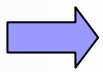
DataCategory Registry, LMF, TMF etc. What is the situation?

Gil FRANCOPOULO
INRIA-loria
(co-editor for ISO-LMF & convenior for morpho-syntactic DatCat Registry)
gil.francopoulo@wanadoo.fr



not a one way presentation, interrupt when you want

Introduction

- One of the crucial aspects impacting HLT is the need to optimize the production, maintenance and extension of linguistic resources. A very frequent mentioned need is: « how to merge resources? ».
- A second crucial aspect involves optimizing the process leading to integration of linguistic resources in applications
- Scope: all NLP based applications, all natural languages.

Family

- A family of standards is being defined within ISO TC37
- Deal with NLP and Terminology Management
- There is not a huge ISO standard but a family of goal oriented ISO standards with common principles
- Various degree of progress: some standards are already published and some are still in their « infancy »

Common principles

- In these complex domains, it's not possible to anticipate everything
- Hypothesis: common best practices can be specified as structural elements
- These structural elements are decorated as the user convenience by ISO 12620 datcats
- And a registry of datcats is proposed
- Key basic standards:
  UNICODE (ISO), UML (OMG), XML (W3C+ISO)

- LMF = Lexical Markup Framework ISO 24613
- TMF = Terminology Markup Framework ISO 16642
- MAF = Morpho-syntactic Annotation Framework ISO 24611
- SynAF = Syntactic Annotation Framework
- Word Segmentation ISO 23679
- DatCat management ISO 12620 revision

High level standards	**TMF (published)** **LMF (cd)** **MAF (wd)** **SynAF (nwip)** **WordSegmentation (nwip)**	**Structural elements. Ex: a word can have various forms**
Intermediate level	**nothing**	**Ex: in English, a noun has a number**
Low level	**DatCat management (draft for 12620 revision)** **With DatCat Registry (starting)**	**Ex-1: the number in English is /singular/ or /plural/** **Ex-2: /singular/ and /plural/ are constants**

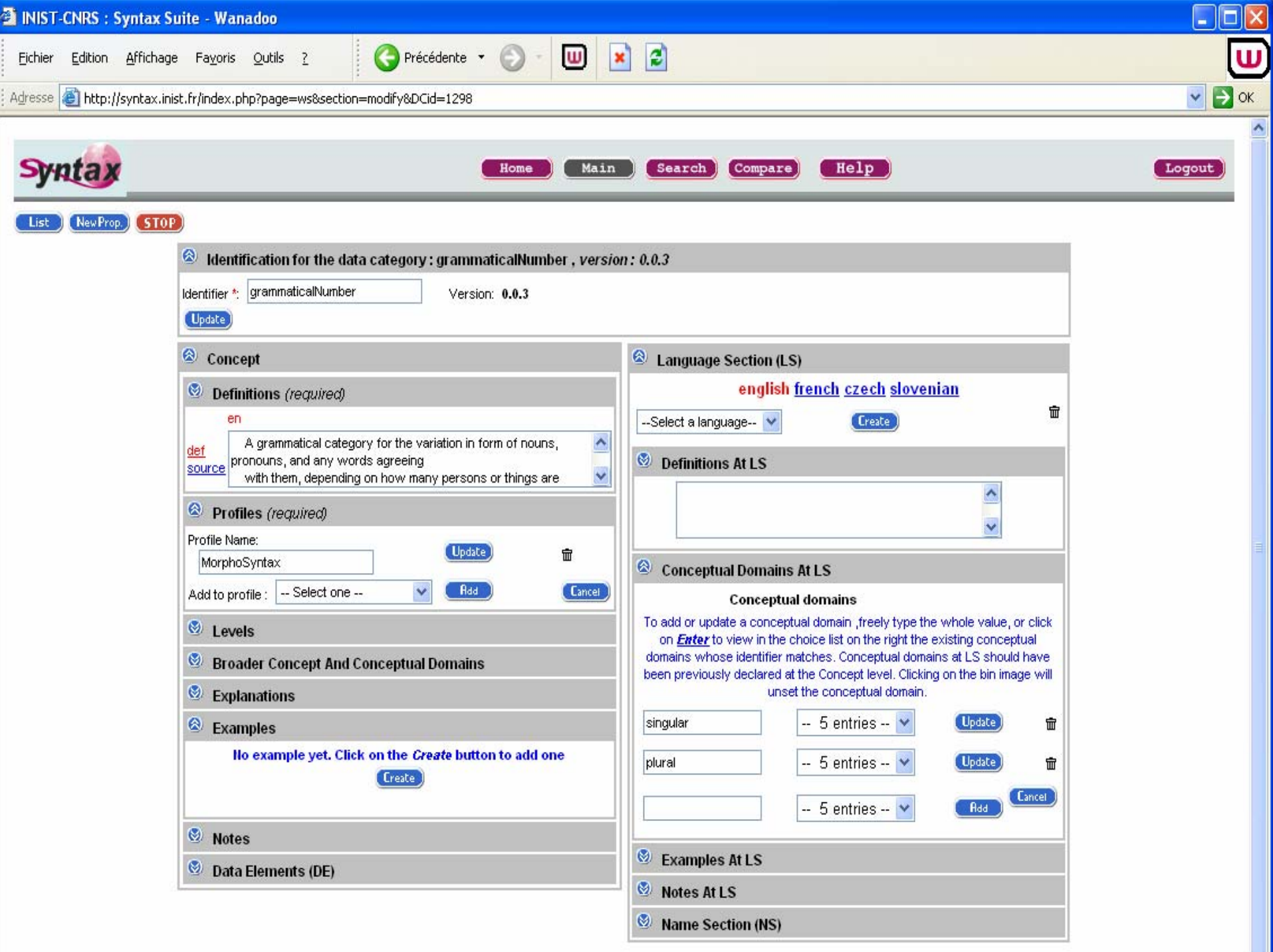
ISO DataCategories

- ISO 12620 revision, work in progress done by Laurent Romary (France)
- A DatCat registry is on the way to be defined according to various subsets:
  - terminology: Sue Ellen Wright (USA)
  - metadata: Peter Wittenburg (Nederland)
  - morpho-syntax: Gil Francopoulo (France)
  - semantics: Koiti Hasida (Japan)
- These conveniors work with the help of a team
- A central database keeps track of the current values

