Symposion KI 2013

# Background Annotation of entities in Linked Data Vocabularies

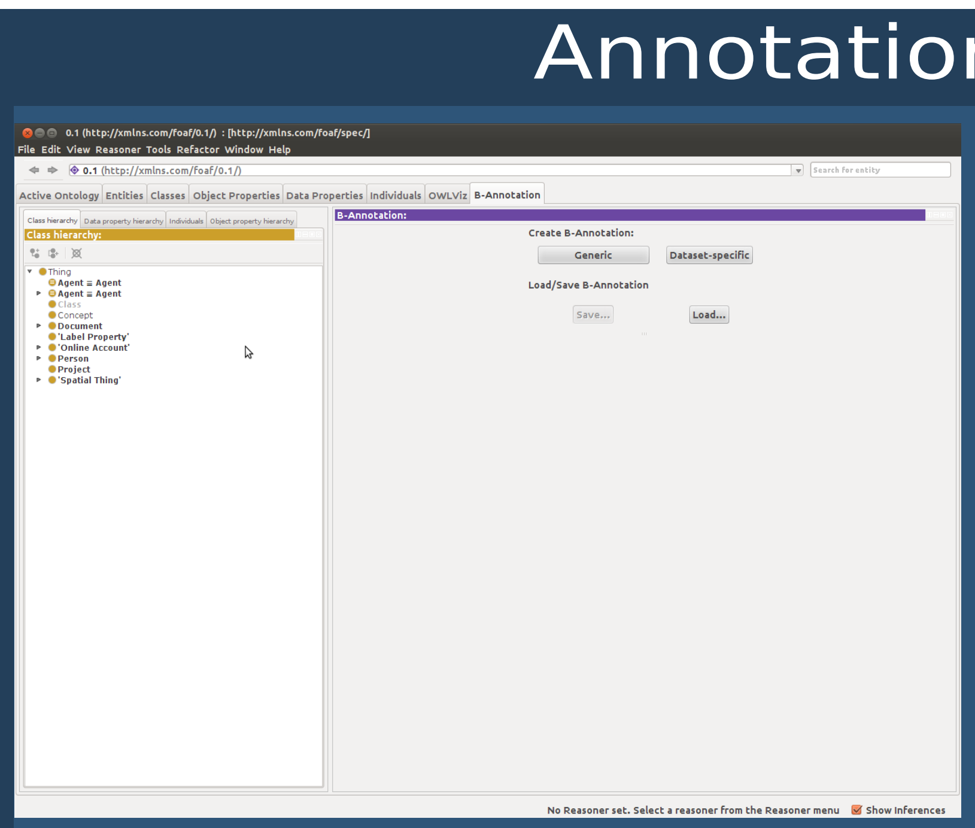
Simone Serra

Supervisor: Vojtěch Svátek

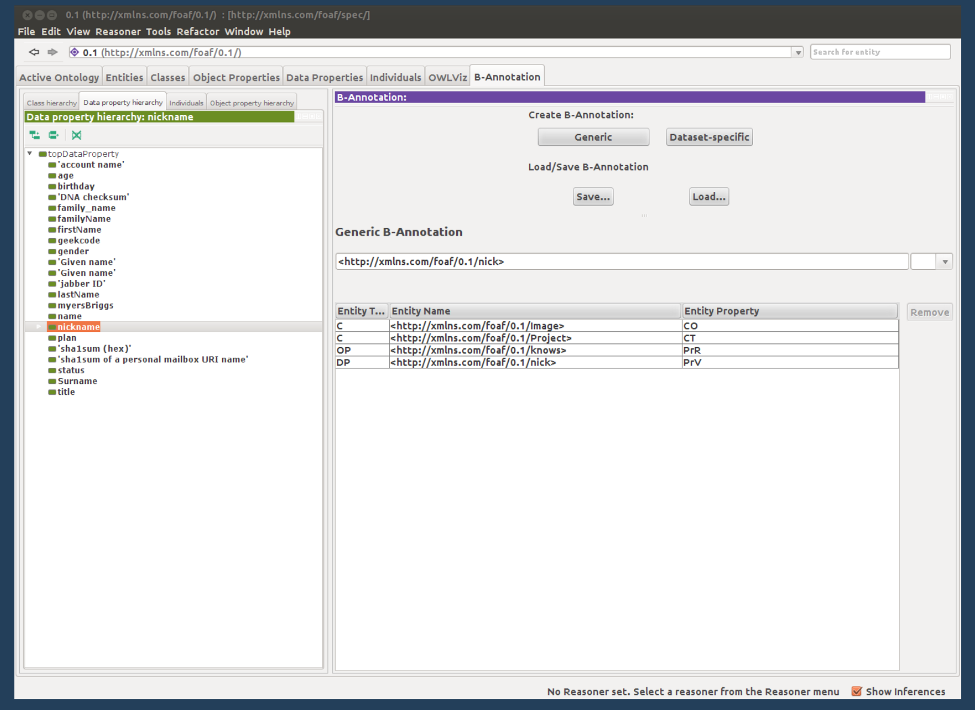
## Motivation

- Address Linked Data vocabularies shortcomings in modelling real-world entities
- Implement a workflow for annotating vocabularies entities via a formal ontology (PURO)
