

**SILT**

Sustainable Interoperability for Language Technology



**Fostering Language Resources Network**  
Grant Agreement No. ECP-2007-LANG-617001

# Data Categories and their Semantics

*Jan Odijk, FLareNet/Utrecht/CLARIN-NL*

November 1-2, 2009

SILT-FLaReNet, Brandeis



# Data Categories and their Semantics

1. Data Categories & ISOCAT
2. Specific Problems (some examples)
  - Semantics of Data Categories
  - Structured Data Categories?
  - Mapping Data Categories
  - Collections of Data Categories (Profiles?)
3. Challenges
4. How to Proceed: Action Items



# Defining the Topic: Data Categories

1. Data Categories (DC) are defined in a Registry
2. (Some) Data Categories are made part of a standard
3. (Contribution to) Semantic Interoperability by
  1. Using the standardized DCs, or
  2. Mapping one's own DC to a standardized DC

# Interoperability by Mapping DCs

- Example (simple, naïve):
  1. LR1: uses DC *myDC*
  2. LR2: uses DC *yourDC*
  3. Standard has *ISODC*
  4. *myDC*  $\Leftrightarrow$  *ISODC*
  5. *yourDC*  $\Leftrightarrow$  *ISODC*
  6. *Therefore: myDC*  $\Leftrightarrow$  *yourDC*
- *ISOCAT standardized DCs serve as a pivot (cf. interlingua)*
- *Gain when >2 resources/tools*



# Interoperability

- Tools must specify
  - Form of input it can apply to
  - How it interprets elements in the input
  - Format and interpretation of output
- Data must specify
  - Form of the data
  - Interpretation of elements in the data
- Interoperability only if there is a match,



# Interoperability

- Tools & Data
  - Applicability of tool onto data
  - Invocation of converter if mismatch
- Data1 & Data2
  - E.g. merging lexicons
  - Merging/combining different annotations of the same source text
- Tool1 & Tool2
  - E.g. in a pipeline
  - More generally: In work flow systems

# ISOCAT

- ISOCAT ([www.isocat.org](http://www.isocat.org)) provides framework for
  - Data Categories (DC) and their Registry
  - Persistent identifier (PID) for each DC (in the form of a ‘cool’ URI)
  - Definitions, examples, ... (in multiple languages)
  - Associated value domains
  - (Administrative) Status (candidate, private/public)
  - With API, input & output formats, Web interface, ...



# ISOCAT

- Goal of ISOcat/DCR:
  - Defining widely accepted linguistic concepts
- Open:
  - User community involvement is encouraged.
- Official standard:
  - Decision making conforms to ISO process.



# ISOCAT

Anyone

1. can register with ISOcat
2. can create data categories
3. can create data category selections (DCSs)
4. can create groups
5. can share DCSs with groups
6. can make DCSs public
7. may submit DCSs for standardization

# ISOCAT

## Thematic Domain Group

- group of [experts](#)
- select and maintain the [data categories](#)
- that are relevant for a [thematic domain](#)
- as part of the [ISO 12620:2009](#) standardization process

# Specific Problems

- Interpretation
  - What is the interpretation of a specific ISOCAT DC?
- Mapping
  - How do I map *myDC* to an ISOCAT DC?
  - How do I map the ISOCAT DC to *yourDC*?



# General Recommendation

- Test ISOCAT
  - against your resource / tool / project and
  - report on virtues / problems / omissions / questions!!
- In NL appr. 10 projects in CLARIN-NL in 2010

# Interpretation

- My (Dutch) resource contains DC
  - *Overgankelijk*
- What is the corresponding DC in ISOCAT?
  - English translation is: *transitive*
- “transitive” occurs in ISOCAT
  - <http://www.isocat.org/datcat/DC-1405>
  - In profile *morphosyntax*
  - *Status: reg.:candidate, admin:private, scope:public*
- Does it mean the same as *overgankelijk*?

