CLARIAH vocab registry: a first sketch

Overview

A CLARIAH vocabulary registry would provide an overview of vocabularies[1] relevant to the CLARIAH community. Currently several overviews and collections exist:

1. https://github.com/CLARIAH/awesome-humanities-ontologies
2. https://github.com/TriplyDB/YALC
3. More general LD registries like https://lov.linkeddata.es/ or https://prefix.zazuko.com/
4. More general vocabulary collections like http://bartoc.org/

For CLARIAH we would like to start by creating a one stop shop, but with a growth path to authoritative recommendations. These recommendations can be based on:

- Actual use of the vocabulary in real world datasets
- Positive or negative experiences of the CLARIAH community with the vocabulary.
- Reliability of the vocabulary, e.g., no semantic drift between versions and persistence of the URIs, sustainability promise/track record of who is publishing it
    - Vocabulary owners should provide a statement on the persistence and maintenance of the vocabulary, to enable users to decide whether or not it fits in their preservation policies.

Recommendation status is impacted by bad persistency, should move to a CLARIAH centre for active maintenance to be recommendable

Additional functionality might make the registry even more valuable:

- Store and provide access to copies of the vocabulary, incl. older versions,
    - but make clear what is the canonical version, that's never the one in the registry
    - provide access via upcoming defacto APIs, e.g. SKOSMOS or NDE termennetwerk or ELMA or TPF or combinations thereof
        - But maybe this is only useful for a subset of the vocabularies
    - align with other vocabulary registries and/or archives, e.g., https://archivo.dbpedia.org
- Have a safety net in place, e.g. the LD Proxy from the Sustainability IG.

Usage scenarios

The CLARIAH community is in need of guidance with regards to L(O)D vocabularies.

- What are the recommended vocabularies, i.e. what should you use related to
    - a specific domain, e.g. linguistics, or
    - functionality, e.g. geo tagging, discoverability, cooperation/data exchange?
- Are there discouraged vocabularies, i.e. what shouldn't you use?

[1] We use vocabularies here in the broad sense, i.e., including both the LD equivalences of a term list for an open or closed controlled vocabularies, most commonly expressed in SKOS, entity lists, most commonly expressed in schema.org and more schema-like vocabularies, most commonly expressed in RDFS, OWL or SHACL.

- Who has experience with a specific vocabulary and can you ask for advice?

The CLARIAH infrastructure needs one common API to access vocabularies so tools/services don't need to implement multiple (vocabulary specific ones).

The CLARIAH infrastructure needs stable access to a specific version of a vocabulary, i.e., changing semantics should be a supervised decision and basic LD characteristics, like resolvability, should be guaranteed as much as possible.

Platform selection

1. Collect requirements and prioritize them
    a. Metadata (= topic of Findability IG, although their aim is broader, i.e. findability for arbitrary datasets)
        i. Description of a vocab
        ii. Reference or resource
        iii. Domain
        iv. Function
        v. License
    b. Sustainability Assessment
        i. Persistency
        ii. Versioning policy
            1. Semantic drift
    c. Experiences
        i. Positive
        ii. Negative
    d. Usage
        i. Links naar datasets
        ii. Statistics
    e. (Facetted) search
        i. On domain and/or function
        ii. Fulltext
    f. Effort
2. Collect platforms
    a. OntoPortal
    b. Prefix server
    c. Bartoc

Development

Pilot

Setup an instance and populate it with vocabularies from YALC and the awesome humanities ontologies list.

Extension 1

Get to know the software and see if we can make the basis for making recommendations visible:

1. Usage info
2. Positive and negative experiences
3. Reliability info

Some of these need some additional tooling, e.g., some way to get usage info and refresh it regularly.

Extension 2

Implement a Common API (based on the DANS/NDE proposal currently under construction) on cached copies of specific vocabularies.

Extension 3

Automatically classify vocabularies so metadating a new vocabulary becomes less manual work.