- Investigate the use of Linked Data dataset statistics and summaries for improving the annotation process

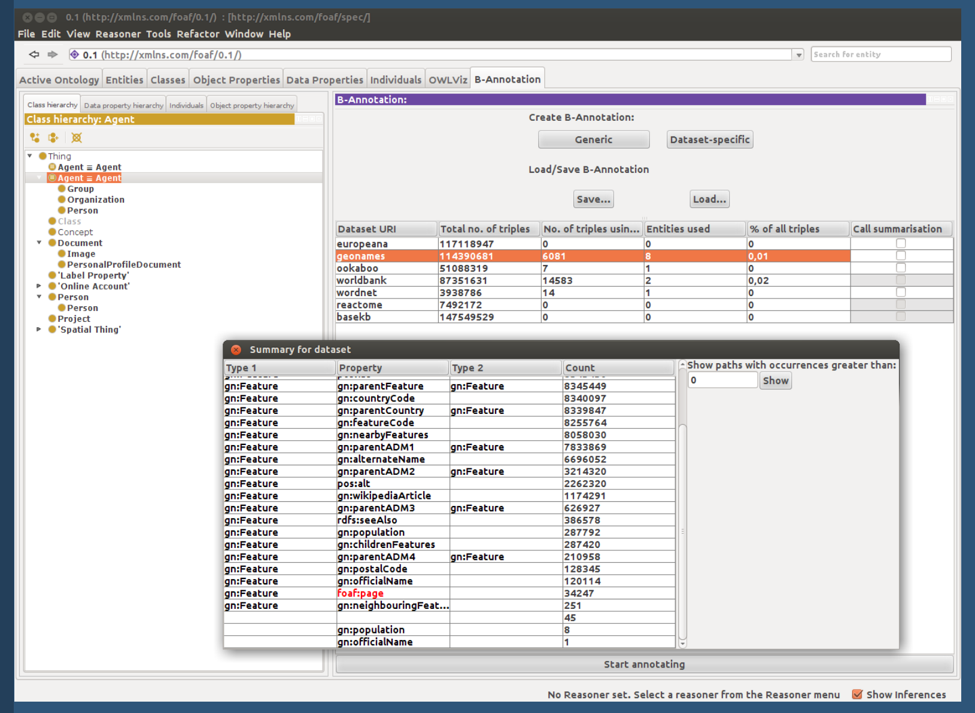
## Annotation workflow



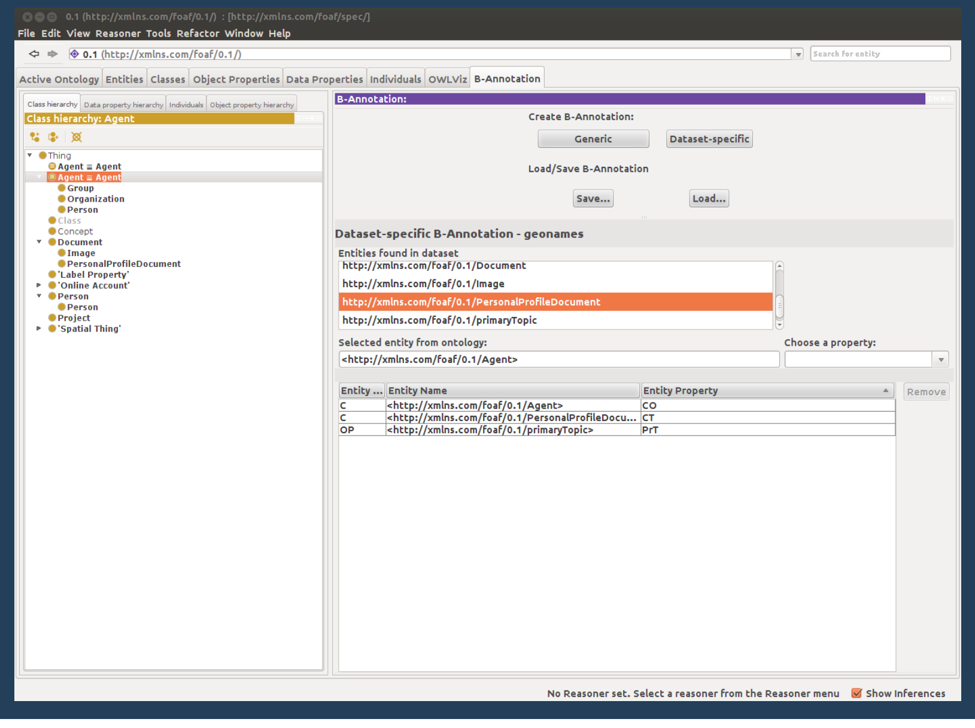
1. The user loads the vocabulary into Protégé and selects Generic or Dataset-specific annotation. A previously saved annotation can be loaded into the plugin