# DC-1405 Definition

- English:
  - *A verb which takes a direct object; that is, a verb that expresses an action which directly affects another person or thing.*
  - *Source:*  
[www.southwestern.edu/~carlg/Latin\\_Web/glossary.html](http://www.southwestern.edu/~carlg/Latin_Web/glossary.html)
  - *Example:*
    - *He **has** a horse*

# DC-1405 Definition

- This definition raised 15 (!) problems for me
- Discussed at NEERI Standards Workshop
- <http://www.csc.fi/english/pages/neeri09/workshop/materials/odijk.pdf>
- Here:
  - some problems that might be prevented by an improved ISOCAT system
  - To derive recommendations

# DC-1405 Problems

- *Problem 1: there are two subdefinitions*
  - *A verb which **takes** a direct object*
  - *That is, a verb that **expresses an action** which **directly affects another person or thing**.*
- It violates the ISOCAT DCR guidelines
- (<http://www.isocat.org/manual/DCRGuidelines.pdf>)
- → add provisions to avoid such duplicate definitions

# DC-1405 Problems

- *Problem 2:*
  - *take* in 1<sup>st</sup> subdefinition is not defined
    - Type reading
      - Obligatory reading: “must be combined with” ([Lexicon of Linguistics](#))
      - Potentiality reading: “can be combined with”
    - Token reading: “is combined with”
- → require link of each term used to a DC

# DC-1405 Problems

- *Problem 3: 2<sup>nd</sup> subdefinition*
  - *defines morphosyntactic DC in (apparently) purely semantic terms*
  - *This is suspect (why then is it a morphosyntactic DC?)*
- → warn for DCs of level A defined exclusively by terms for DCs from level B

# DC-1405 Problems

- Problem 4:
  - *Most of the terms used in the second definition do not correspond to a DC (hence have no definition) in ISOCAT*
  - *Person does occur in ISOCAT (in the sense of grammatical person, [DC-399](#)) but obviously not in the sense intended here*
- → require link of each term used to a DC

# DC-1405 Problems

- Problem 5:
  - *A plausible interpretation of 2<sup>nd</sup> definition:*
  - *→ 1<sup>st</sup> and 2<sup>nd</sup> subdefinition are different (→ internal contradiction)*
  - *→ example given is incorrect*
    - *He **has** a horse* (state v. action)
    - *He **baked** a cake* (effected .v affected thing)
    - *He **bought** a car* (nonaffected v. affected thing)
    - *He **considers** her smart* (no argument of *consider*)
    - *He **believed** her to be happy* (no argument of *believe*)
  - *(the definition source distinguishes “action” from “state” in verbs)*

# DC-1274

- Definition of “direct object” ([DC-1274](#))
  - A direct object is a **grammatical relation** that exhibits a combination of certain independent syntactic properties, such as the following:
    - The usual grammatical characteristics of the patient of typically transitive verbs
    - A particular case marking
    - A particular clause position
    - The conditioning of an agreement affix on the verb
    - The capability of becoming the clause subject in passivization
    - The capability of reflexivization
  - Source:  
[www.sil.org/linguistics/GlossaryOfLinguisticTerms/WhatIsADirectObject.htm](http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/WhatIsADirectObject.htm)
- Example: A *book* is the direct object in the sentence *They bought Anne a book*
  - Source (incorrect source given in ISOCAT)

# DC-1274

- Observation
  - This is a pretty reasonable definition (though not without problems)
  - It illustrates that pure definitions are often not possible for linguistic concepts
    - One can only place the DC relative to a superordinate category (*grammatical relation*), and
    - List some properties/criteria that can be used to argue for or against assigning this DC to a particular phenomenon
    - But the actual assignment is a hypothesis/theory, that usually can only be evaluated by taking into account the grammatical description of the whole range of relevant phenomena (i.e. in conjunction with other grammatical relations such as *subject*, *indirect object*, *head*, etc.)
    - **Recommendation** Therefore definitions of DCs from the same “conceptual Domain” should preferably originate from one and the same common source

