print dictionary chapter

The Dictionary base tagset

- primarily for printed dictionaries, rather than lexica or dictionary production systems
- **<*entry*>, <*entryFree*>,** and **<*superEntry*>**
- **<*sense*>** and **<*hom*>**
- logical structure vs. typographic fidelity
  - **E.g. <*scrap*>**

Constituents of a dictionary entry

- the form group
- the grammatical-information group
- the definition or translation
- Etymology
- Examples
- usage information
- cross-references to other entries
- notes and related entries

Dictionary components (1)

- **<*form*>** grouping element for one or more of **<*orth*>** **<*pron*>** **<*hyph*>** **<*syll*>** **<*stress*>** etc.
- **<*gramGrp*>** groups specialised grammatical tags **<*gen*>, <*number*>** etc.
- **<*def*>** for definition text, **<*trans*>** for translation
- **<*etym*>** for etymology

Dictionary components (2)

- examples **<*eg*>**
- usage note **<*usg*>**
- label **<*lbl*>**
- related entries **<*re*>** and specialized pointers **<*oRef*>, <*pRef*>** etc

Simple example

```
<entry>
    <form>
        <orth>OATS,</orth>
    </form>
    <gramGrp>
        <pos>n.</pos>
        <number>s.</number>
    </gramGrp>
    <etym>[aten, Sax.]</etym>
    <def>A grain, which in England is
generally given to horses; but in
Scotland supports the people.</def>
    </form>
</entry>
```

Declaring a schema including DI

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0">
   <teiHeader>…</teiHeader>
   <text>
   <body>
   <schemaSpec ident="myTEISchema">
        <moduleRef key="header"/>
        <moduleRef key="core"/>
        <moduleRef key="tei"/>
        <moduleRef key="textstructure"/>
        <moduleRef key="dictionaries"/>
   </schemaSpec>
   </body>
   </text>
</TEI>
```

The structure of the Print Dictionary chapter

Classes in ODD

- Groups together all elements with the same role in the TEI architecture
  - Same syntactic behaviour
    - The elements in the class will appear in the same content models
  - Semantic similarity
    - The class defines a group of elements belonging to the same family of concepts
- Principle:
  - elements declare themselves as belonging to a class

Main classes in DI

- tei.dictionaries
  - Groups all elements defined in DI
  - Declares general attributes: `expand`, `norm`, `split`, `value`, `orig`, `location`, `mergedin`, `opt`
- tei.dictionaryParts
  - Groups all elements defined in DI
- tei.dictionaryTopLevel
  - Elements occuring at entry level

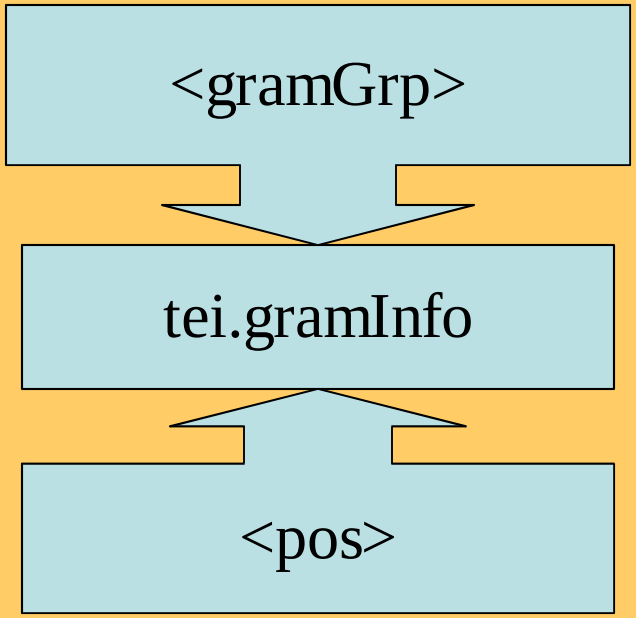
Specialized classes: tei.gramInfo

- 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>
```

Specification of a dictionary entry

The <entry> element

- Content model
  - Choice of:
    - hom, sense
    - tei.dictionaryTopLevel, tei.Incl
- Anchoring:
  - Belongs to tei.entries (+ entryFree + superEntry)

Declaring <entry>

```
<elementSpec module="dictionaries" id="ENTRY" ident="entry">
    <classes>
        <memberOf key="tei.entries"/>
    </classes>
    <content>
        <rng:oneOrMore
          xmlns:rng="http://relaxng.org/ns/structure/1.0">
            <rng:choice>
                <rng:ref name="hom"/>
                <rng:ref name="sense"/>
                <rng:ref name="tei.dictionaryTopLevel"/>
                <rng:ref name="tei.Incl"/>
            </rng:choice>
        </rng:oneOrMore>
    </content>
    <desc>contains a reasonably well-structured dictionary
   entry.</desc>
</elementSpec>
```

Toying with ODD

(collaboration with Susanne Salmon-Alt,
Susanne.Salmon-Alt@atilf.fr)