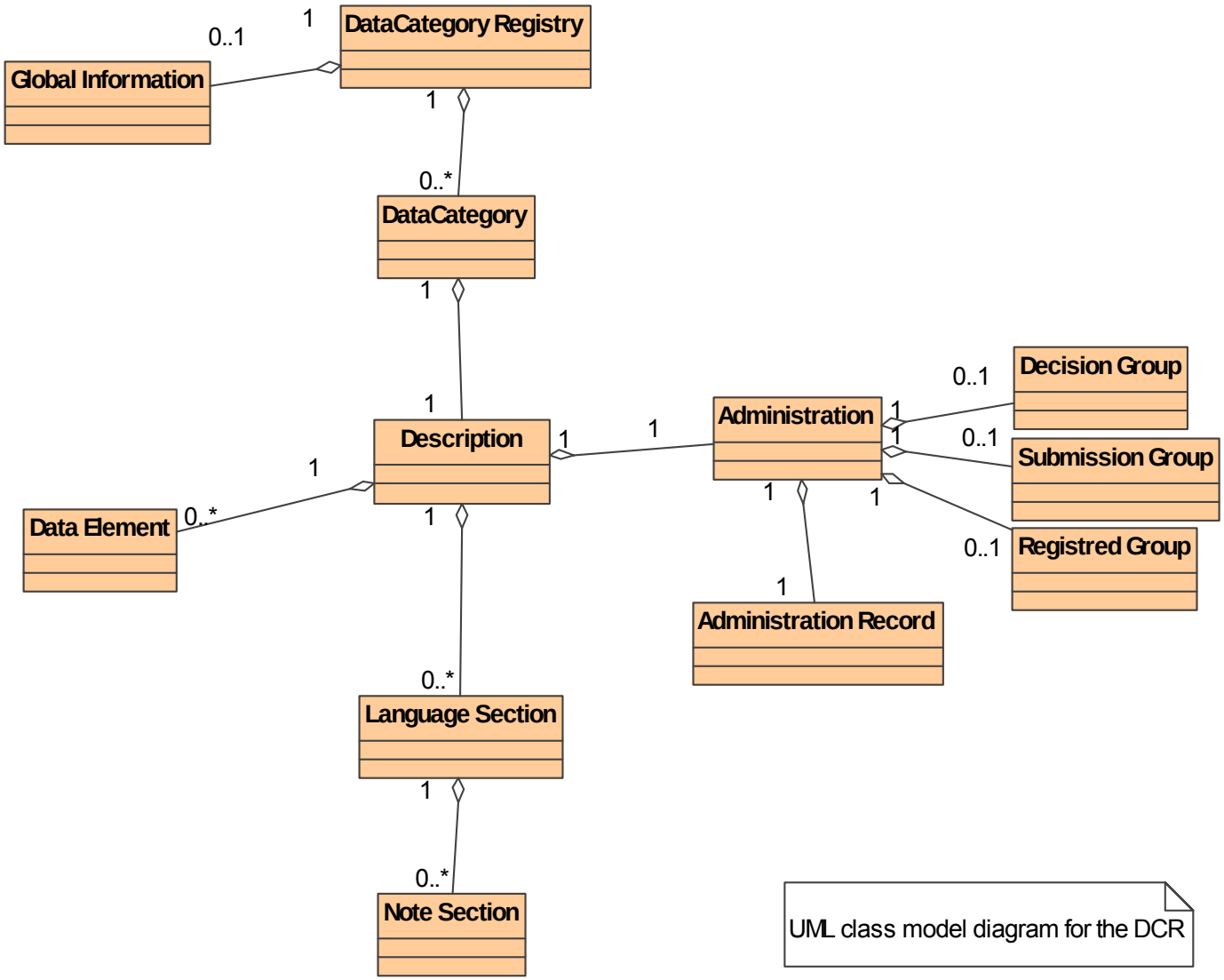
Morpho-syntax DatCats (starting)

- Currently not an ontology of DatCats but a very flat organization of constants
- Monica Monachini & Thierry Declerck gathered a great deal of values from Eagles, Multext-East & STTS
  ► we plan to study that during this summer
- I personally began to record for English & French
  - part of speech
  - morphological features (like /grammatical number/)

  constants
  attributes with language specific pick list
  ex: in English, /grammatical number/ can be valued by /singular/ or /plural/
- But, rather hesitant on this last point? Is the goal of the registry to record a beginning of organization? An other strategy could be to stick to constants?


INIST-CNRS : Syntax Suite - Wanadoo
Fichier Edition Affichage Favoris Outils ?
Précédente
Adresse http://syntax.inist.fr/index.php?page=ws&section=modify&DCid=1298
OK
Syntax
Home Main Search Compare Help Logout
List NewProp. STOP
Identification for the data category : grammaticalNumber , version : 0.0.3
Identifier *: grammaticalNumber
Version: 0.0.3
Update
Concept
Definitions (required)
en
def
source
A grammatical category for the variation in form of nouns, pronouns, and any words agreeing with them, depending on how many persons or things are
Profiles (required)
Profile Name:
MorphoSyntax
Update
Add to profile : -- Select one --
Add
Cancel
Levels
Broader Concept And Conceptual Domains
Explanations
Examples
No example yet. Click on the Create button to add one
Create
Notes
Data Elements (DE)
Language Section (LS)
english french czech slovenian
--Select a language--
Create
Definitions At LS
Conceptual Domains At LS
Conceptual domains
To add or update a conceptual domain ,freely type the whole value, or click on Enter to view in the choice list on the right the existing conceptual domains whose identifier matches. Conceptual domains at LS should have been previously declared at the Concept level. Clicking on the bin image will unset the conceptual domain.
singular
-- 5 entries --
Update
plural
-- 5 entries --
Update
-- 5 entries --
Add
Cancel
Examples At LS
Notes At LS
Name Section (NS)

