Using Ontologies as Heuristic Tools: Sources in the History of Philosophy and Their Interpretation in the Semantic Web

In my presentation, I will first criticise the ongoing debate on whether folksonomies or ontologies are more powerful tools for the organisation and representation of knowledge. In this, I rely on a basic insight from the philosophy of language regarding the mutual dependence of the meaning of concepts and the meaning of propositions. Based on this criticism, I want to argue for a new methodology of building ontologies ‘inductively’. I will finally apply this methodology to a use case, namely the semantic representation of sources in the history of early modern philosophy.

In the debate between ‘ontologists’ and ‘folksonomists’, ‘ontologists’ base their view mostly on the conviction that the world as it is is ‘classifiable’. Their ‘folksonomist’ opponents have some reservations, mostly because they operate in domains in which the practical applicability of ontological schemata is at best dubious. I contend that this debate is based on a fundamentally flawed semantic assumption: a 'building block' account of the meaning of propositions as a mere sum total of the meaning of their parts. This overemphasis on the 'compositionality' of propositional meaning leads to a neglect of the complementary aspect of 'contextuality', i.e. the insight that the meaning of concepts as the constituent parts of propositions is at least to a certain extent determined by the very propositions they appear in. This means that nonpropositional uses of concepts and singular terms are parasitical on the use of concepts and singular terms in propositions.

Hence, an act of folksonomist 'tagging' contains at best the assertion of an incomplete proposition, so that, from the point of view of the philosopher of language, 'tags' are systematically underdetermined, because in themselves such labels cannot clarify the exact relation between the tagged entity and the label used to describe or characterise it. Conversely, ontologists cannot presume that the meaning of the concepts they employ is fully stable over uses of these concepts in different propositional contexts. This is particularly true as soon as we assess the practical value of semantic web technologies in the context of the digital humanities.

Prime examples for the practical success of the top-down approach favoured by the 'ontologists' can be found in the sciences, based on their clear commitment to the existence of natural kinds. But as soon as we enter the sphere of the cultural, the surplus of such a methodological approach is more questionable, the impressive achievements of ontologies like CIDOC CRM notwithstanding. The practical value of such ontologies for the digital humanist is limited, since the complexity of the domain is mirrored in a similarly complex conceptual schema that may be accessible only to 'classification experts' like librarians or cultural heritage

professionals, experts that often share the conviction of the natural scientist that their respective domain is 'classifiable in principle'.

But although both the ontologist and the folksonomist argue from faulty premises, there is no need to abandon all attempts to represent knowledge about the cultural in machine-readable form. These insights should rather prompt us to revise our common strategies of ontology building for cultural artifacts. Changes to how we devise representations of conceptual structures would allow us to respect the semantic interdependence of concepts and propositions, to use ontologies as heuristic tools for the exploration of an unknown conceptual space, and to combine the explicitness and precision of ontologies with the flexibility of folksonomies. My talk discusses the basic features of such a methodology under the heading of 'inductive ontology building'.

For this, it is first helpful to contemplate again the question why we need (or want) an ontology of a given domain. We can use an ontology as an instrument for structuring information about a domain or we can apply it for automated reasoning tasks. However, in both cases, the basic unit of information within an ontology is the proposition, so that the primary class of entities in an inductive ontology are propositions and the facts they express. So we first need a basic and sparse 'ontology of propositions' in order to represent these particulars. Such an ontology would know only one class, namely propositions. It would contain three relations, namely 'has-subject-term', 'has-relation-term', and 'has-object-term'. In this model concepts in a proposition are not to be interpreted as atomic building blocks, but as dependent aspects of the proposition they are used in. Therefore, a proposition in this model should be expresssed not in one, but in three RDF triples:

- 'S:Proposition-p R: 'has-subject-term' O:literal
- 'S:Proposition-p R: 'has-relation-term' O:literal
- 'S:Proposition-p R: 'has-object-term' O:literal

Such a schema respects the basic syntactical requirement of RDF that subject terms and relation terms in RDF statements must be resources, i. e. identifiable via a URI. At the same time, such a form of representation need not make any assumptions whatsoever about the conceptual structure of the underlying domain.

We can then begin to identify e. g. terms within a given domain that are predominantly employed in the subject position. We can extract facts in which different subject terms appear with identical relation or object terms, inferring either identical extensions or class hierarchies. By identifying 'outliers' standing in no relation to other terms, we can circumscribe the domain in question more clearly and ask whether facts that do not share terms with any other collected fact really belong to the domain under examination.

An example may help to clarify the advantages of such a strategy. At http://emto-nanopub.referata.com, I have begun to extract doxographical content from early modern sources, notating these excerpts as triples. In my presentation, I will talk about 62 propositions that were asserted by 16th and 17th century Spanish philosophers while debating the proper definition of philosophy and relate them to the vocabulary available in the ontology language OWL for describing the relations between the concepts employed. OWL offers resources to describe and characterise the relation between classes, properties, subclasses, subproperties, equivalent properties, or instances. An ontology of how 17th century Spanish philosophers conceptualised philosophy would start from the assertion that philosophy is a habit, i.e. as an acquired property of the soul, so that we could tentatively ‘philosophy’ as an OWL ‘subproperty’ of OWL ‘property’ 'habit', 'habit' as a subproperty of the property 'quality', and 'quality' as a property of the class 'soul'. Philosophical disciplines could again be introduced as subproperties of the property 'philosophy'. I will then show how to model the relation between a proposition and an author entertaining this proposition and how this model can be used to make ontologies author-relative.

In closing, I will briefly compare this ‘digital’ way of reading sources to traditional approaches of reconstructing historical sources using the complete apparatus of first-order predicate logic. The main advantage of the methodology proposed here is a reduction of complexity. We are not concerned with the logical reconstruction of arguments, but only with the collection and analysis of statements. This reduction of complexity is particularly relevant when charting unknown territory. Since mass digitisation has made available vast amounts of completely unknown sources in the last decade, the history of early modern philosophy faces enormous challenges in coming to terms with this material. In the light of current research on OCR technologies for early modern prints it can be expected that a significant percentage of these sources will soon be available as full text, so that we may be able to develop tools for an automatic extraction of RDF triples describing their content. The methodology proposed here may then prove to be a fruitful strategy for turning this content into semantically rich information.