Two testbeds

- Applying “standard” constraints to the <gen> (grammatical gender) element

- Introducing a complex element to deal with historical notes (<diachrony>)

Constraining the values of <gen>

- Basic content model for <gen>
  - Cf. gen.odd

```
<content>
  <rng:ref
        xmlns:rng="http://relaxng.org/ns/structure/1.0"
        name="macro.paraContent"/>
</content>
```

- Two aspects
  - Changing the content model
    - Account for the local editorial practices

```
<gen>m</gen>
```

  - Modify the `norm` attribute (tei.dictionaries)
    - Relate to some kind of standard set of values

```
<gen norm="masculine">m</gen>
```

The underlying picture

tei.dictionaries

```
<classSpec
  module="dictionaries-decl"
  id="DIGLOBAL »
  type="atts »
  ident="tei.dictionaries">
  <attList>
    <attDef ident="norm">
      <datatype>
        <rng:text
xmlns:rng="http://relaxng.org/ns/
structure/1.0"/>
      </datatype>
  <desc>gives a normalized form
of information given by the
source text in a non-normalized
form</desc>
    </attDef>
  </attList>
</classSpec>
```

gen.odd

```
<elementSpec
 module="dictionaries”
 id="GEN »
 usage="rec" ident="gen">
  <equiv
    name="grammaticalGender”
 uri="http://www.tc37sc4.org"/>
  <gloss>gender</gloss>
  <classes>
    <memberOf
    key="tei.dictionaryParts"/>
      <memberOf
    key="tei.morphInfo"/>
      <memberOf
    key="tei.dictionaries"/>
  </classes>
  <content>
  </content>
</elementSpec>
```

mySchema.odd

```
<schema>
  <moduleRef key=“dictionaries”/>
  <elementSpec
    module=“dictionaries””
    ident=“gen“
    mode=“change”>
    <content>
    </content>
    <attList>
    </attList>
  </elementSpec>
</schema>
```

Constraining the values of <gen>

```
<elementSpec xmlns="http://www.tei-c.org/ns/1.0"
   ident="gen" mode="change">
        <content>
            <valList>
                <valItem ident="m"/>
                <valItem ident="f"/>
            </valList>
        </content>
</elementSpec>
```

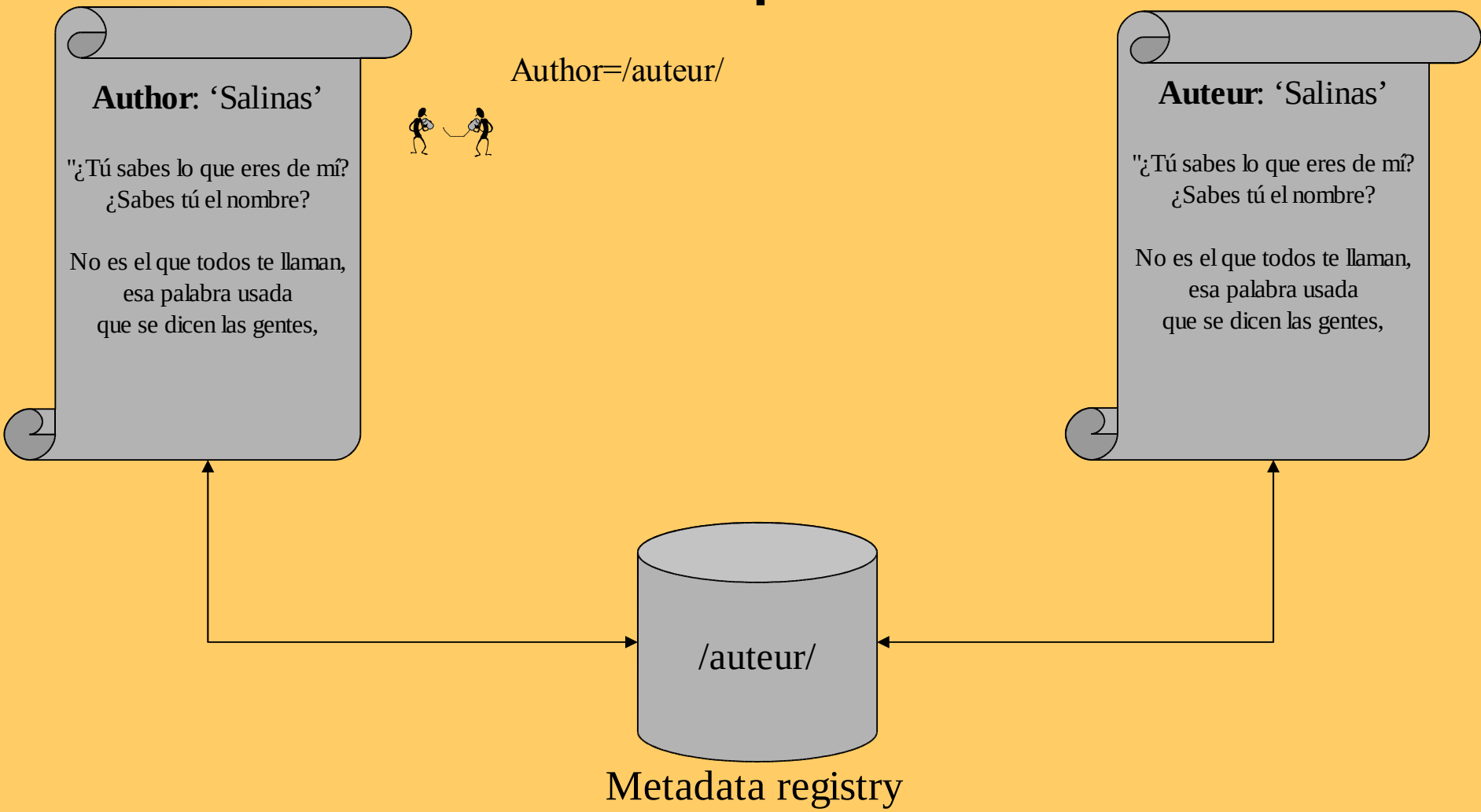
Example

```
<entry>
   <form>
        <orth>pamplemousse</orth>
   </form>
   <gramGrp>
        <pos>noun</pos>
        <gen>m</gen>
   </gramGrp>
</entry>
```