DCR structure



UML class model diagram for the DCR

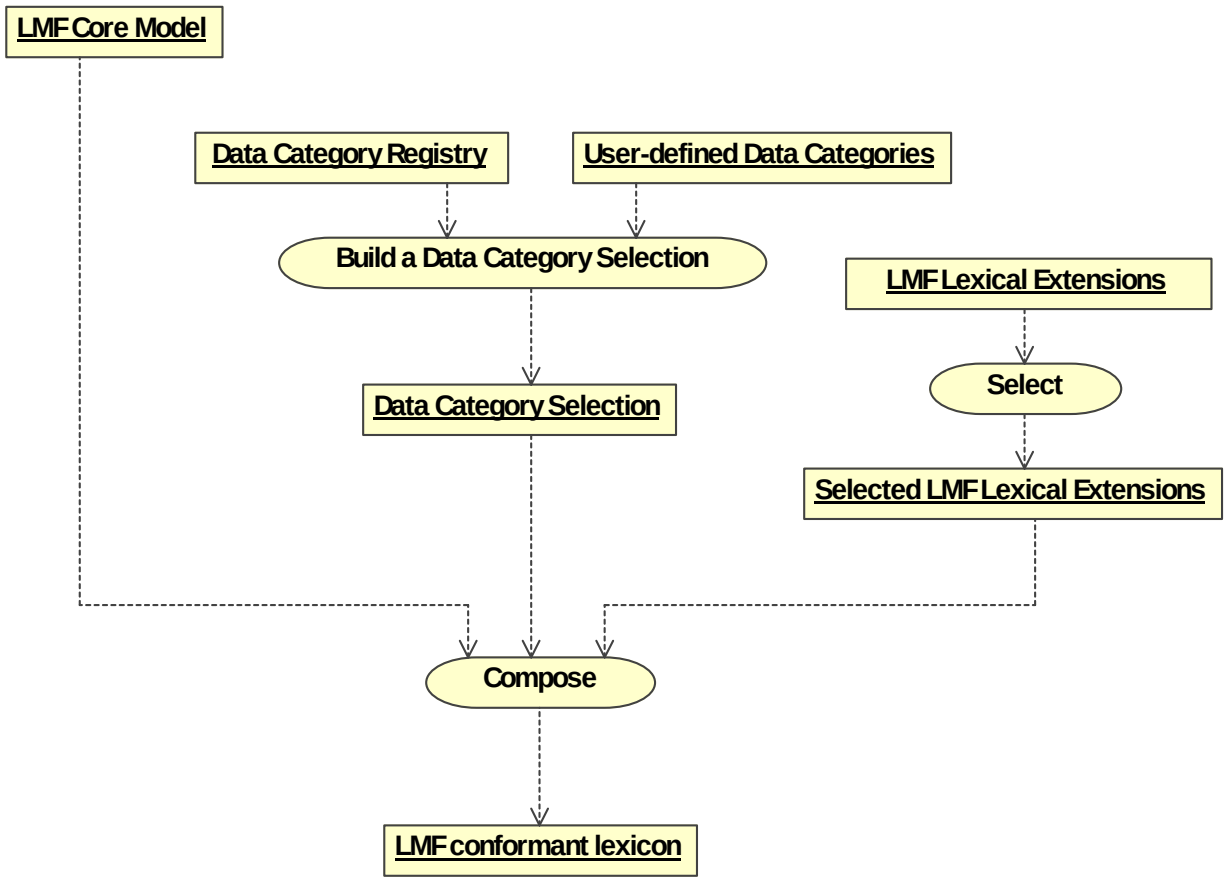
Lexical Markup Framework (LMF) ISO 24613

- Scope: a standard for MRD & NLP lexicons
- Target: ease management and interchange of lexicons
- Work started in Autumn 2003
- Convenior : Nicoletta Calzolari (Italy)
- 2 co-editors : Gil Francopoulo (France) & Monte George (USA)
- LMF Rev-6 document produced on 15th June 2005

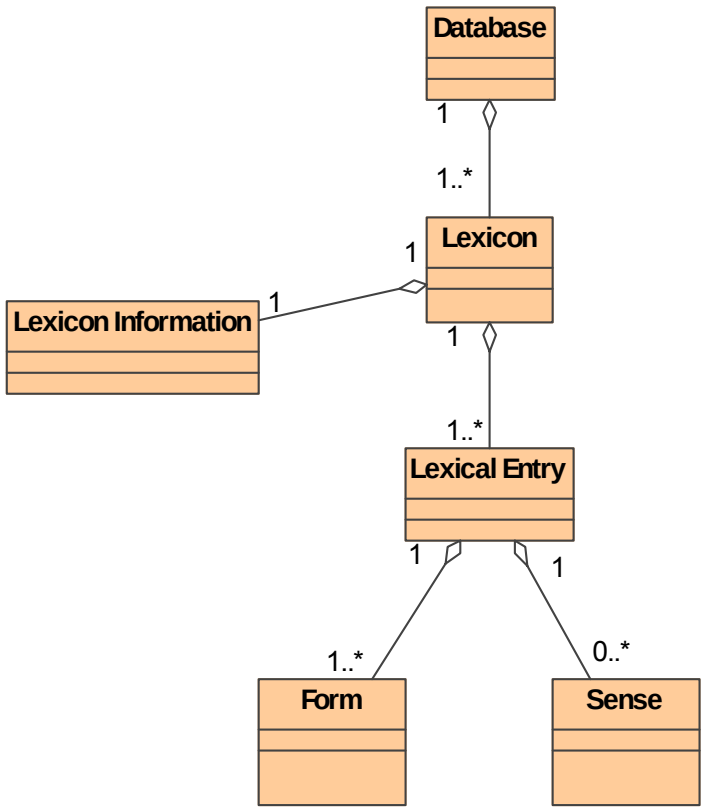
Quite a lot of work (and meetings) has been devoted to this task
Collaboration of experts from Europe, America and Asia
(intellectual) testing against most famous NLP lexicons

Quite a complex task: one core model and extensions (currently 5 extensions: MRD, NLP morpho, NLP syntax, NLP semantics, NLP multilingual notations)

