Structured (CMDI) Vocabularies

Jan Odijk

2021-12-14

Clariah

Overview

- Vocabularies
- Example 1: ToolTasks
- Example 2: CLAPOP faceted search
- Example 3: GOLD Ontology
- More examples
- References

Vocabularies

- Vocabularies containing vocabulary items (labels for concepts or data categories) with an explicit meaning are required (ISOCAT, CLARIN Concept Registry)
- Their meaning is in the form of a URL (Linked data), precise but long and difficult to memorise
- There are other properties, e.g., a definition, …
- Vocabulary Items are short strings that are mnemonically useful
  - but ideally not real words of English or other common natural languages to avoid the horrors of natural language

Vocabulary Use

- Using the vocabulary:
  - stimulate reuse of existing vocabulary items
  - avoid unnecessary proliferation of vocabulary items
  - easy for a user to find the right vocabulary item
  - user needs to look only at relevant vocabulary items to select from,
  - avoid the user to be bothered by irrelevant vocabulary items
  - long (>15) list of vocabulary items, and repeated search: forget it
    - User will create his/her own new vocabulary items
  - See Odijk, 2009, 12-13!!, who already pointed this out

Additional Structure

- Creating and maintaining the vocabulary
  - List too long ➔ the vocabulary creator gets lost and will create unwanted duplicates (e.g. both lexicon lookup and lexicon search in ToolTasks)
- Additional Structure (e.g., a hierarchical ontology/taxonomy, or additional properties):
  - A small hierarchical taxonoy can be used to reduce the items to be searched for to short lists (<=15), even though this has to be done in a number of steps (which should therefore be kept small, so the depth of the taxonomy should be < 5).
  - The ontology is there just for organising the vocabulary items, facilitating their maintenance and search in it. It has no meaning and nothing is claimed with it.
  - Multiple mutually incompatible taxonomies can exist in parallel.
  - The meaning of the vocabulary items is not dependent of the taxonomy
- Interfaces
  - Interfaces that enable creating, editing or searching for vocabulary items must use these ontologies to support these processes, e.g. in CMDI editors, Component Registry, search portals, ...

Example 1: ToolTasks

- CMDI Profile ClarinSoftwareDescription (CSD, clarin.eu:cr1:p_1342181139640) uses a component ToolTasks (clarin.eu:cr1:c_1505397653781) with an element toolTask with a closed vocabular of more than 90 values (and growing). [Odijk, 2019]
- Additional Structure is desired to be able to select a value correctly
- Proposal for a small ontology
- Not implemented in the element, not even using the poor man's option (as slash-separated strings) as an ad-hoc and temporary solution.

Example 2: annotationInfo (MetaShare)

- Element annotationType in component annotationInfo (clarin.eu:cr1:c_1381926654461) > 50 values
- E.g.
  - discourseAnnotation-audienceReactions, discourseAnnotation-coreference, discourseAnnotation-dialogueActs
  - semanticAnnotation-semanticRelations, semanticAnnotation-semanticClasses, semanticAnnotation-semanticRoles
- Implemented using the poor man's option (as hyphen-separated strings) as an ad-hoc and temporary solution. Alphabetic sorting groups the values then semantically
- Similarly: AnnotationType element in AnnotationType component (clarin.eu:cr1:c_1527668176048) in CSD

Example 3: CLAPOP

- CLAPOP offers faceted search for software
- For a few facets a small hierarchical taxonomy was defined (e.g. History/Art History, History/Oral History)
  - Used in the interface
  - Reduces the number of options
  - Restricts additional options to a particular category
  - Implemented in Drupal by Daan Broeder
  - Less important because used in combination with faceted search

Clariah

Example 3: LIDIA & Excalibur: Gold Ontology

- LIDIA: database and search interface into a database of Linguistic **Dia**gnostics: a database with arguments from the linguistic literature that have been adduced to argue in favor of or against a linguistic property or construction (grammatical relations, syntactic categories, part of speech tags, etc. etc.)
- EXCALIBUR: Glossing service based on a database of glossed examples
- Probably will use the GOLD linguistic ontology http://linguistics-ontology.org/, supplemented by other vocabularies, and use the ontology in the interface to facilitate searching for linguistic concepts.

More examples

- Standards Committee Google sheet : long list of data formats
- Classified by category (column a) and family (column b).
  - E..g Matriska has category Audio and family TEI (row 15): Audio/TEI. Makes maintenance of the list and searching in it much easier
- Profiles and Components in the component registry:
  - Long unstructured lists ➔ unnecessary proliferation ➔ even longer lists
  - With more structure this could have been avoided /reduced

Clariah

Thanks for Your Attention!

References

Jan Odijk 2009. Data Categories and ISOCAT: some remarks from a simple linguist. NEERI, FLaReNet/CLARIN Standards Event, Helsinki September 30 2009.

Jan Odijk 2019. Discovering software resources in CLARIN. *Selected papers from the CLARIN Annual Conference 2018.* Linköping Electronic Conference Proceedings 159: 121–132

Clariah

DO NOT ENTER

The Horrors of Natural Language

- Words have associations (slightly different for everyone)
- Words have a (common-sense) meaning
- Words are often ambiguous / polysemous
- Words are too long ➔ redundant (➔ abbreviations, acronyms)
- Words have synonyms
- Words are specific to a language

- ➔use codes that are non-words instead! (cf. the ISO language codes)