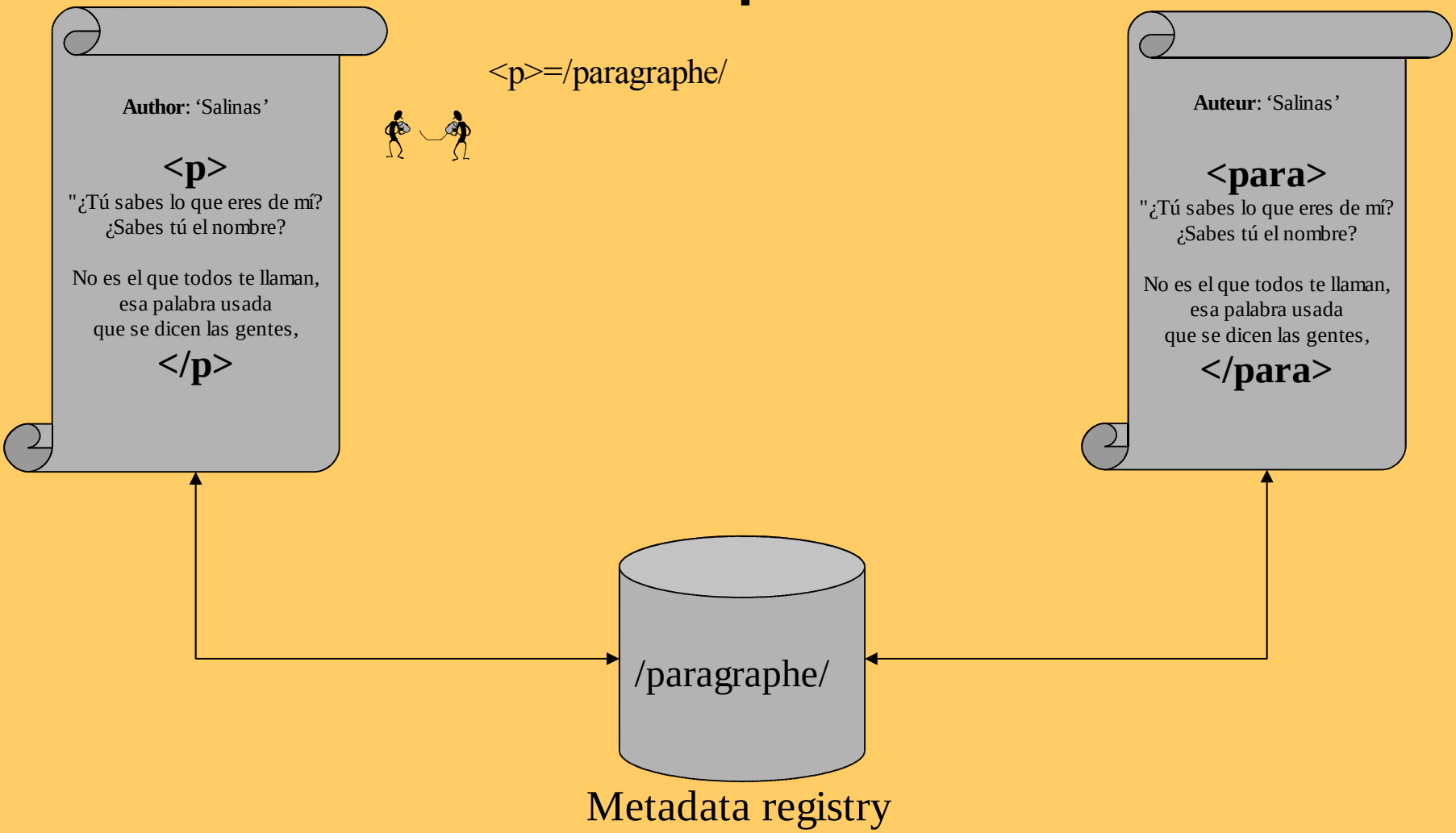
Before we go any further…

- Which normative reference for the values of grammatical gender?
  - Not an issue specific to dictionary design
    - Cf. linguistic annotation at large (e.g. POS tagging)
  - Not an issue specific to the TEI community
    - Such values and their semantics should be defined independantly of any specific tagset
- A possible answer: the ISO/TC 37 data category registry (DCR)