2. The user annotates the vocabulary by selecting the entities from the left pane and choosing the appropriate PURO annotation property. Available PURO properties depend on the kind of entity selected (class, data property, object property)



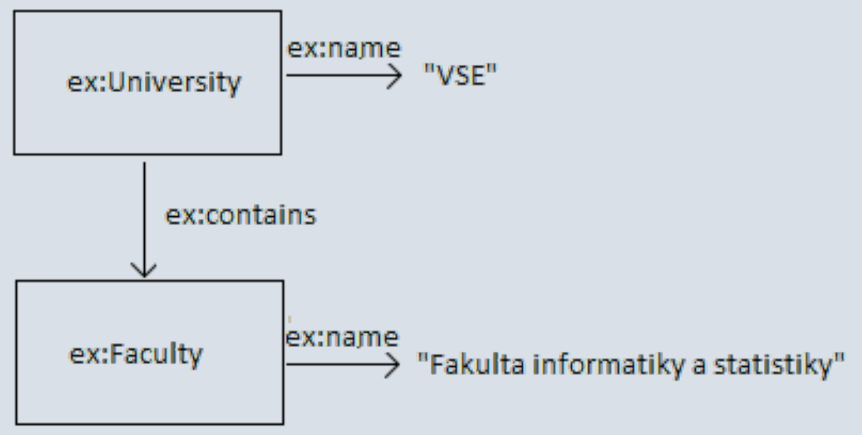
3. In dataset-specific mode, the list of currently available dataset statistics is shown. The dataset summary shows the most recurrent paths and their occurrences. Paths can be filtered according to the number of occurrences.Path elements in red show that the entity is present in the vocabulary.