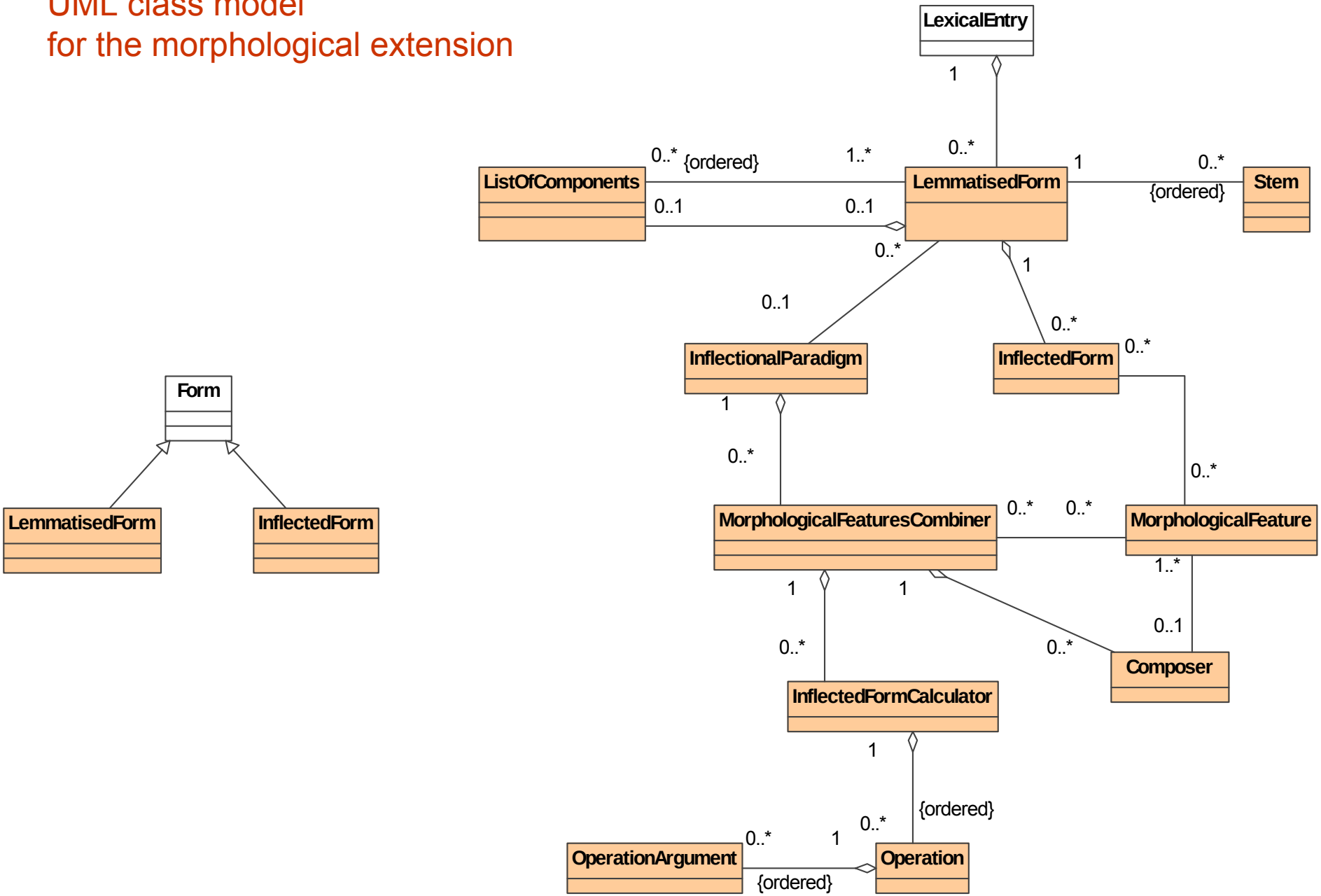
Process


LMF Core Model
Data Category Registry
User-defined Data Categories
Build a Data Category Selection
LMF Lexical Extensions
Select
Data Category Selection
Selected LMF Lexical Extensions
Compose
LMF conformant lexicon

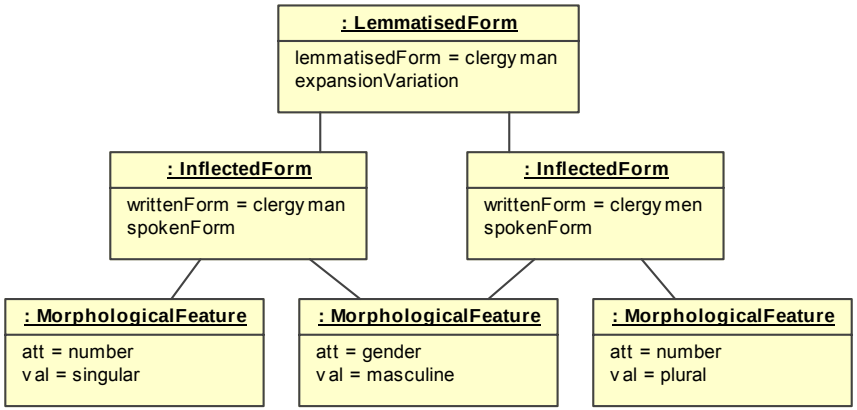
UML class model for core LMF (rev-6)

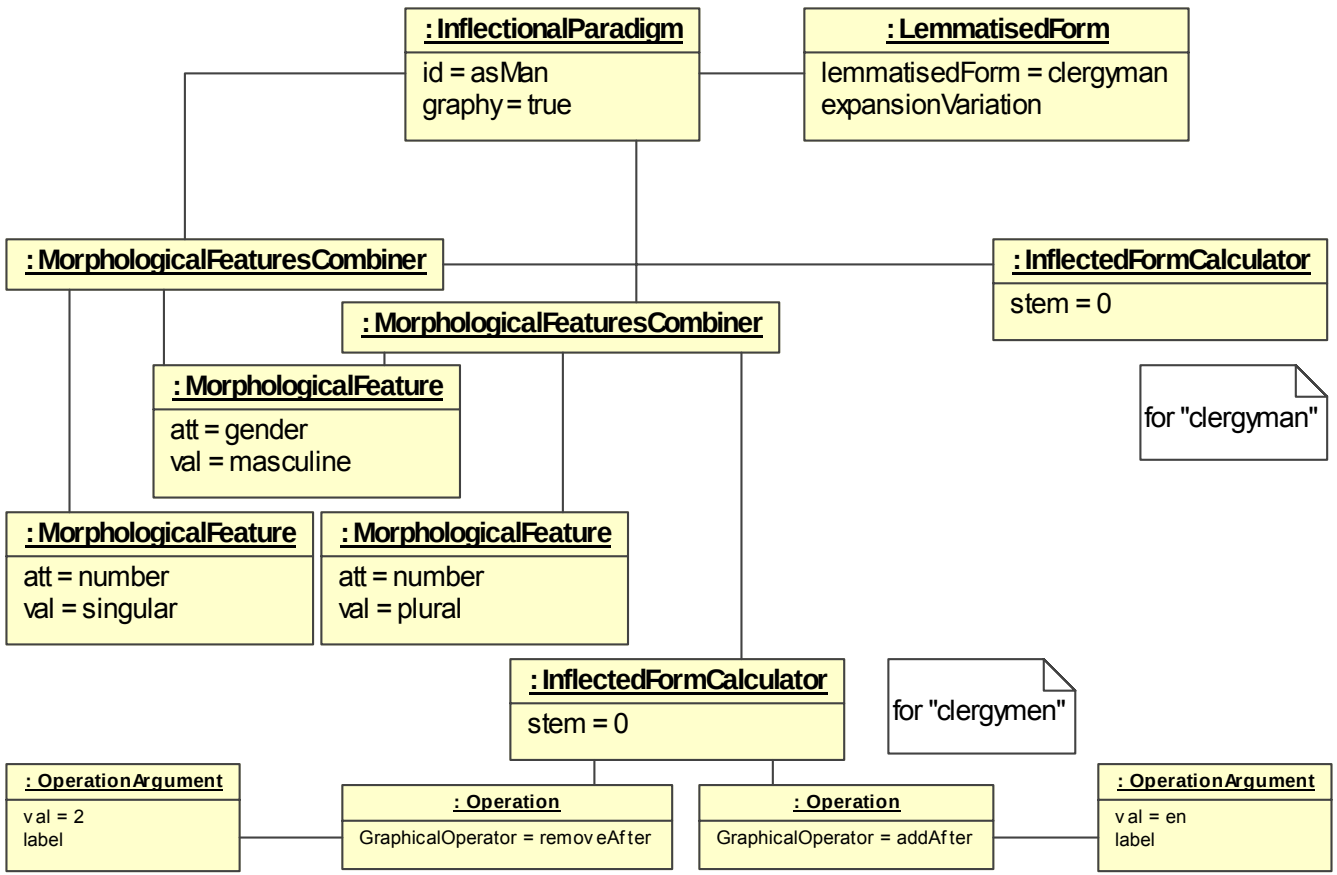


UML class model
for the morphological extension



UML object model example:
Two ways to describe « clergyman »