Meta data for content description



Meta data for structural description



Data Category

- Definition
  - Elementary descriptor used in a linguistic description or annotation scheme
- Example
  - Placeholders: /part of speech/, /grammatical gender/
  - Values: /feminine/, /plural/, /ablative case/
- Role
  - Characterize an annotation scheme (specification)
  - Identify its linguistic coverage (documentation, variation of scope)
- Market place for annotation scheme designers
  - Wide choice of possible descriptors
    - Provision of the semantics and condition of use
  - On-going experiment: http://syntax.loria.fr

Technical background

- ISO 11179 (ISO JTC 1/SC 32): metadata
  - Management of data categories
- OWL (W3C Sem. Web activity): ontology
  - Hierarchies and constraints on data categories
    - fr: /noun/ => /grammatical gender/ + /grammatical number/
- ISO 16642 (ISO TC 37/SC 3): terminology
  - Multilingual issues
    - Multiple names, language specific values

Documenting data categories

Entry Identifier: grammatical gender

Profile: morpho-syntax

Definition (fr): Catégorie grammaticale reposant, selon les langues et les systèmes, sur la distinction naturelle entre les sexes ou sur des critères formels (Source: TLFi)

Definition (en): Grammatical category… (Source: TLFi (Trad.))

Conceptual Domain: {/feminine/, /masculine/, /neuter/}

Object Language: fr
Name: genre
Conceptual Domain: {/feminine/, /masculine/}

Object Language: en
Name: gender
Name: grammatical gender

Object Language: de
Name: Geschlecht
Name: Genus
Conceptual Domain: {/feminine/, /masculine/, /neuter/}

TEI goes ISO with `equiv`

```
<elementSpec xmlns="http://www.tei-c.org/ns/1.0" ident="gen" mode="change">
        <content>…</content>
         <attList>
            <attDef ident="norm" mode="change">
                <valList>
                    <valItem ident="masculine">
                        <equiv
                          name="masculine"
                          uri="http://www.tc37sc4.org"/>
                    </valItem>
                    <valItem ident="feminine">
                        <equiv
                          name="feminine"
                          uri="http://www.tc37sc4.org"/>
                    </valItem>
                </valList>
            </attDef>
        </attList>
</elementSpec>
```

Going further…

- Should not we say that <gen> means /grammaticalGender/?
- We actually did it:

```
<elementSpec module="dictionaries” id="GEN” usage="rec"
   ident="gen">
   <equiv
        name="grammaticalGender”
        uri="http://www.tc37sc4.org"/>
   <gloss>gender</gloss>
   <classes> … </classes>
   <content> … </content>
</elementSpec>
```

Etymology vs. diachrony

Two types of organization in the
Trésor de la Langue Française

“Pamplemousse”:Etymology

Empr. au néerl. pompelmoes, fém., au sens 1 a, qui est prob. comp.de pompel «gros, enflé» et de limoes «citron» (Boulan, p.148; König, pp.159-160). Apparaît d'abord dans des textes fr. qui le donnent comme mot néerl.: 1665 pompelmoes (J. Le Carpentier, L'Ambassade de la Compagnie orientale des Provinces Unies... [trad. d'un ouvrage néerl.], II, p.88 ds Arv.); 1666 pompelmous (M. Thévenot, Relation de divers voyages curieux... t.3 ds König).

“Pamplemousse”: diachrony

1. a) 1677 pampelmous «fruit comestible, peu juteux, d'un arbre épineux originaire des îles de l'océan Indien» (Fr. de L'Estra, Relation ou Journal d'un voyage fait aux Indes Orientales, p.107 ds König);

   ca 1685 pamplemousse (J. Bouvet, Voiage de Siam, éd. J. C. Gatty, Leiden, 1963, p.68);

b) 1772 «arbre qui produit ce fruit» (Chambors, Dissertation sur le jardinage de l'Orient, p.77 ds Fr. mod. t.6, 1938, p.255);

2. a) 1946 «fruit du Citrus paradisi de grande taille, jaune, de goût acide» (J. Brichet, Pamplemousse ou Pomelo ... ds Fruits d'outre-mer, no 10, p.297 d'apr. M. Chauvet ds Journal d'agric. traditionnelle et de bot.appl., t.27, 1980, p.67);

b) 1962 «arbre qui produit ce fruit» (Rob.).

