Observations from Development of Social Reality Ontology for a KDD Application

Vojtěch Svátek

Department of Information and Knowledge Engineering,
University of Economics, Prague, W. Churchill Sq. 4, 130 67 Praha 3, Czech Republic
e-mail: svatek@vse.cz

Abstract. We describe the process of building an OWL ontology of social reality, which serves as support for generating explanations of hypotheses discovered via KDD. Three modelling problems related to existing or potential ontology design patterns are discussed in more depth: the way of expressing descriptive features amounting to 'social phenomena' and 'aspects of life', the issue of using a single property with not entirely homogeneous domain/range, and the degree of abstraction of domain/range restrictions.

1 Problem and Context

Althought the span of reported ontology *application types* and *domains* is growing, there are still large areas that are almost untouched by existing research.

Among the application types, let us focus on ontological support for *discovery* of new knowledges in masses of *tabular data*, i.e. KDD (Knowledge Discovery in Databases). One of the main outcomes of the first *Workshop on Knowledge Discovery and Ontologies* [2] was that the role of prior knowledge is underestimated by the KDD community, and even if this knowledge is used, it is rarely underpinned by a clear conceptual model. However, [3] demonstrated that ontologies can be beneficial in nearly all phases of the KDD (more specifically, association mining) cycle, starting from domain and data understanding, through the semantic interpretation of discovered hypotheses, and ending by exposing the hypotheses on the semantic web, e.g. in the form of annotated textual reports [5].

Here we pay main attention to the middle phase, in which the ontology is to provide guiding 'templates' for the human expert[1] who attempts to interpret the discovered knowledge (though it is desirable that the same ontology is used in all phases of the KDD cycle). Such a 'non-canonical' type of intended application may have some impact on the required structure of ontology in question. For example, a suitable feature of UMLS Semantic Network and Metathesaurus used in our earlier medical data mining experiment [3] was the existence of

[1] A more ambitious goal would be to generate plausible explanations fully automatically; we are however afraid that this is beyond the capabilities of state-of-the-art knowledge engineering techniques.

named non-taxonomic relations. However, these relations were mostly defined at a higher degree of abstraction, which decreased the usefulness of results. In general, it seems that the degree of formality of the ontology required for such purpose is, in most application scenarios, a medium one. Since we mostly deal with data schemas and with (induced) general hypotheses rather than with individual data objects, we do not always care for fully transparent set-theoretic semantic. On the other hand, what matters is the existence of a sensible number of distinct concepts and named relations among them: chains of concepts and relations then can serve the human interpreter as 'templates' for possible explanations of hypotheses. A language such as RDF/S will have the required expressivity. However, if we do not preclude the future use of mined hypothesis by more sophisticated reasoners, even advanced (OWL) constructs could be worth employing.

Among the application domains of semantic web ontologies, domains from within *social science* are extremely rarely encountered, especially when compared to e.g. business or medicine. The main reason (aside the lower level of financial support) might be the lack of clear borders for individual domains, which makes the traditional 'hugeness problem' of ontology design particularly discouraging. Designing a sufficiently complete formal ontology reflecting the subject of e.g. history or sociology potentially usable for an unforeseen application would amount to formalising many volumes of encyclopaedic knowledge.

We however believe that some sort of ontology could still be designed with a particular *application* in mind, which helps filter out concepts that are theoretically relevant but practically unusable. Moreover, given a restricted set of start-up concepts explicitly marked as relevant for an application, we can build a (clearly, incomplete) ontology in a *bottom-up* manner, by adding just the 'connecting' components needed to link the start-up concepts. The set of start-up concepts can be determined e.g. with respect to an existing database scheme, which is likely to beavailable in the KDD application type mentioned above. An ontology with better coverage could then arise by gradual merging/mapping of independently created models.