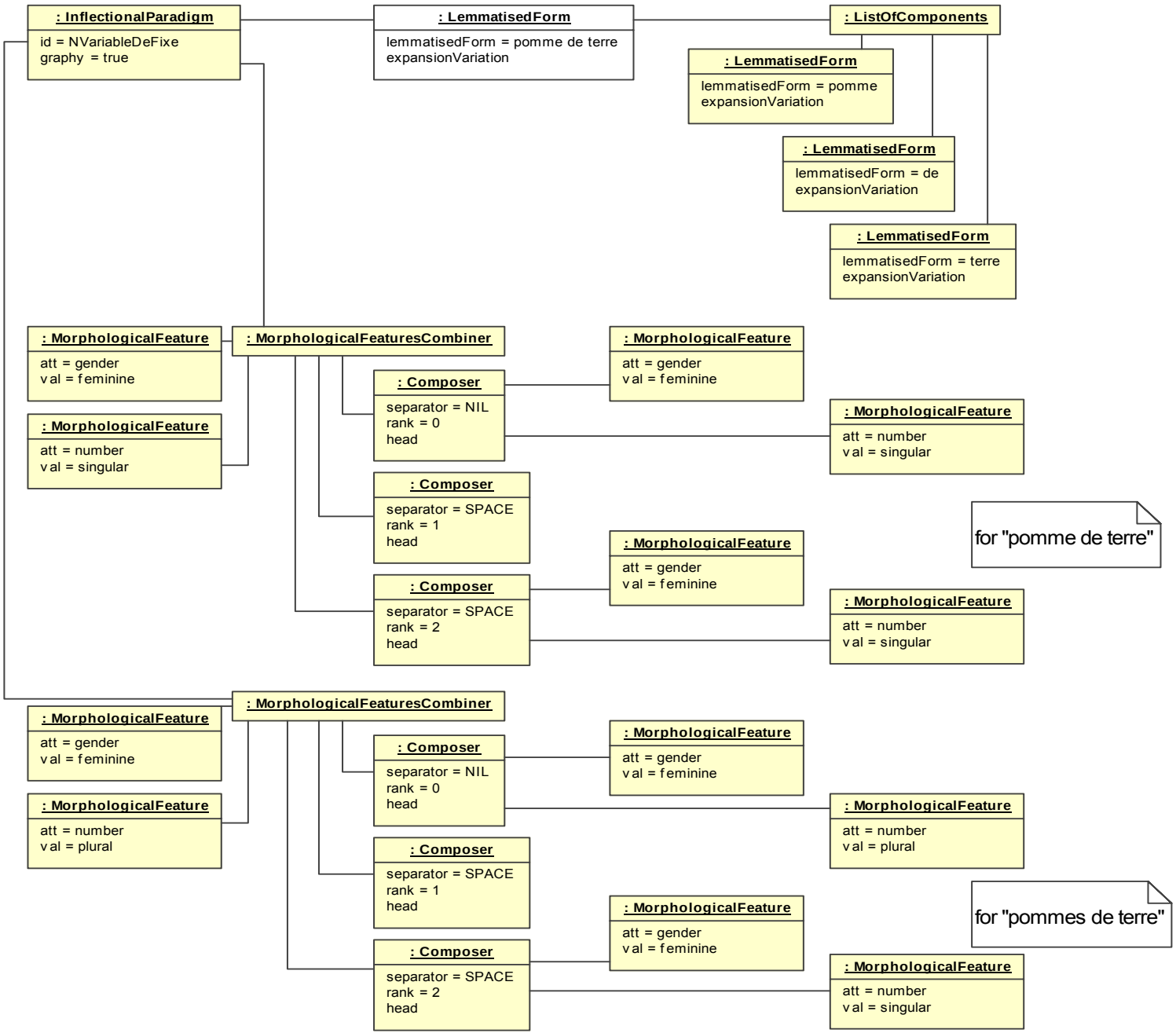




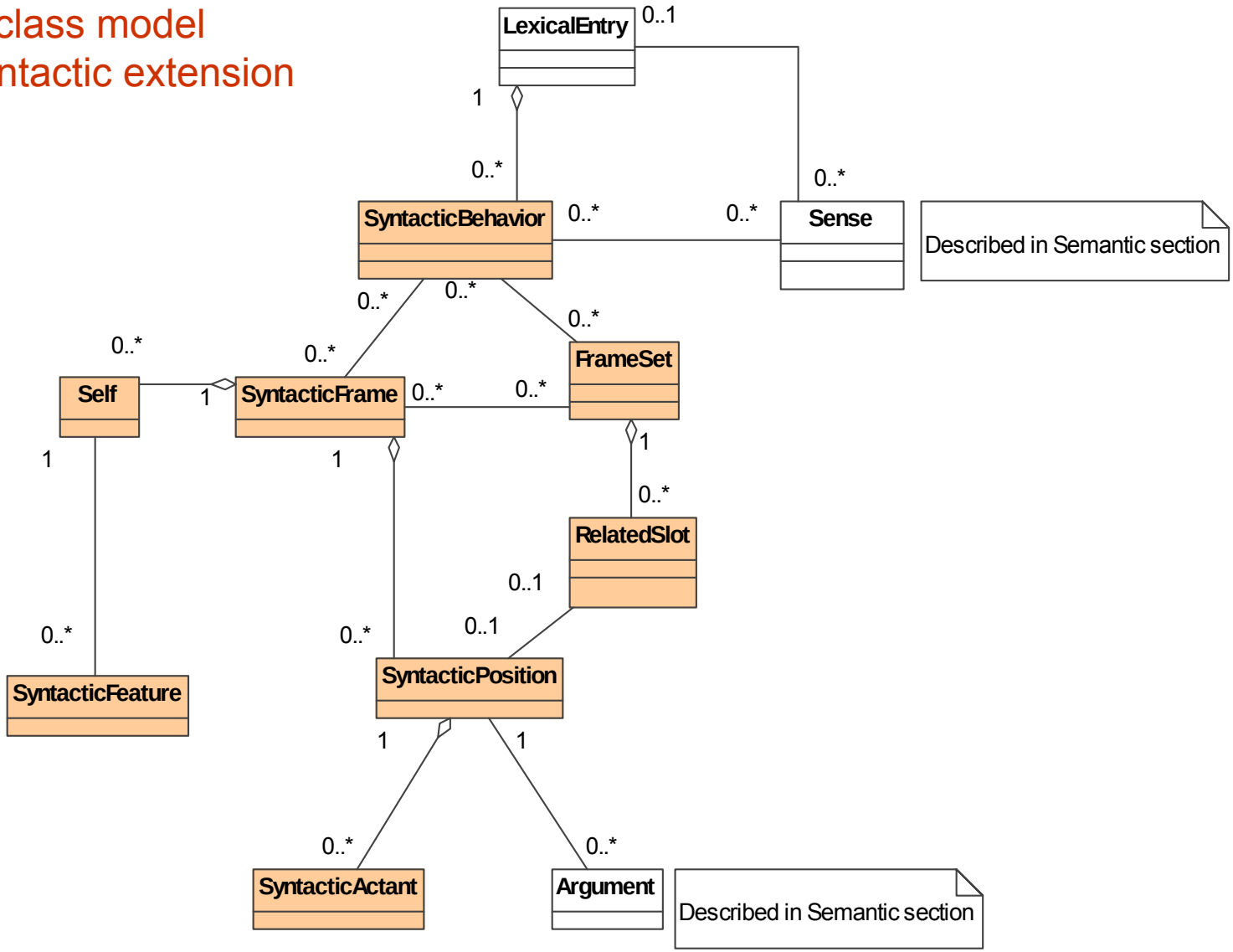
This could give the following external XML format