Main issues

- An entry-like organization of historical notes
  - Senses and sub-senses
  - Information related to the form of the word
    - E.g.: *pampelmous*
  - Sense related information:
    - Gloss, definition, collocation, usage information, etc.
    - E.g.: *fruit comestible, peu juteux, d'un arbre épineux originaire des îles de l'océan Indien*
- Some specific information
  - Testimonial date
    - E.g.: *ca 1685*
  - Bibliographical sources
    - E.g.: J. Bouvet, Voiage de Siam, éd. J. C. Gatty, Leiden, 1963, p.68

ODDifying this

- Declaring a <diachrony> element
  - Inspired from <re>
    - Entry-like structure
    - May appear in an entry
  - Add the necessary features for:
    - Dates
    - Bibliographical descriptions
- Change <sense> to acount for dates and bibliographical descriptions

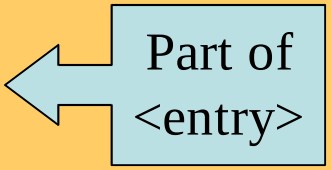
And now for the concrete stuff…

Schema declaration with extensions

```
<schemaSpec ident="romary">
    <moduleRef key="header"/>
    <moduleRef key="core"/>
    <moduleRef key="tei"/>
    <moduleRef key="textstructure"/>
    <moduleRef key="dictionaries"/>

    <elementSpec
          module="dictionaries" id="DIACHRONY”
          ident="diachrony" mode="add">
          …
    </elementSpec>

    <elementSpec module="dictionaries" id="SENSE”
          ident="sense" mode="change">
    </elementSpec>

</schemaSpec>
```

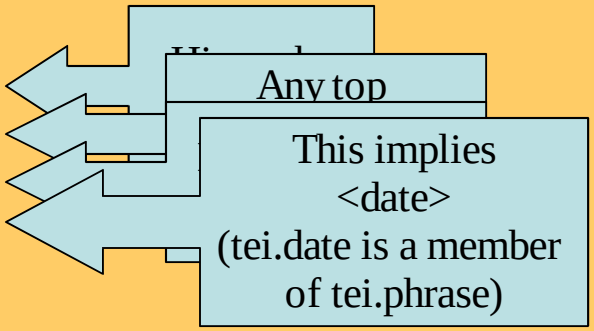
<diachrony>: behaviour

```
<elementSpec
   module="dictionaries" id="DIACHRONY”
   ident="diachrony" mode="add">
   <classes>
        <memberOf key="tei.dictionaryTopLevel"/>
        <memberOf key="tei.dictionaryParts"/>
        <memberOf key="tei.dictionaries"/>
   </classes>
   <content>
        …
   </content>
</elementSpec>
```



<diachrony>: content model

```
<elementSpec
    module="dictionaries" id="DIACHRONY”
    usage="opt" ident="diachrony" mode="add">
    <classes>…</classes>
    <content>
       <rng:zeroOrMore
          xmlns:rng="http://relaxng.org/ns/structure/1.0">
       <rng:choice>
          <rng:text/>
          <rng:ref name="sense"/>
          <rng:ref name="tei.dictionaryTopLevel"/>
          <rng:ref name="tei.bibl"/>
          <rng:ref name="tei.phrase"/>
          <rng:ref name="tei.Incl"/>
       </rng:choice>
       </rng:zeroOrMore>
    </content>
</elementSpec>
```



<sense>: old content model (sense.odd)

```
<elementSpec module="dictionaries" id="SENSE" ident="sense">
   …
   <content>
        <rng:zeroOrMore
           xmlns:rng="http://relaxng.org/ns/structure/1.0">
               <rng:choice>
                      <rng:text/>
                      <rng:ref name="sense"/>
                      <rng:ref name="tei.dictionaryTopLevel"/>
                      <rng:ref name="tei.phrase"/>
                      <rng:ref name="tei.Incl"/>
               </rng:choice>
        </rng:zeroOrMore>
   </content>
   …
</elementSpec>
```

<sense>: adding bibliography

```
<elementSpec module="dictionaries" id="SENSE" ident="sense"
   mode="change">
   <content>
        <rng:zeroOrMore
               xmlns:rng="http://relaxng.org/ns/structure/1.0">
               <rng:choice>
                    <rng:text/>
                    <rng:ref name="sense"/>
                    <rng:ref name="tei.dictionaryTopLevel"/>
                    <rng:ref name="tei.bibl"/>
                    <rng:ref name="tei.phrase"/>
                    <rng:ref name="tei.Incl"/>
                </rng:choice>
           </rng:zeroOrMore>
       </content>
</elementSpec>
```

Applying this to an entry…

```
<entry>
   <form>
      <orth>pamplemousse</orth>
   </form>
   <diachrony>
      <sense n="1">
         <sense n="a">
            <date>1677</date>
            <form>
               <orth>pampelmous</orth>
            </form>
            <def>fruit comestible, peu juteux, d'un arbre épineux originaire des îles
   de l'océan Indien</def>
            <bibl>Fr. de L'Estra, Relation ou Journal d'un voyage fait aux Indes
   Orientales, p.107 ds König</bibl>
         <sense>
            <date>1685</date>
            <form>
               <orth>pamplemousse </orth>
            </form>
            <bibl>J. Bouvet, Voiage de Siam, éd. J. C. Gatty, Leiden, 1963,
   p.68</bibl>
         </sense>
      </sense>
      <sense n="b">
         …
      </sense>
   </diachrony>
```

Where are we
going from here?