# DC-1274 Problems

- Problem 6:
  - *Grammatical Relation* is **not** a DC in ISOCAT
  - *Grammatical Function* **is** a DC in ISOCAT ([DC-1296](#))
    - With only a French definition (name: *fonction grammaticale*)
    - In profile Morphosyntax
    - And its “conceptual domain” data type “String”, but no list of possible grammatical functions
    - Name status: admitted name
  - *syntacticFunction* **is** a DC in ISOCAT ([DC-1507](#))
    - With only an English definition
    - In profiles Syntax and MorphoSyntax
    - Its “conceptual domain” datatype “String” and 3 possible values are distinguished
      - Subject ([DC-1391](#)), indirectobject ([DC-1310](#)), directobject ([DC-1274](#))
      - Twice: once for profile “Syntax”, once for profile “Morphosyntax”
    - Name status: nothing specified

# DC-1274 Problems

- Problem 6: (cont.)
  - “**Syntactic function**” **is** a DC in ISOCAT ([DC-2244](#))
    - With an English and a French definition (French name: *fonction syntaxique*)
    - In profile Syntax
    - Conceptual Domain: data type string, no values listed
    - Name status: standardized name
  - **Dependency** **is** a DC in ISOCAT ([DC-2323](#)) (related, not identical)
  - **Should DC-1296, DC-1507, DC-2244 be considered as synonymous/equivalent?**
- → require link of each term used to a DC
- → ensure that a coherent set of definitions is included in the standard

# DC-1274 Problems

- Problem 7
  - How do I find alternative names for the same concept in ISOCAT?
  - How do I find closely related DCs in ISOCAT?
    - It currently requires a linear manual search... , even across different profiles!!
    - Is there a way of grouping closely related concepts together?
- → add relations between DCs

# DC-1274 Problems

- Problem 8 “direct object”:
  - grammatical relation (relation-reading)
  - word or phrase that bears the grammatical relation “direct object” (phrase-reading)
  - Definition of *transitive* uses the phrase-reading

# DC-1274 Problems

- Problem 9: Even stronger, an “alternative” definition is given:
  - *A noun, pronoun, or noun phrase whose referent receives the direct action of a verb*  
Where the *phrase-reading* is given as the only one possible
  - *Such polysemy shifts occur very often in natural language*
    - E.g. morphosyntax = (1) a particular level of grammatical description, (2) study of this level of grammatical description, and (3) properties at this level of grammatical descriptions (e.g. “*the morphosyntax of this word is quite interesting*”, and idem for other levels of grammatical description)
- **Recommendation:** Conventions on how to deal with such cases (preferably in a systematic manner) must be agreed upon and applied systematically
- **Recommendation** ISOCAT should provide means for supporting a systematic use of polysemous variants.

# DC-1274 Problems

- Problem 10
  - There are often also definitions and examples in other languages, e.g. French.
  - These have similar problems as the English definitions, and
  - they usually differ from the English definitions (i.e the English and French definitions are not just translations of each other), e.g. in the case of “transitive”
- → require that definitions in different languages are translations of each other, not completely new definitions

## Answer?

- Can I map my DC *overgankelijk* to [DC-1405](#) (E. *transitive*, Fr. *transitif*)?

I really do not know!

- So I will probably be forced to make a new DC in ISOCAT (which would be bad for semantic interoperability!)

# Structured Elements (1)

- ISOCAT has no provisions for this except for Strings (sequences of Characters) and REs over strings
- Attribute Value Pairs (AV-Pairs)
  - Attribute is a DC
  - Value must be
    - of attribute DC type and
    - from attribute DC Conceptual Domain
- Records/AV matrixes
  - Which AV-Pairs are possible/mandatory for noun, verb etc

## Structured Elements (2)

- Lists
  - e.g. HPSG SUBCAT attribute:  $[NP_{nom}, NP_{acc}]$
- Trees/Tree Models
  - E.g. [DUELME](#) database (Dutch Multiword Expressions)
  - [SAID](#) (LDC2003T10)
  - Treebanks