- <LemmatisedForm lemmatisedForm=« clergyman » ip=« asMan »/>
<InflectionalParadigm id=« asMan » graphy=« true »>
<MorphologicalFeaturesCombiner mfs=« mf1 mf2 »>
<InflectedFormCalculator stem=« 0 »/>
</MorphologicalFeaturesCombiner>
<MorphologicalFeaturesCombiner mfs=« mf2 mf3 »>
<InflectedFormCalculator stem=« 0 »/>
</MorphologicalFeaturesCombiner>
</InflectionalParadigm>

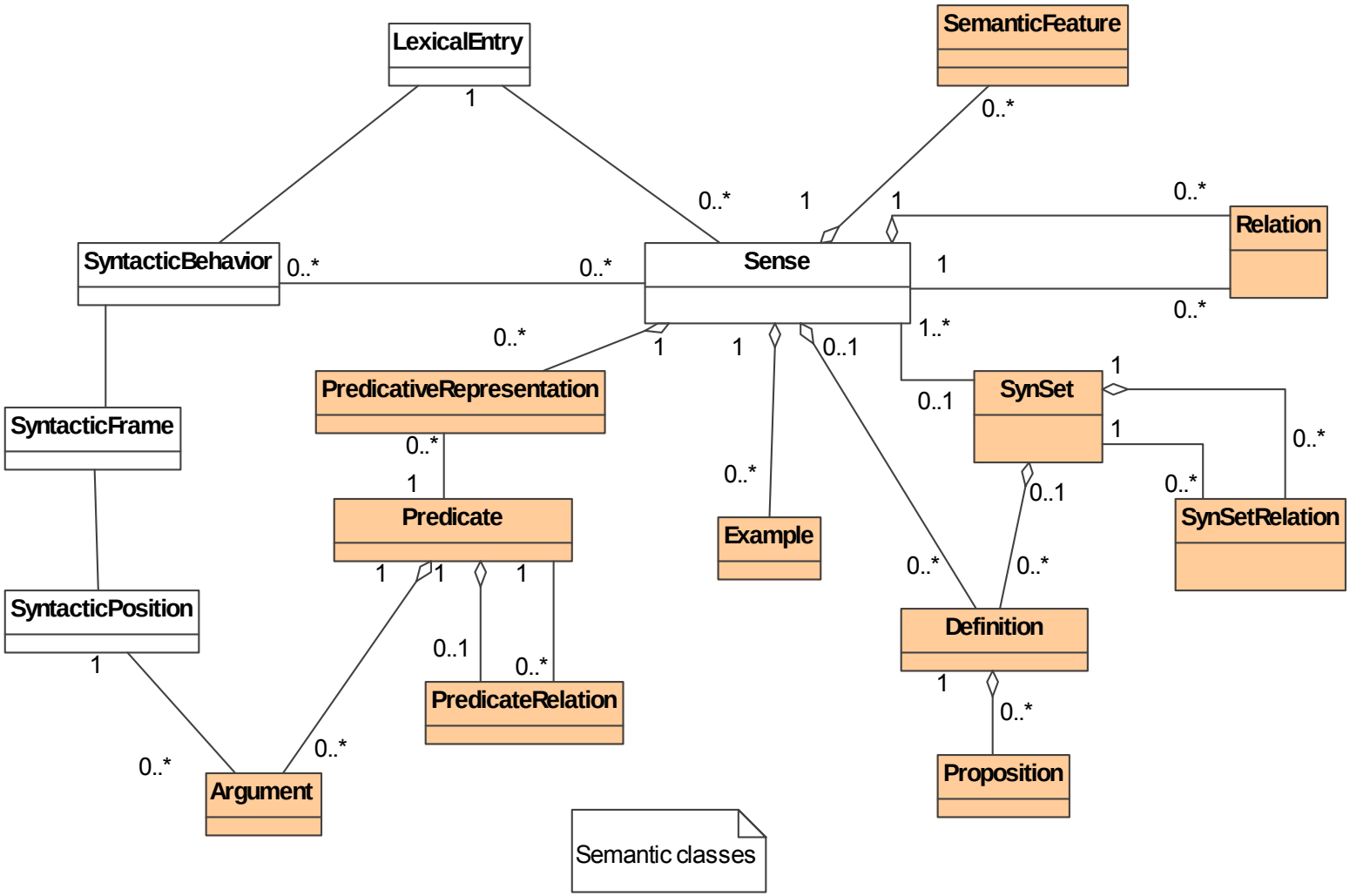