Where we could work together

- Joint TEI-ISO working group on feature structures
- On-going work on LMF (Lexical Markup Framework)
- TEI header and other metadata initiatives
- SO working group and LAF (linguistic annotation framework)
- Link between the TEI tagset and the ISO/TC 37 Data Category Registry

A good start…

- Joint working group on Feature Structure established in early 2003
  - FSR: close to its termination
    - New P5 chapter
    - FSR at DIS stage in ISO/TC 37/SC 4
  - Ongoing discussions on joint proposal for FS declaration;

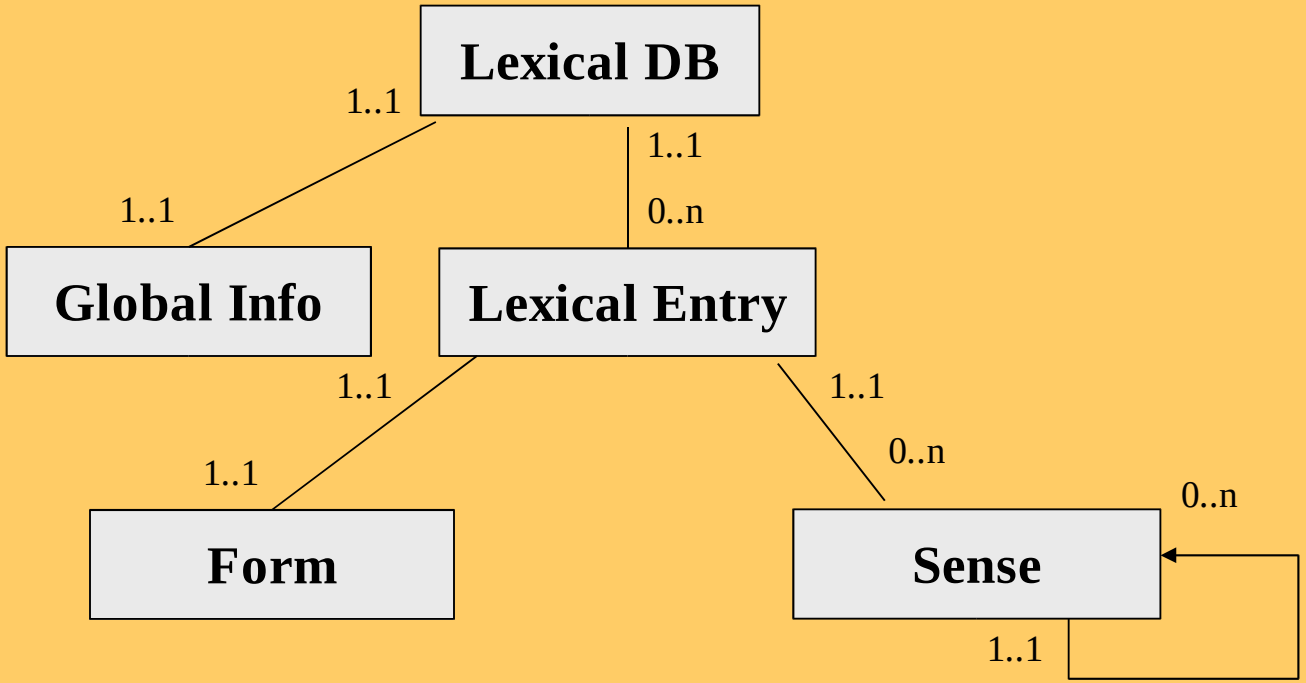
Where we could work together

- Joint TEI-ISO working group on feature structures
- On-going work on LMF (Lexical Markup Framework)
- TEI header and other metadata initiatives
- SO working group and LAF (linguistic annotation framework)
- Link between the TEI tagset and the ISO/TC 37 Data Category Registry

