Data Category Registry: Morpho-syntactic profile

Gil FRANCOPOULO

INRIA-Loria (LIRICS project)

Summary

- **1 What is a datcat?**
- **2 What is a profile?**
- **3 What is the situation in the morpho-syntactic thematic domain group?**

  - what has been done?
  - what is left to do?

1 What is a datcat?

A datcat is basically a symbol

- Context:
- In TC37/SC3+SC4, we have and we are on the way to specify two sorts of standards

- Low level standards
- The set of data categories: this is the pair:
  - Revision of ISO12620 that specifies how the datcats are described and maintained
  - Registry of datcats (DCR) endorsed by ISO-12620
- This registry provides all the linguistic symbols that we need
- Of course, there are also some other important low level standards that we need, but we are not going to define them because they already exist. So we will use them. These are for character codes, language codes, script codes, country codes.

High level standards

- These are structural models that specify how to represent linguistic resources.
- The structural model provides the classes (in UML terminology) and links between classes.
- And the registry provides the symbols for attribute names and constant values. The attribute/value pairs decorate the classes.
- These structural specifications deal mainly with: word-segmentation, morpho-syntactic annotation (MAF), syntactic annotation (SynAF) and lexicon (LMF).

Objectives: the goal is to propose to the user a coherent family of standards

- All these standards share this property: the user defines a model of linguistic resource by combining structural elements with symbols taken from the datcat registry (DCR).
- So all these resources share the same set of symbols. The goal is to provide a good interoperability between segmentation, annotation and lexicon.

2 What is a profile?

- A profile is a set of datcats related to a specific type of information
- The current profiles are:
  - for TC37/SC3: Terminology (for TMF)
  - for TC37/SC4: NLP
    TDG1 meta-data
    TDG2 morpho-syntax
    TDG3 semantics
    TDG4 syntax
- Each profile is managed by a committee and a chairman

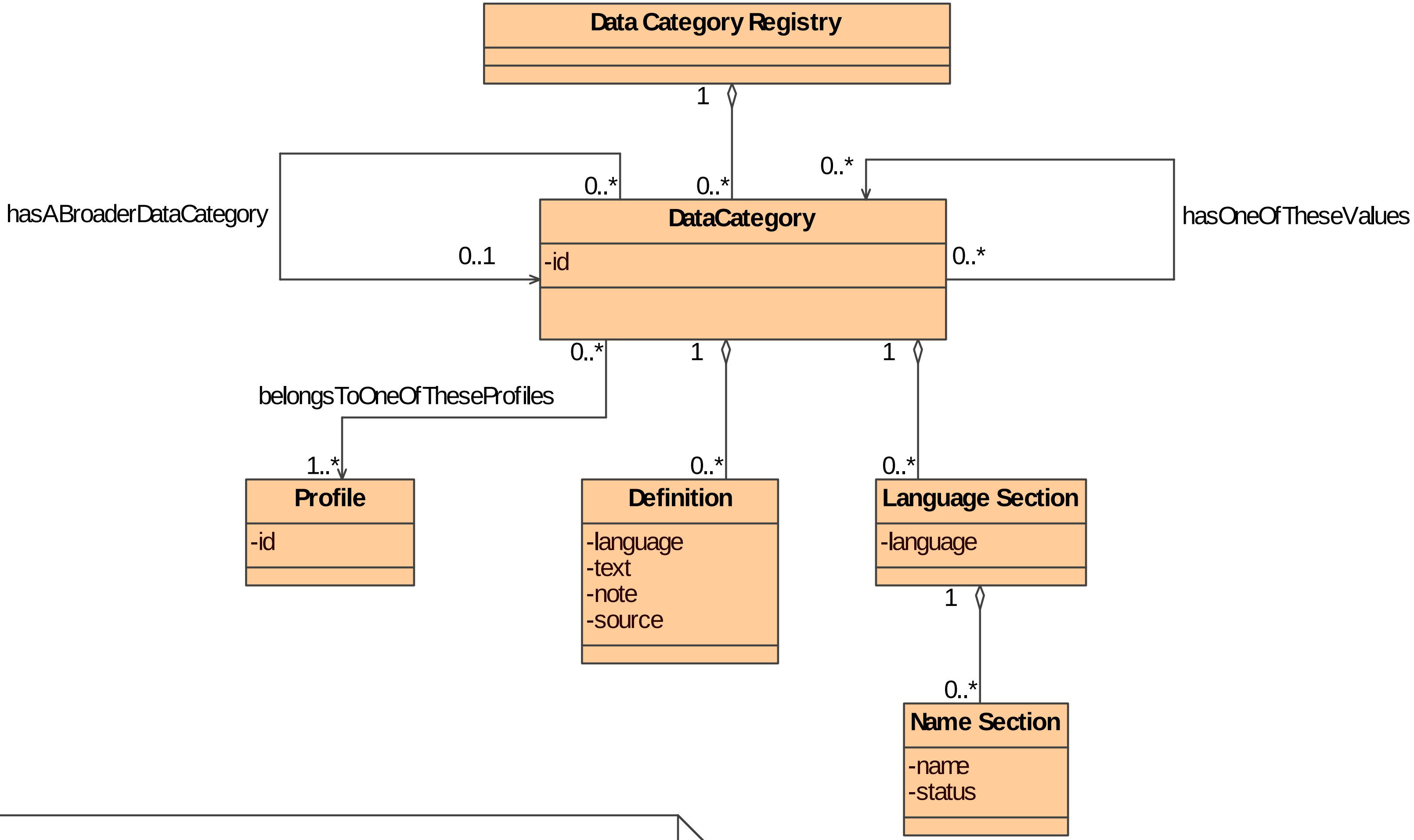
- Note-1: in order to ensure a good interoperability between WS, Annot & Lexicon, the profiles are the same. The split is made on the linguistic criteria, not the resources
- Note-2: a datcat may belong to several profiles but in fact we try to avoid this (in order to avoid conflicts).