UML object model example: MWE



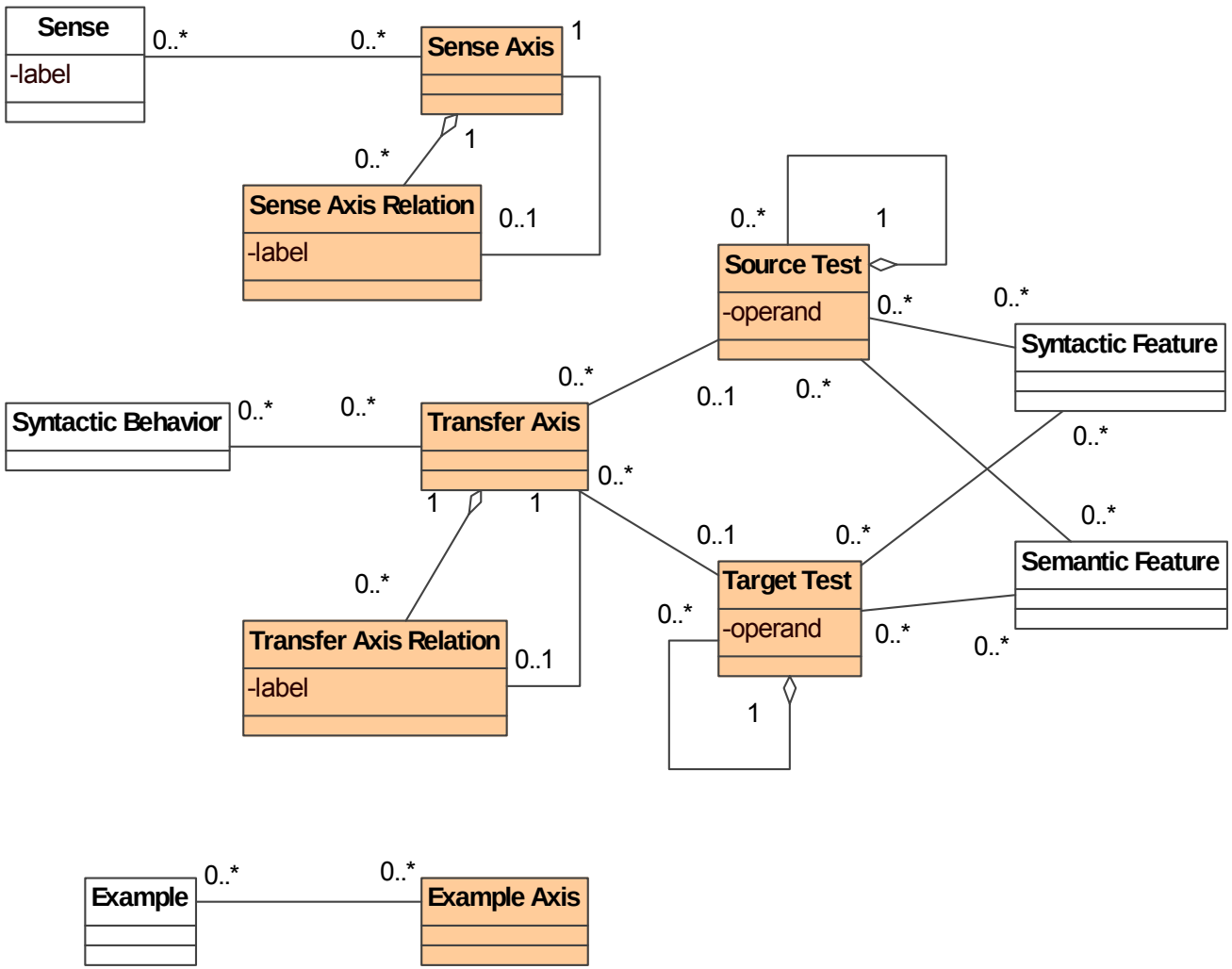
UML class model for syntactic extension



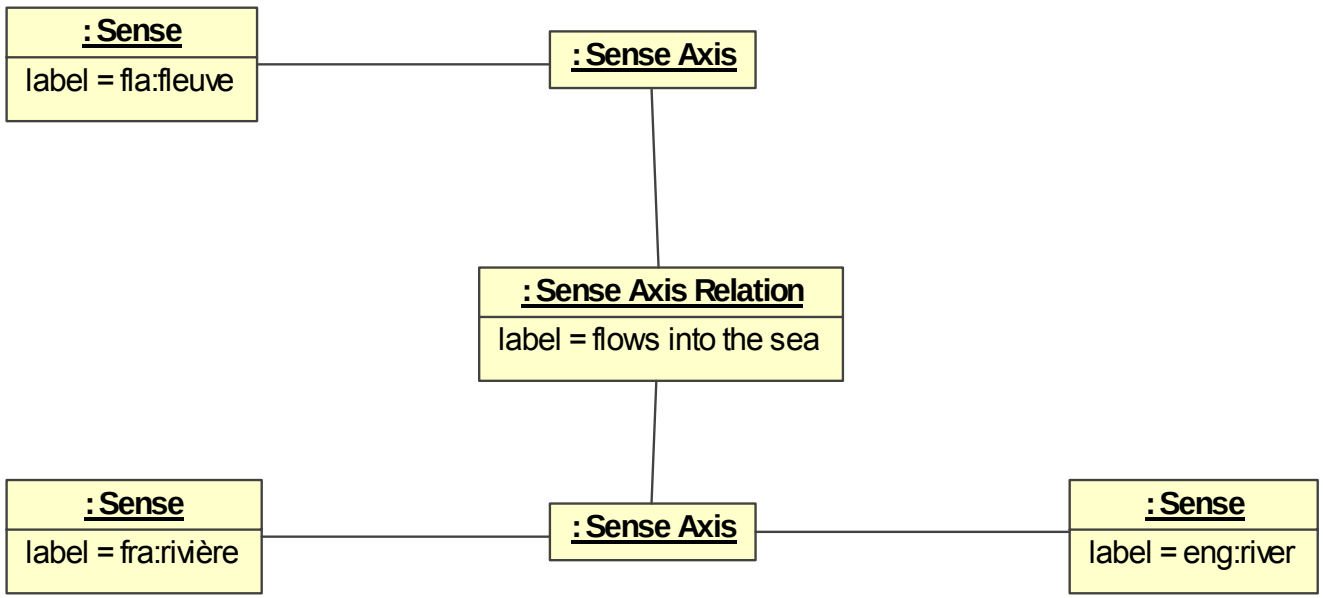
UML class model for semantic extension



UML class model for multilingual notations



UML object model example for multilingual notations



Not a tool for an ontology: but 2 possibilities to hook nodes belonging to ontologies