# Structured Elements (3)

- Sets
  - E.g. set of verbpatterns in Rosetta
  - Subcat patterns Alpino:
    - {intransitive, transitive, *pc\_pp(aan)*} (*breien* 'to knit')
- Parameterized values
  - E.g. Alpino: *pc\_pp(aan)*
  - i.e. prepositional complement of syntactic category PP with *aan* as head

# Mapping & Structure (1)

- Mapping of DCs often requires structure
- Structures are also needed if there is to be a pivot
- Examples
- Combination: Atomic DC → A-V pair combination:
  - *Transitive* maps onto `thetavp=vp120` & `synvps=[synNP]` & `caseAssigner=True` (ISOCAT → Rosetta)
  - RBR → `partOfSpeech=adjective` & `degree=comparative` (Penn Treebank → ISOCAT)
  - VVIMP => `partOfSpeech=verb` & `verbClass=main verb` & `mood=imperative` (STTS Tagset => ISOCAT)

## Mapping & Structure (2)

- List: Atomic DC → List:
  - Transitive* → [NP<sub>nom</sub>, NP<sub>acc</sub>] (ISOCAT, Alpino → HPSG)
- Combination → parameterized value
  - synPREPNP in synvps & prepkey1=*aan* → *pc\_pp(aan)*  
(Rosetta → Alpino)
  - (*in fact : subcats U= {pc\_pp(aan)}* )

## Mapping & Structure (3)

- Union: German Adjectives
- Morphosyntactic features:
  - Gender (3), Case (4),
  - Number (2), Declension type (3)
- In theory  $3 \cdot 4 \cdot 2 \cdot 3 = 72$  distinctions
  - Gender is neutralized in plural
  - So:  $3 \cdot 4 \cdot 1 \cdot 3 + 4 \cdot 1 \cdot 3 = 36 + 12 = 48$  distinctions
- Only 5 forms are used: *e*, *er*, *es*, *em*, *en*

## Mapping & Structure (4)

- Map *en* to a union of a combination of morphosyntactic features:

- *En*  $\Leftrightarrow$

- m sg acc str V m sg gen str V n sg gen str V dat pl str V
- dat sg mixed V gen sg mixed V pl mixed V m sg acc mixed V
- dat sg weak V gen sg weak V pl weak V m sg acc weak
- (using underspecification for gender in some cases)

## Mapping & Structure (5)

- Conclusion:
  - One can often not map DCs in isolation
  - But must map whole entry (record) to a new entry (set of entries)
    - Entry= lexicon entry or word occurrence+annotations in text corpus
  - Or even complexer: multiple entries → multiple sets of entries
  - Additional means are needed to provide structures
    - Does LMF provide part of this?
    - What about annotated text corpora (incl. treebanks)?

# Data Category Sets

- For each coherent data category set a DC must exist to identify it. E.g. *STTS*, *Penn tagSet*, *CGN tagSet* in the value domain of the DC *morphosyntacticTagSet*
- ISOCAT must represent/group them as a set
  - Is *Profile* suited for this and sufficient?
  - Can multiple profiles be assigned? (it appears so)
  - Do we need finer distinctions
    - E.g. profile is currently used mainly for different levels of grammatical description

# Implicit Semantics: 'Mime'-like approach

- Pragmatic option
- Resource/Tool 1 specifies: tagSet=STTS
- Resource/Tool 2 specifies tagSet=STTS
- Match is found → interoperability
- Semantics of STTS is left implicit
  - identity of semantics suffices
- Occurs often, is simple and must be supported

# Semantics Explicit: Mapping:where?

- Option 1: Directly in an XML file

- “PID can be embedded in the schemata of linguistic resources”
- <http://www.csc.fi/english/pages/neeri09/workshop/materials/windhouver.pdf>, slide 8
- Is that sufficient?
  - XML element must be mapped onto DC, but also
  - Attributes and values inside XML elements

- Option 2: in separate files

- Needed for commonly used coherent subsets (e.g. Penn Treebank Tagset, STTS, CGN Tagset, etc.)
- To avoid duplication, inconsistency, etc.
- Is that possible now?