3.1) What has been done in the morpho-syntactic thematic domain group ?

- The goal is to extablish a first set of data categories
- Progress has been rather slow: 4 phases

- PHASE-1: to collect (done)
- PHASE-2: to revise = to group, to structure and complete the definitions (done)
- PHASE-3: to extend to Semitic languages (done)
- PHASE-4: to extend to Asian languages (to be done)

- PHASE-1: an initiale flat list of 281 datcats has been collected from:
  - current ISO-12620
  - Eagles
  - Multext-East
  - a couple of values for LMF
- The symbols coming from ISO-12620 were general purpose values like « language » or « derivation ». But it's not enough for NLP resources because they cover just terminological resources. For instance, for /part of speech/ the only values are /noun/, /adjective/ and /verb/. By comparison, in NLP we need all linguistic values like /preposition/ and /pronoun/.
- In fact, most linguistic values come from Eagles. And extension for slavic languages comes from Multext-East.



Data Category Registry

1

0..*

DataCategory

-id

hasABroaderDataCategory

0..*

0..1

0..*

hasOneOfTheseValues

0..*

0..*

1

1

belongsToOneOfTheseProfiles

1..*

Profile

-id

0..*

Definition

-language
-text
-note
-source

0..*

Language Section

-language

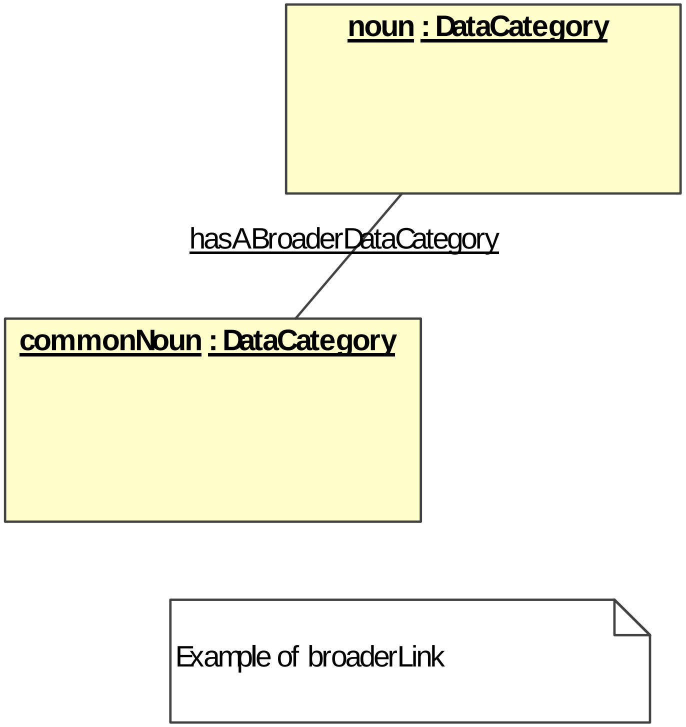