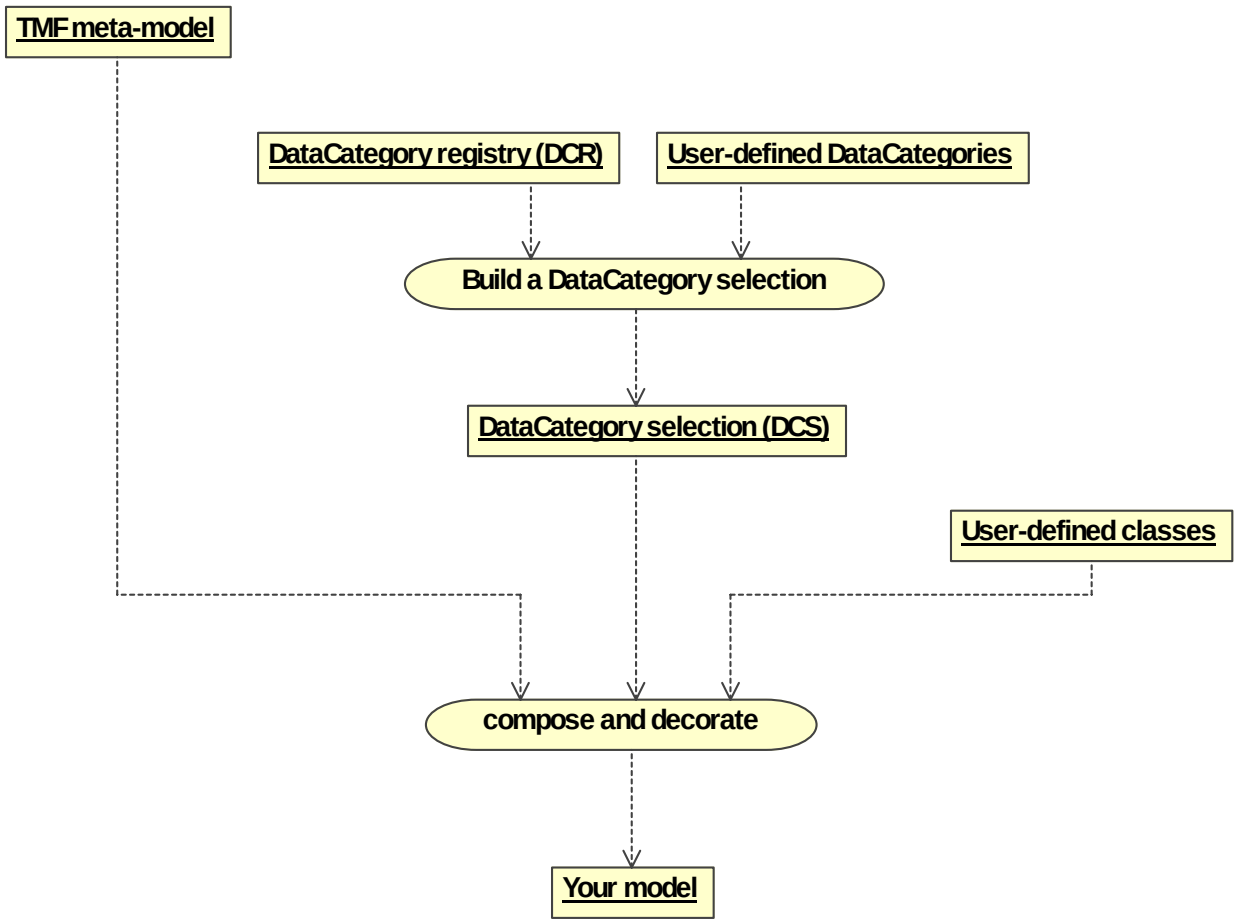
Conceptual vs external physical model

- Until now, we worked only on the conceptual model with UML
- No real decision concerning the external physical model: just some tests.
- If we decide something, of course the specifications will be based on XML
- We have 4 different options:
  - traditional XML validation: DTD
  - modern XML validation: RelaxNG schema
  - no XML validation: GMT
  - RDF-OWL description for Semantic Web guys
- Open issue section (at the end)

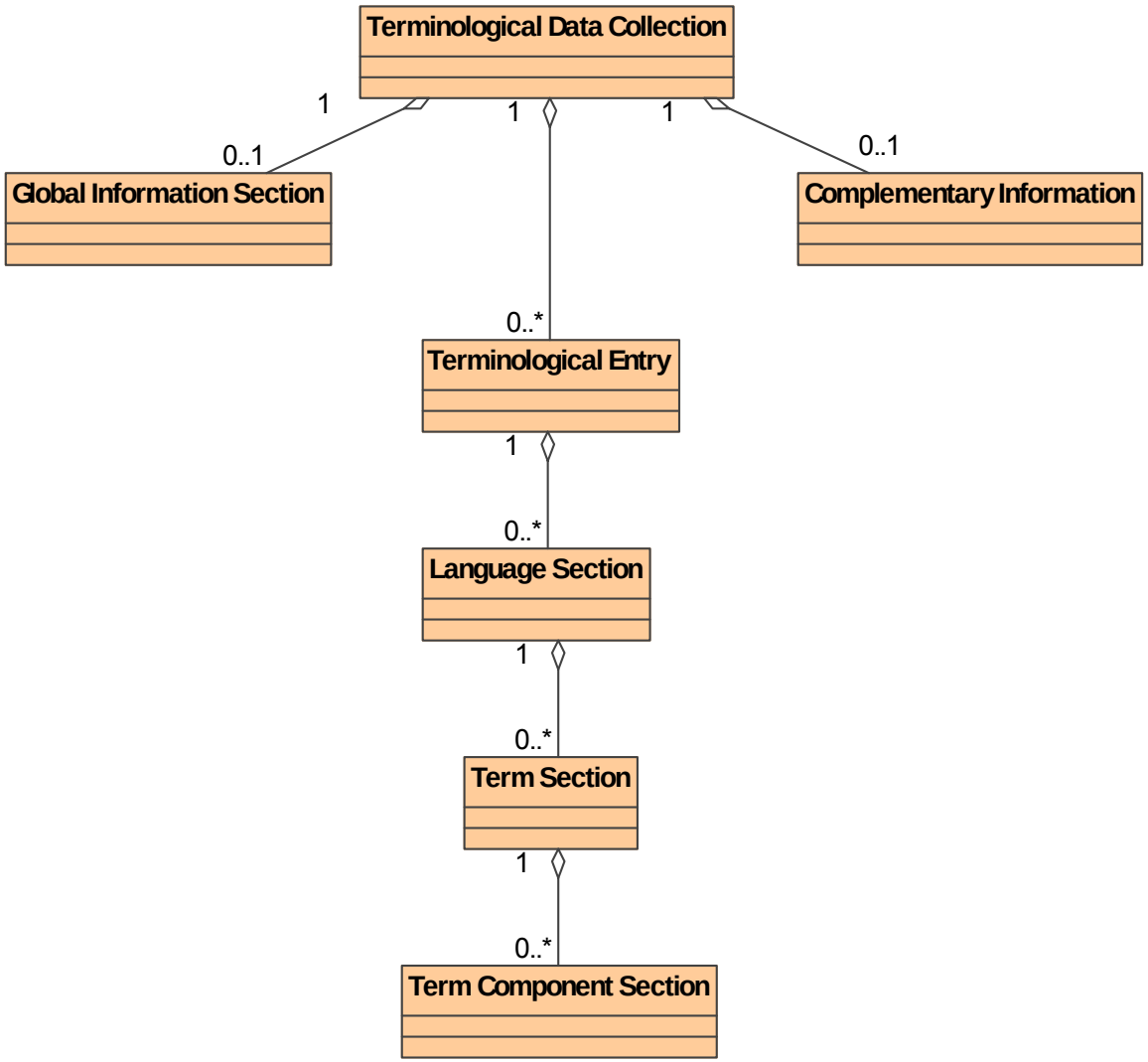
Terminological Markup Framework (TMF) ISO 16642 published in 2003

- standard dedicated to terminology management
- for one or more specialized domains
- mono or multilingual
- no linguistics
- onomasiologic: starts from the sense to reach the word form
- a means to cover 2 concurrent standards: MARTIF & GENETER
- in multilingual context. hypothesis: there exists an interlingual concept that links the terms from different languages

Process


TMF meta-model
DataCategory registry (DCR)
User-defined DataCategories
Build a DataCategory selection
DataCategory selection (DCS)
User-defined classes
compose and decorate
Your model

TMF meta-model



Conclusion

- As you can see, these standards do not progress at the same speed.
- Specifications do not come from the « sky »
- Specifications must be explicited thru consensus. We must listen from foreseen users, from people coming from different cultures and languages. And sometimes, a good consensus is not very easy to reach.
- For sure, the process takes time.
- The overall objective is to provide good and suited documents in a reasonable time.