# Interlingua is not enough!

- S1's DCs map to ISOCAT DCs A B C D E
- S2's DCs map to ISOCAT DCs D E F G H
- Then there is only a translation/interoperability for DCs mapping to ISOCAT DCs D and E
- Others should be cases of real (not just notational) differences
  - Granularity is essential here
  - Approximations between different but closely related (combinations of) DCs should be supported

# Interlingua is not enough!

- Extreme Example: syntactic selection
  - LFG selects arguments in terms of Grammatical Function (GF) (and NOT syncat)
  - HPSG selects arguments in terms of Syntactic Categories (syncat) (and NOT GFs)
  - → Intersection in ISOCAT IL will be empty

# Interlingua is not enough!

- Extreme Example: syntactic selection (continued)
  - Rosetta selects arguments in terms of **GF and syncat**
  - Alpino selects arguments in terms of **GF and syncat**
  - But:
    - Alpino always selects for subject
    - Rosetta never selects for subject
  - → Intersection in ISOCAT IL will be empty

# Interlingua is not enough!

- Extreme Example: syntactic selection (continued)
  - Split syntactic selection into two parts in IL
    - Syntactic selection for subject
    - Syntactic selection for complements
  - Then:
    - Intersection will still be empty for subject selection
    - Intersection will be large for complement selection

# Mapping: theoretical differences (1)

- Rosetta classifies intransitive verbs into 2 subclasses
  - Unaccusatives
  - Unergatives
- Alpino makes no such distinction
- What will the standard require? Leave it open?
- Rosetta lexicon entry => Alpino lexicon entry (via ISOCAT)
  - Easy (less specific), just drop some info
- Alpino entry => Rosetta entry
  - Strictly spoken not possible (insufficient info)
  - But *intransitive* & conjugated with *zijn* gives a very good approximation (I expect >>95% correct)
  - Such approximative mappings must be supported

## Mapping: theoretical differences (2)

- Rosetta

- Any dependency between syntactic property and meaning/translation => separate lexical entries
- Metal (Alpino) can have 1 entry with multiple semantics/translation conditioned upon syntactic property

- Mapping from Metal (Alpino) → Rosetta requires creation of a set of Rosetta entries for each Metal entry

- What are the conventions here in the pivot? Does LMF prescribe anything here?

## Mapping: theoretical differences (3)

- Example (in English)
- *Look at, look after, look for, etc,*
  - can all be in one entry in Metal/Alpino
    - Are distinguished only in transfer
  - Must be separate entries in Rosetta
    - Are distinguished already in syntax



# Challenges

## 1. Challenges



# How to Proceed: Action Items

1. How to Proceed: Action Items
2. To be filled during the workshop

# Recommendations (Linguist)

- Test ISOCAT against your resource / tool / project and report on virtues / problems/ omissions / questions!!
- Be hesitant in adding new DCs and definitions. It is not easy to do this correctly.
- If you add a new DC, provide 1 and only 1 definition!
- Clarify and illustrate your definition
- For DCs and their definitions: Use reliable and respected sources that are relevant to the point
- Stick to the [DCR Style Guidelines](#)
- Try to find DCs that have other names but the same meaning (spelling variants, synonyms)
- Ensure that all data categories used in your definition are defined as a DC and refer to their PIDs
- Take good care of polysemy shifts in meaning of one and the same expression. Make the conventions you adopt explicit!
- Use the “Conceptual Domain” facility much more than is done so far (it is currently underutilized)
- Select definitions of DCs preferably from one and the same common RELIABLE source to increase overall coherence of the DC set in ISOCAT, and certainly do this for DCs from the same “conceptual Domain” .