1

0..*

Name Section

-name
-status

classes and attributes that are used in the morpho-syntactic profile

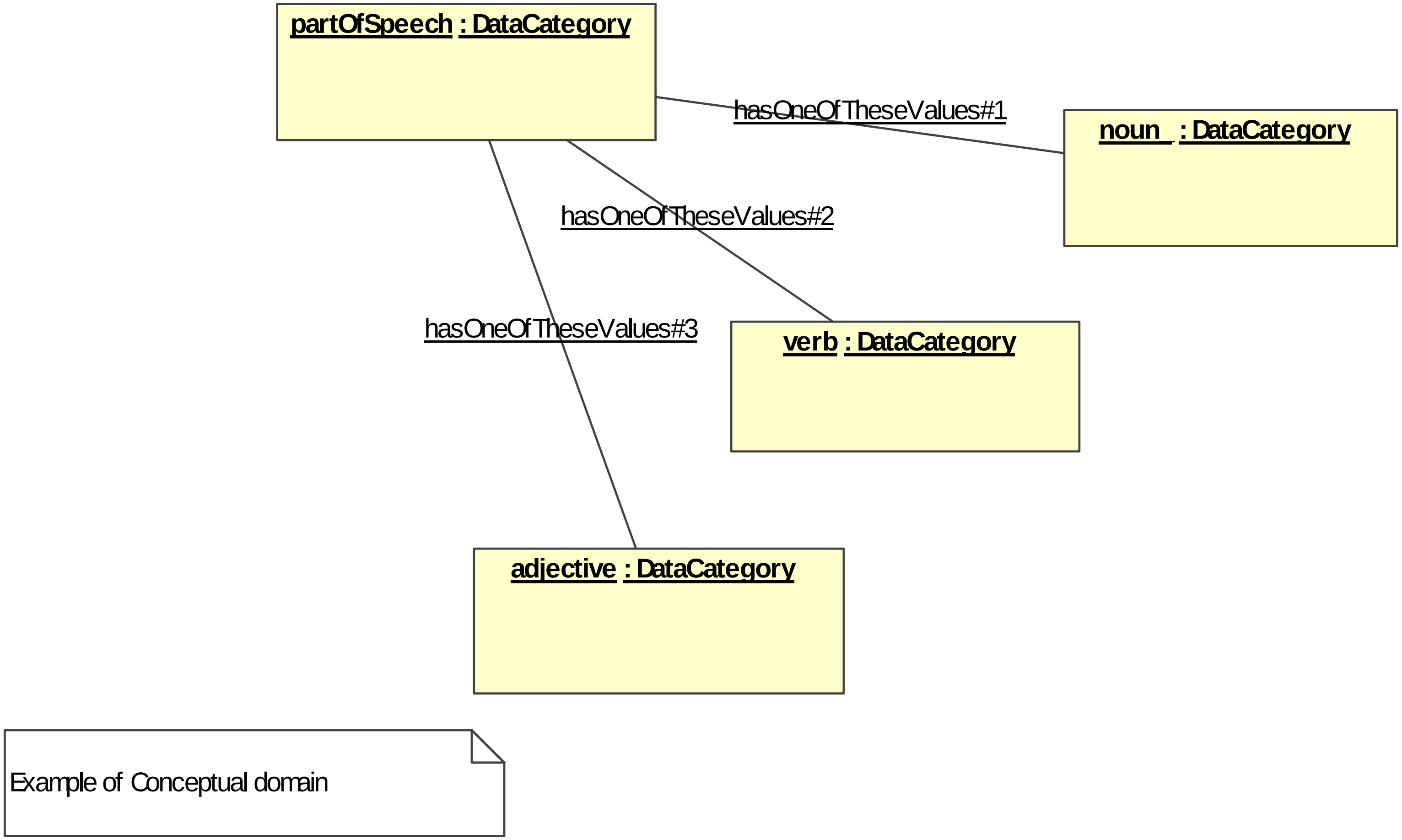


noun : DataCategory

hasABroaderDataCategory

commonNoun : DataCategory

Example of broaderLink



partOfSpeech : DataCategory

hasOneOfTheseValues#1

noun_ : DataCategory

hasOneOfTheseValues#2

verb : DataCategory

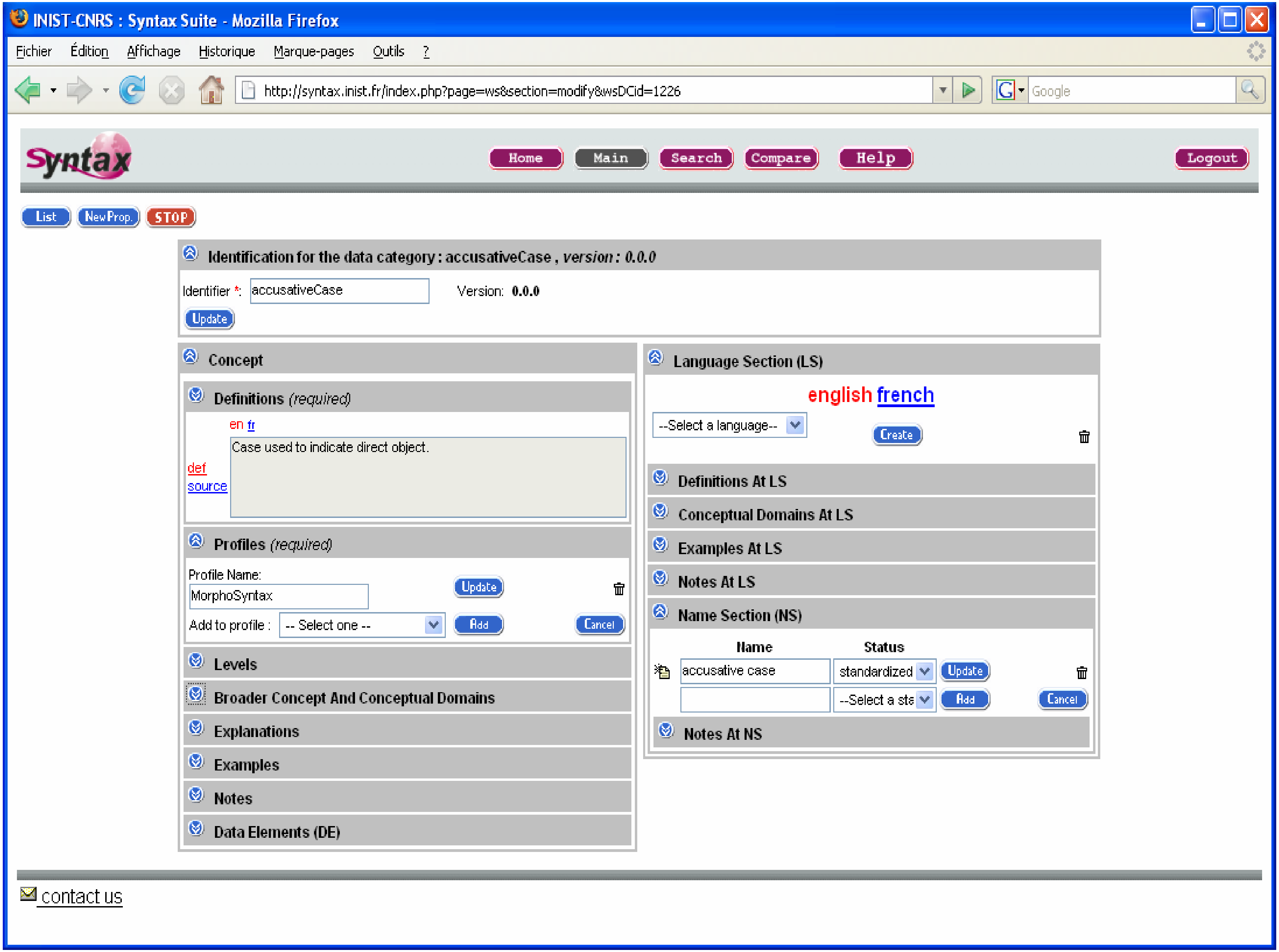
hasOneOfTheseValues#3

adjective : DataCategory

Example of Conceptual domain

What has been recorded so far in the DCR?

- we use the Syntax software hosted by INIST in Nancy: http://syntax.inist.fr
- the software is not very « fancy » but it works
- the list being rather huge (341 items now), the software directories are used in order to help datcat categorization



http://syntax.inist.fr/index.php?page=ws&section=modify&wsDCid=1226

Identification for the data category : accusativeCase , *version : 0.0.0*

Identifier *: accusativeCase

Version: 0.0.0

Concept

Definitions *(required)*

en fr

def

source

Case used to indicate direct object.

Profiles *(required)*

Profile Name: MorphoSyntax

Add to profile : -- Select one --

Levels

Broader Concept And Conceptual Domains

Explanations

Examples

Notes

Data Elements (DE)

Language Section (LS)

english french

--Select a language--

Definitions At LS

Conceptual Domains At LS

Examples At LS

Notes At LS

Name Section (NS)

Name	Status
accusative case	standardized
	--Select a sta

Notes At NS

•PHASE-2: revision => 7 Directories

	#
Basics	49
Cases	33
Form related symbols	33
Morphological Features excluding Cases	73
Operations	27
Part Of Speech	107
Register Dating Frequency	19
Total=	341

Basics (extract)

- abbreviation
  comment
  derivation
  elision
  expansionVariation
  foreignText
  label
  native

- **Cases**
- attribute = case
- values like accusativeCase, dativeCase, genitiveCase etc.

- **Morphological features (excluding cases)**
- attributes like grammaticalGender, verbFormMood, grammaticalTense etc.
- values like feminine, indicative, present etc.

Part Of Speech

- Attribute = partOfSpeech
- A small hierarchy of values in order to provide two levels of detail:
    - commonNoun vs noun
    - preposition vs adposition

Recently we studied what was missing with regards to semitic languages and we arrived at an amount of 7 data categories to be added

What is left to be done?

- The two Asian experts in our group must provide the missing data categories for Asian languages. They are working on it, but did not deliver yet
- I asked also two experts in African languages, but they didn't reply

Thank you ...