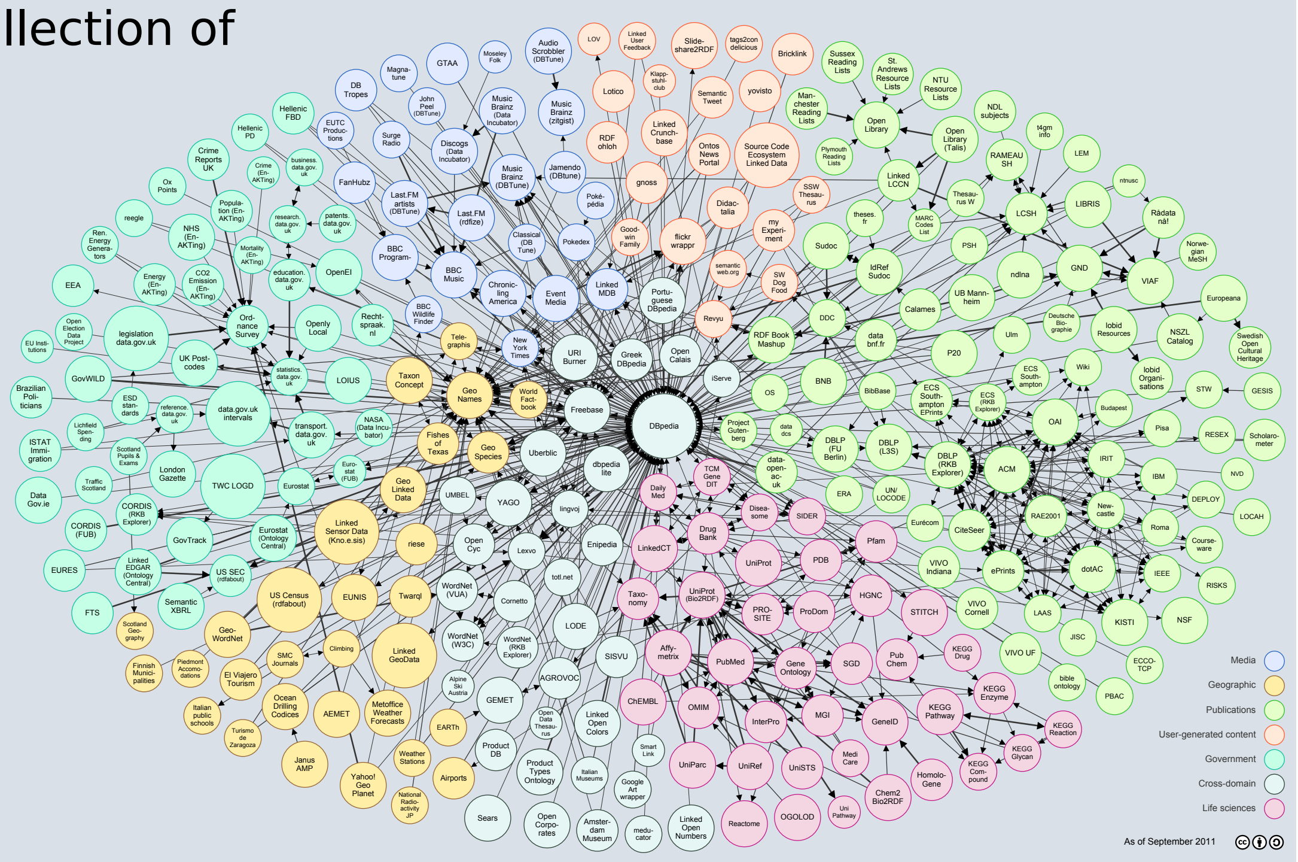
4. When the user clicks on "Start Annotating", the annotation continues as (2). The list of vocabulary entities found in the dataset is also shown.
5. User can save the annotation locally in RDF/XML format

## Background Information and terminology

- Semantic Web: A single model of data integration and representation of real-world entities so that information can be shared and re-used across different platforms and communities [1].
- Web of Documents vs Web of Data: Instead of connecting web pages through the use of URLs (Uniform Resource Locator), in the Semantic Web one data item in one page to point to another data item through global identifiers called Uniform Resource Identifiers, or URI.
- Linked Data: a method for publishing and sharing such structured data on the Web.
- RDF (Resource Description Framework): Model for representing data on the Semantic Web in the form of a subject - predicate - object statement (RDF triple). Example: *ex:University ex:name "VSE"*
- RDF statements can be linked together to create a RDF graph



- Linked Data datasets: RDF graphs containing millions of RDF triples, each representing knowledge about a specific field, interlinked via RDF links
- Linked Data vocabularies: collection of definitions of useful terms, relationships, properties and constraints on the use of these terms
- Analogue to the notion of "ontology"
- Vocabularies help resolving ambiguities by providing previously agreed-on "ontologies" of a specific knowledge domain
- Defined using RDF
- Self-descriptivness: vocabulary terms are identified by URIs i.e. they are dereferenceable and linked to their own definitions
- Links between vocabularies help mapping terms that refer to the same concept



## The PURO ontology

- Development of Linked Data vocabularies not driven by rigorous ontology design methodologies
- Example from the Music Ontology (mo)