In this paper, we focus on the intersection of both novel problems (KDD application type and social reality domain), and attempt to draw more general conclusions and potential guidelines from our ontology-building experience. Published or potential ontology patterns used to express descriptive features and domain/range constraints over certain properties are discussed in more extent.

2 Existing Resources and Past Work

As mentioned above, the general notions of *social reality* have rarely been subject of application-oriented (formal) ontology engineering. One exception is the recent work by Boella & van der Torre [1], who developed an upper-level model of social reality centred around the concept of 'agent'. Although we preferred a bottom-up ontology building approach, in view of soon obtaining solid match with data

analysed via KDD, we plan to adopt this or similar upper-level model as root of our ontology in the future.

Some small parts of our ontology are similar to existing more specific resources. For example, the properties of *countries* are covered by the CIA Factbook ontology[2]; this is however a high-level data model rather than a structured ontology and covers few aspects of the social life in the country itself.

As far as the research on OWL *design patterns* is concerned, we repeatedly refer below to the SWBPD WG Note on 'Representing Specified Values in OWL...' [7]. The current work also informally follows up from earlier general considerations on OWL design patterns presented in [8].

3 Formalisation in OWL

3.1 Overview

A specific aspect of our ontology was the (pre-dominantly) *bottom-up* style of its construction. Both the *ontology* and the *dataset* used for association discovery had the same seed material: the questionnaire posed to respondents during the *opinion poll* mapping the 'social climate' of the capital city of Prague in Spring 2004. The questionnaire contained 51 questions related to e.g. economic situation of families, ways of earning money and dwelling, attitude towards important local events, political parties or media. Some questions consisted of aggregated sub-questions each corresponding to a different 'sign', e.g. "How important is X for you?", where X stands for family, politics, religion etc. Other questions corresponded each to a single 'sign'. While the *dataset* was straightforwardly derived from the individual 'signs' (each becoming a database column), the *ontology* first had the form of *glossary* of candidate terms from the text of the questions. In conformance with most ontology engineering methodologies [4], the terms were then divided into candidates for *classes*, *relations* and *instances*, respectively. Then a *taxonomy* and a structure of *non-taxonomic relations* was built, while filling additional entities when needed for better connectivity of the model or just declared as important by domain expert. Since our main intended application was interpretation of hypotheses discovered via KDD, we did not specifically target at complex axioms. However, since the use of OWL as more expressive language does not represent an obstacle to reasoning in RDFS as less expressive language, we decided for the former. This allowed us to represent some useful constructs not available in RDFS, in particular, disjointness of classes, non-identity of instances, and functionality of properties. The validity of the constructs was quite obvious and did not bring (much) additional modelling effort.

The fact that the same textual questions were basis for both the ontology and the dataset made the subsequent *mapping* between the two straightforward[3]. A bottleneck of this approach could obviously be *over-specialisation* of

[2] `http://www.daml.org/ontologies/245`

[3] This contrasted to our earlier experiments with UMLS (as independently developed, pre-existent ontology) and a medical dataset [3].