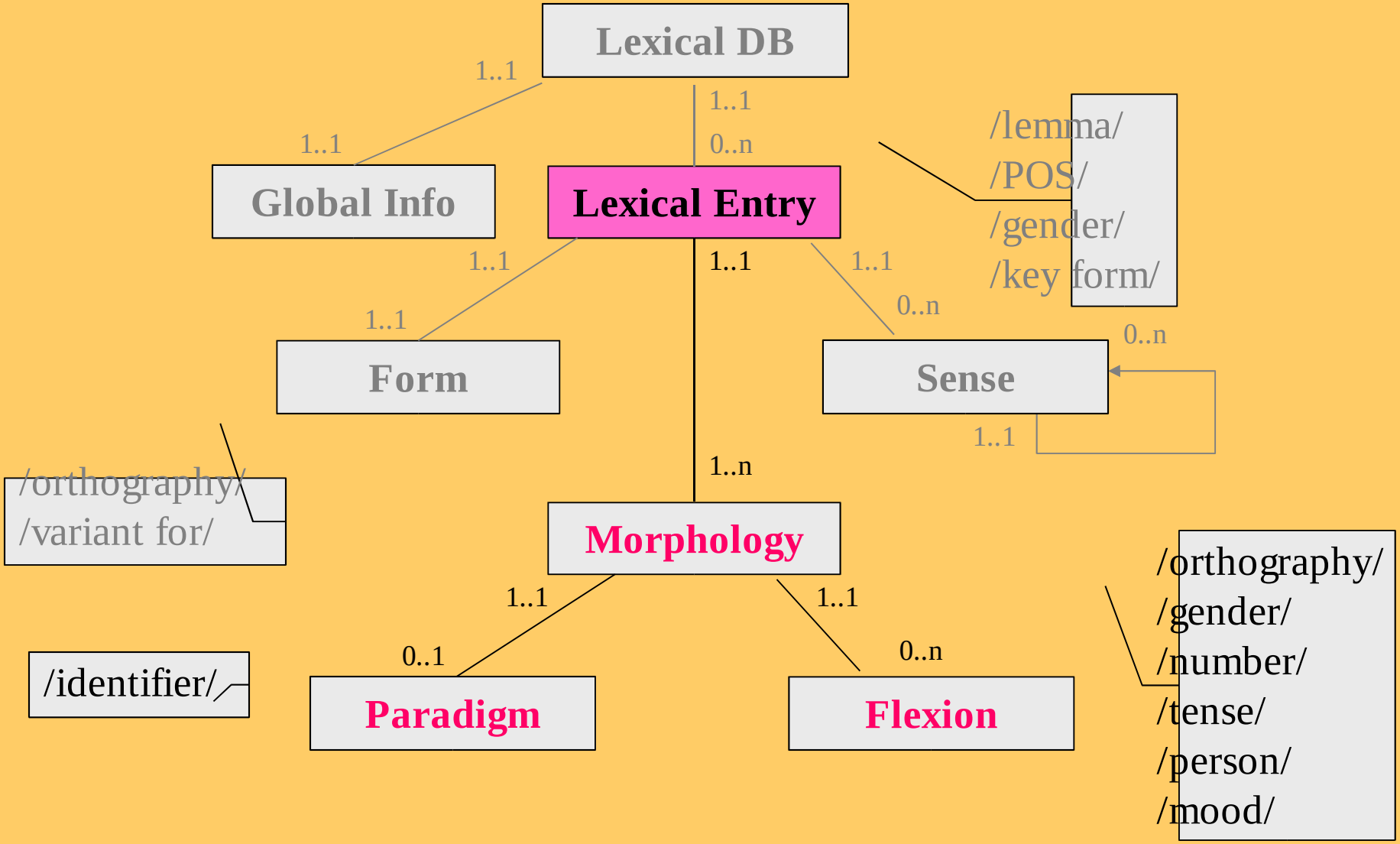
Lexical structures and print dictionaries

- E.g. print dictionary chapter as an LMF application;
  - General organizations are very similar
  - More modeling tasks are required
    - Deep structure of etymology
      - » Etymons and information attached to them
    - Links with NLP lexical
      - » What constrained should we apply to a DI entry to be mappable back and forth to a morphological lexicon

LMF – core metamodel



Core metamodel with morphological extension



```
<struct type="lexical entry">
   <feat type="lemma">profesor</feat>
   <feat type="grammatical category">common noun</feat>
   <struct type="morphology">
      <struct type="flexion">
         <feat type="word form">profesor</feat>
         <brack>
            <feat type="tag set">SRPcor</feat>
            <feat type="POS tag">ms1v</feat>
         </brack>
         <feat type="grammatical gender">masculine</feat>
         <feat type="grammatical number">singular</feat>
         <feat type="grammatical case">nominative</feat>
         <feat type="animate">yes</feat>
      </struct>
      <struct type="flexion">
         <feat type="word form">profesorom</feat>
         <brack>
            <feat type="tag set">SRPcor</feat>
            <feat type="POS tag">ms6v</feat>
         </brack>
         <feat type="grammatical gender">masculine</feat>
         <feat type="grammatical number">singular</feat>
         <feat type="grammatical case">instrumental</feat>
         <feat type="animate">yes</feat>
      </struct>
      …
   </struct>
```

Where we could work together

- Joint TEI-ISO working group on feature structures
- On-going work on LMF (Lexical Markup Framework)
- TEI header and other metadata initiatives
- SO working group and LAF (linguistic annotation framework)
- Link between the TEI tagset and the ISO/TC 37 Data Category Registry

Relating meta-data activities

- General purpose initiatives
  - Dublin Core
- Domain specific initiatives
  - OLAC: Open Language Archive Community
  - IMDI: Isle Metadata Initiative
  - *MARC
- …and the TEI