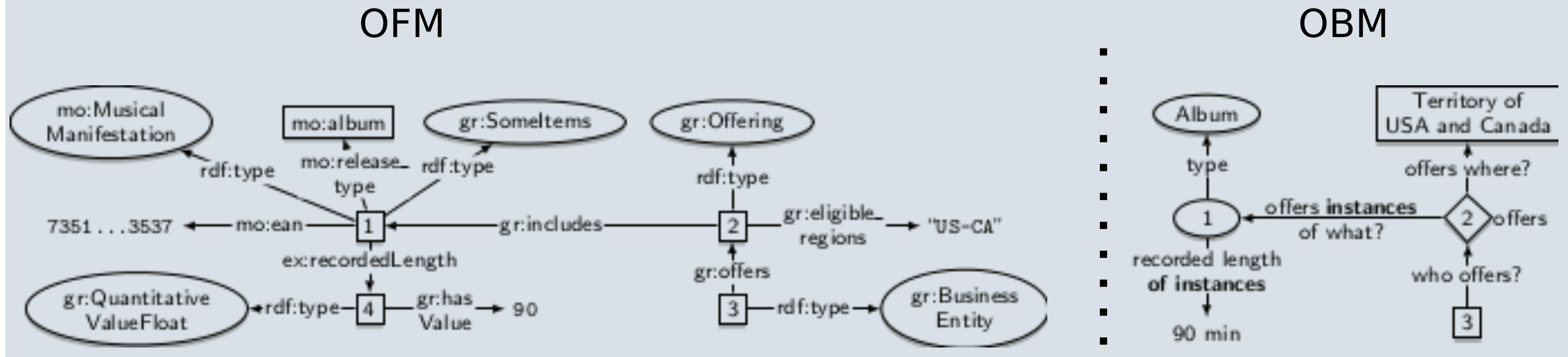
***Artist mo:primary_instrument mo-mit:Cello***

- Intended use mo-mit:Cello: instrument type, not a concrete instrument
- mo-mit:Cello is an ontology individual, but its is treated as a class modelled by an ontology individual
- Vocabulary is used and linked to a variety of datasets. A dataset may then contain the following statement:

***"Yo-Yo Ma" mo:primary_instrument "1712 Davydov Stradivari"***

- The original intent was to state that the artist's primary instrument is the cello, but instead, it specifies a concrete cello by its name => incoherence
- A formal ontology modelling methodology can be used to "cure" this kind of problems

- Given the mass of data present in dataset, their direct refactoring is nearly impossible
- Use annotations: assign labels to entities from Linked Data vocabularies that indicate useful ontological distinctions
- PURO: An annotation ontology that represents the structure of ontological distinctions via an Ontological Background Model (OBM) and the associated visible model of vocabulary as an Ontological Foreground Model (OFM)[2]



| | *Object* | *Relationship* | *Valuation* |
|---|---|---|---|
| ***Particular*** | $\mathcal{B}$-object<br>(individual) | $\mathcal{B}$-relationship<br>(*3 possibilities*) | $\mathcal{B}$-valuation<br>(data prop. assert.) |
| ***Universal*** | $\mathcal{B}$-type<br>(class) | $\mathcal{B}$-relation<br>(*3 possibilities*) | $\mathcal{B}$-attribute<br>(data property) |

- Two basic distinctions: **Particulars** vs **Universals** and **Objects** vs **Relationships**
- Valuation distinction takes care of LD vocabularies modelling the assignment of quantitative data values to individuals

- Two annotation types:

  Generic: based on the vocabulary textual description

  Dataset-specific: based on the specific use of a vocabulary in a LD dataset
- Statistical indicators for a sample of datasets are provided:
  - total number of entities
  - number of vocabulary entities in the dataset
  - percentage of vocabulary entities with respect to total number of entities
- Dataset summaries shows the most recurrent triple paths of length 1 and 2 in the dataset, along with the number of occurrences

  ***subject-property-object*** (length 1 path)

  ***subject-property-object(subject)-property-object*** (length 2 path)
- Most recurrent paths (knowledge patterns) represent the dataset core knowledge

## Conclusions

- Too early to evaluate benefits and real-world usage. The application is still in a prototypical form. It will be released to a close group of Linked Data experts for trial
- Major areas of improvements in the summarisation (paths of length 3) and statistics for a larger number of datasets of smaller size
- Further collaboration with the Sindice initiative (The Semantic Web Index) for gathering dataset statistics
- Experimental support for a wizard-style decisional tree in assisting the annotator during the annotation workflow
- Two foreseen areas of application

  1. Test the effectiveness of mappings between vocabularies of different datasets in approaches such as the R2R framework

  2. The annotation of dependent resources in Concise Bounded Descriptions of entities (CBD) to be used in a data sampling process for data mining purposes

[1] ALLEMANG, D., HENDLER, J. Semantic Web for the Working Ontologist. Effective modeling in RDF, RDFS and OWL. 2nd ed. San Francisco, CA: Morgan Kaufmann Publishers Inc., 2011. ISBN:0123735564 9780123735560.
[2] SVÁTEK et al. Mapping Structural Design Patterns in OWL to Ontological Background Models in Proc. K-CAP. 2013

Thanks to **Alexander Erlich** for the Inkscape template, found at http://www.inkscapeforum.com/viewtopic.php?f=8&t=3622