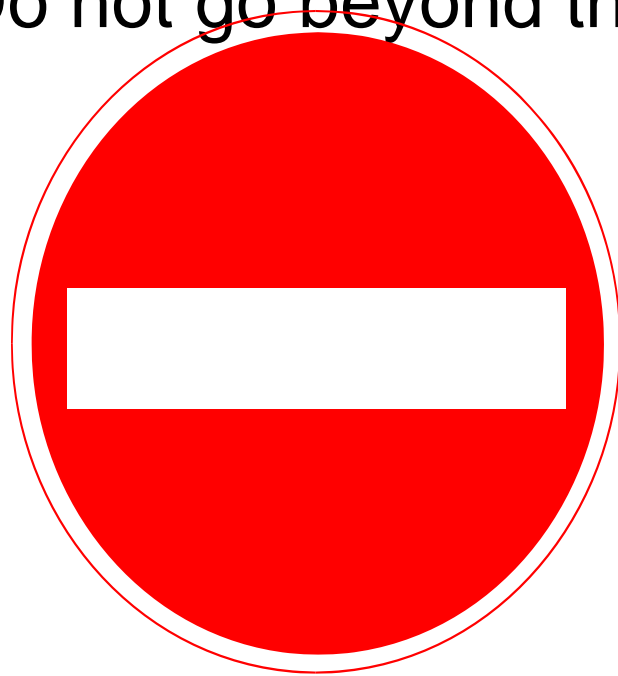
# Recommendations (designers)

- Add [a system](#) that checks whether all terms used in a definition refer to a DC in ISOCAT and if not mark the DC as “incomplete”
- Add facilities that allow the user to deal with systematic cases of polysemy in a systematic manner
- Warn for DCs in the same “Conceptual Domain” that originate from different sources
- Provide means to group closely related concepts together (to find synonyms, and to prevent creation of new synonyms)
- Add relations between DCs (I understand this is already in the plan), esp. to implement an “is very similar to” relation
- Warn if all terms in the definition come from grammatical level (profile?) A when defining a term in grammatical level (profile?) B
- Add a formal link to the superordinate DC (can be done with relations)
- Add automatic notification to TD groups upon change
- Make the search better (search on **comparative** does not find **degree**)
- Make (the whole interface but esp.) text search in the web interface much faster!!!
  - **Google is faster on the whole WWW than the search facility is on ISOCAT!**
- Correct the source references (which are often wrong)

# Recommendations (TD Groups)

- Provide more documentation on the definition of fields and their possible values in ISOCAT DC entries
  - The website has **DCR Style Guidelines and Web Interface — Overview** (and some presentations)
- Check sticking to the [DCR Style Guidelines](#) (currently often violated)
- Be organized for many additions/modifications/questions in 2010Q2!
- Respond quickly and adequately to newly added/modified DCs!
- Study and make recommendations on grouping of related concepts
- Study and make recommendations on how to deal with (systematic) metonymic shifts in meaning in linguistic data categories
- Be very critical before approving proposed DCs!
- Be very critical for “alternative definitions” in the same DC-entry
- Do not accept differences in definitions between different languages (adapt them or split the DC up into two DCs)

Do not go beyond this slide!



# Are all definition terms in ISOCAT?

- Use a list of function words (stop list) for a language to identify the content words in a definition
- Each content word in a definition must
  - Be part of a span marked as *definiendum*, or
  - Be part of a span marked as *connector*, or
  - Be part of a span hyperlinked to an ISOCAT DC, or
  - Be part of a span marked as “intentionally not defined”
- Otherwise the definition is marked as incomplete

# Are all definition terms in ISOCAT?

[BACK](#)

- Example:
- a `<definiendum>` transitive verb `</definiendum>`
- `<connector>` is defined as `</connector>`
- a `<a href=DC-1424>` verb `</a>` that
- `<a href=DC-****>` syntactically selects `</a>`
- a `<a href=DC-2256>` noun phrase `</a>`
- with the
- `<a href=DC-1296>` grammatical function `</a>`
- `<a href=DC-1274>` direct object `</a>`