A concrete example

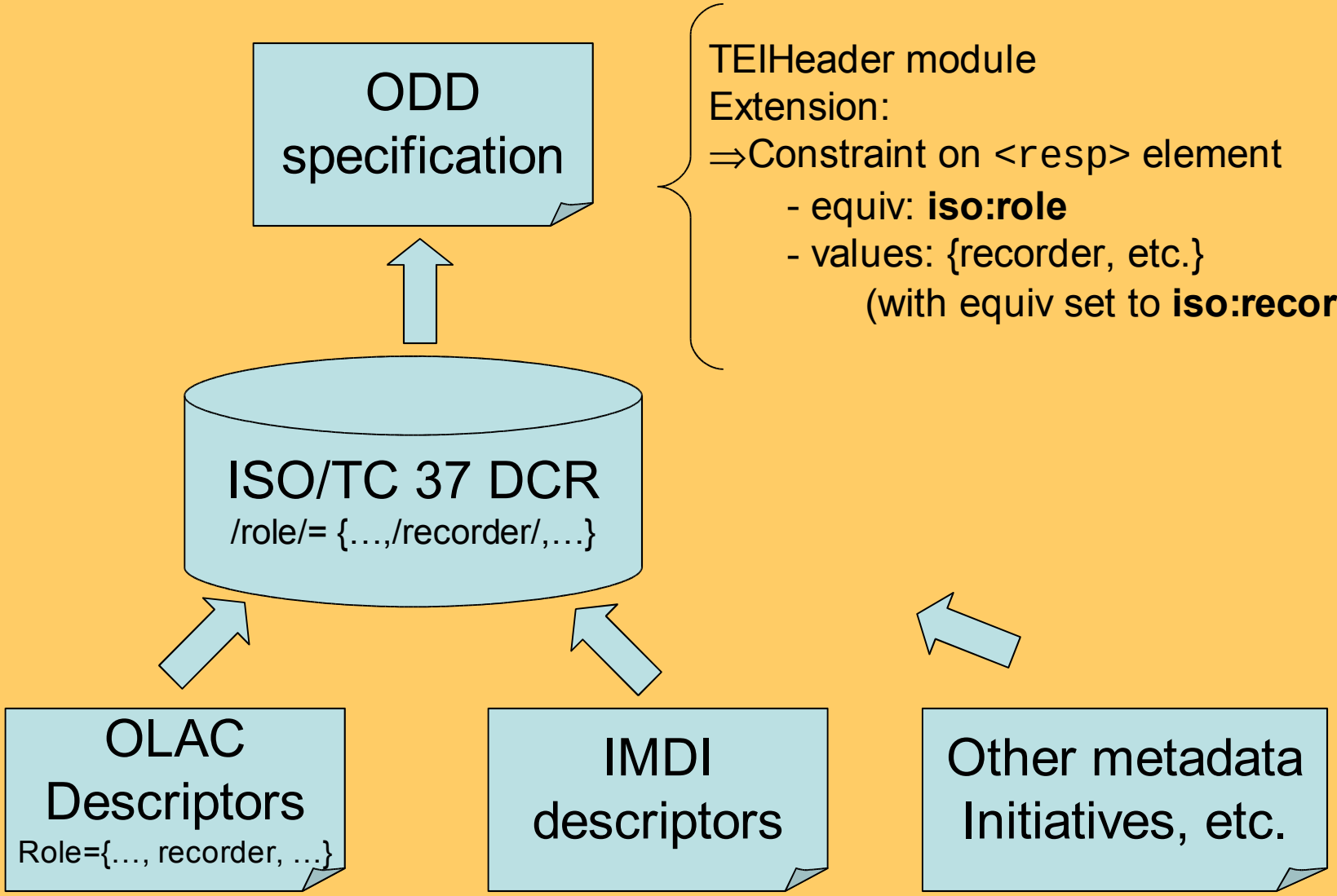
- Responsability statements in the TEI header

```
<respStmt>
   <resp>did the very nice recording</resp>
   <name>Lou Wittern</name>
</respStmt>
```

- The OLAC role typology
  - depositor, compiler, editor, researcher, sponsor, editor, researcher, annotator, data inputter, developer, sponsor, author, translator, Interpreter, interviewer, responder, participant, performer, signer, **recorder**, research participant, singer, speaker
  - Note: clear semantics associated to each value
- How do relate these?

```
<respStmt>
   <resp>recorder</resp>
   <name>Lou Wittern</name>
</respStmt>
```

From OLAC to the TEI



Where we could work together

- Joint TEI-ISO working group on feature structures
- On-going work on LMF (Lexical Markup Framework)
- TEI header and other metadata initiatives
- SO working group and LAF (linguistic annotation framework)
- Link between the TEI tagset and the ISO/TC 37 Data Category Registry
- And much more…
  - SC 4 provides the semantic background in linguistic structures to be used in TEI contexts
  - TEI as the default model for SC 4 activities