Vocabularies for CMDI
The modeler's perspective

The CMDI Sessions - Session 2

Penny Labropoulou

penny@athenarc.gr

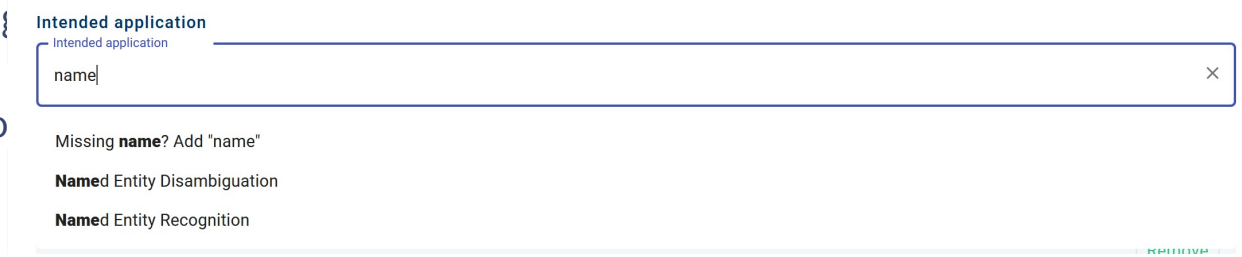
Athena RC/Institute for Language and Speech Processing

www.athenarc.gr

Using controlled vocabularies

Supporting flexibility (1)

- Finding the right balance between normalization/standardization and flexibility in response to providers' preference for free text and the emergence of new values
  - Closed vocabularies: typically a short list of values; values could be added straight into the metadata profile
  - Open vocabularies: e.g. tool/service functions, annotation types, data formats, size units
  - a set of recommended values
  - new values added by users in the metadata editor used as (a) suggestions for next users and (b) collected as candidates for next releases of the vocabulary
  - ➔ need for continuous updatin track of the versions
  - ➔ need for (a) an editing tool fo



clarin:el

Using controlled vocabularies

Supporting relations and other information

- Flat lists vs. thesaurus vocabularies: taking advantage of
  - relations for the benefit of users (e.g. synonyms for free text search, hierarchical relations for structuring, etc.)
  - definitions, but also other hyperlinks and notes for extra information
  - names in other languages in support of multilinguality

  → need for importing vocabularies in CMDI with all properties, even if these are not displayed in the CMDI profiles

  → for vocabularies created by modelers, need for a separate editing module; for standardized formats (e.g. SKOS, RDF/OWL), an open-source editor could be used

  → for all vocabularies and properties, a common representation format is required in order to be fed into the CMDI-based applications

Selecting vocabularies

External vs. Own vocabularies (1)

- Created by modelers
  - If it's simple and with a limited set of values, this can be done through the current CMDI feature for adding a flat list of values BUT it cannot be shared with others

  ➔ need for sharing even flat lists of values

- Selecting the right vocabulary from among external vocabularies (maintained by other organizations)
  - e.g. languages: ISO 639-3, BUT BCP47 relies on ISO 639-1 and there are more vocabularies with attracting features: lexvo, glottolog, ethnologue, …
  - freedom for modelers to use different vocabularies?

  ➔ need for supporting multiple vocabularies for the same domain in CMDI

clarin:el

Selecting vocabularies

External vocabularies vs. Own vocabularies (2)

- "Extending" the vocabulary with values that are not covered or combining similar vocabularies
  - languages: regional variants, language varieties, … ➔ adding free values on another element? but consumers want to view all values under the same element
  - licences: SPDX list of licences, and proprietary licences but also community-specific licences (e.g. CLARIN licences) ➔ creating a mixed recommended vocabulary with values from an external vocabulary and values added by the modeler
  - organizations: combining GRID with national registries of organizations
- "Pruning" the vocabulary to a subset of values
  - LT taxonomy: fine tuned to the requirements of different audiences (e.g. focusing on text applications vs. all types of LT tasks) by selecting items and feeding the subset to the schema

  ➔ need for an editing module that outputs vocabularies in the CMDI-required format

clarin:el

Deploying vocabularies

Integration of vocabularies

- Deployment options into CMDI
  a. copy & paste of values (but what about relations?)
  b. direct link to an endpoint (e.g. REST API, SPARQL endpoint, etc.) offered for the controlled vocabulary
  c. link through a CMDI-supported vocabulary service
- Option b
  - Not all vocabularies offered via an endpoint; some are offered as downloadable files (e.g. SPDX licences)
  - Risks of broken links for deprecated values, vocabularies no longer maintained
  - Speed issues
  - Different formats: SKOS, JSON, XML, CSV, OWL, … , different metadata models, different access protocols & mechanisms
- Option c
  - Hybrid import of vocabularies (which ones?) into a CMDI shared space with links to the source external vocabularies? (e.g. as CCR concepts with a "sameAs" relation to the original values)
  - Who is responsible for adding the vocabularies and maintaining them?
  - How de we keep them updated with the original sources?
  - How to establish relations between similar or derived vocabularies?

Modeler's requirements

Summary

- editing module for creating, editing, curating vocabularies
  - doesn't have to be CMDI-integrated but should output CMDI-compliant format
  - *support for collaborative editing in a shared space?*
- service for deploying vocabularies in metadata profiles and applications
  - support for import of multiple vocabularies in a common format into the CMDI vocabulary service
  - *support for adding other endpoints if they satisfy format and security specifications?*
  - supporting thesaurus relations passing through CMDI profiles to the applications that use them
  - keeping track of versions of vocabularies and their use in the metadata profiles
  - keeping up-to-date with new releases of external vocabularies

clarin:el

Thank you!

Questions?

clarin:el