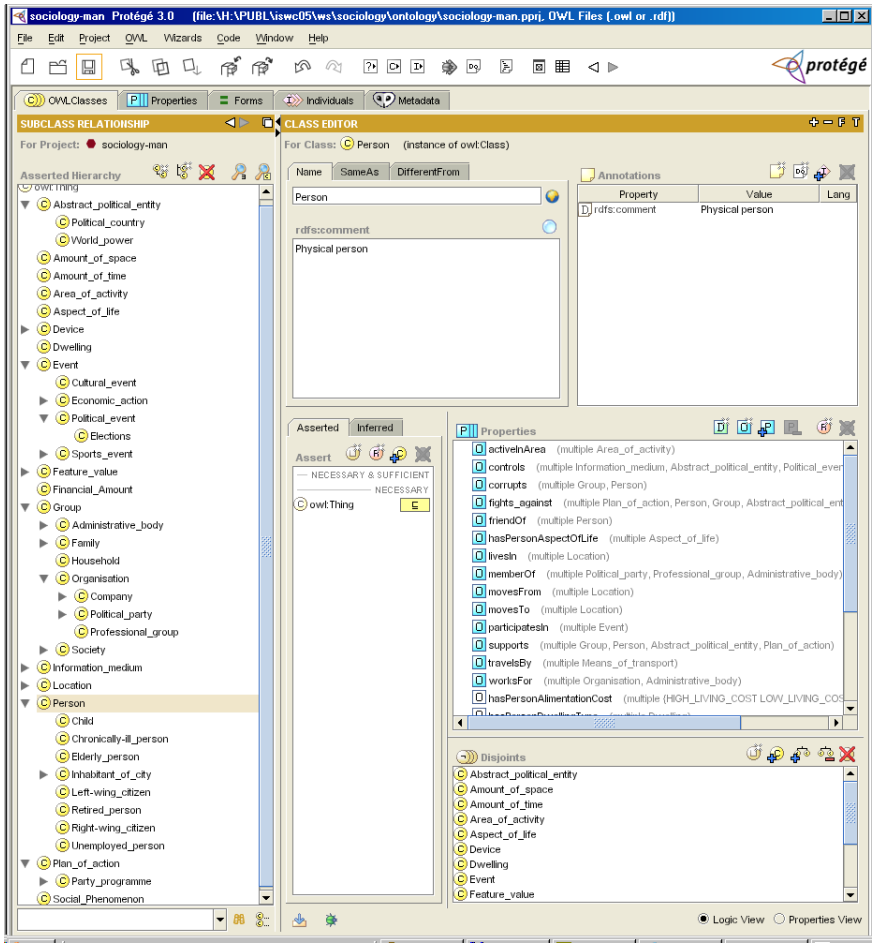

Fig. 1. View of the ontology in Protégé

the ontology with respect to the concrete sociological study addressed by the poll. However, the author developed the ontology following informal consultations with *two* sociology experts, of whom one was unfamiliar with the study. This hopefully reduced the bias and enabled to yield an ontology that to some degree approximates the 'mega-domain' of (municipal) social reality.

The current version of the ontology, eventually formalised in Protégé, consists of approx. 100 classes, 40 relations and 50 individuals; although the principles outlined in this paper are relatively stable, it is still occasionaly being enriched with new entities. A Protégé window showing parts of the class hierarchy plus the properties of class `Person` is at Fig. 1. Actual experiments with matching the ontology with the *empirical associations* discovered by the LISp-Miner tool [6] are described in another paper [9].

During the development of the ontology, we took into account the already published 'design pattern' documents by W3C SWBPD working group`http://www.w3.org/2001/sw/BestPractices/OEP`. The following two subsections reflect our experience with applying certain patterns in our particular case, and also identify possible missing aspects of the patterns.

3.2 Modelling Descriptive Features and their Relationships

Various (transient) features of a society, i.e. *social phenomena* such as 'richness', 'economic growth' or even 'quality of enterpreneurship' are clear candidates for modelling via *properties with specified values*, see the SWBPD note [7]. The domain of such properties (e.g. `hasRichness`, `hasEconomicGrowth`) is `Society`, and the range is always a class that is subclass of the generic `Feature_value` class (such as `Richness_value`). For our purposes, binary value sets (consisting of instances such as `HIGH_RICHNESS`/`LOW_RICHNESS` or `GOOD_JOB_AVAILABILITY`/ /`POOR_JOB_AVAILABILITY`) are sufficient, since they can often be identified with polarised answers of respondents of opinion polls. As suggested for the design pattern [7], we can capture the mutual exclusion of opposing values using `owl:differentFrom` and declaring the property as functional. From the point of view of our application, the data column corresponding to answers of a poll question such as "Do you think that economic growth is beneficial for job availability?" can easily be mapped on properties `hasEconomicGrowth`, `hasJobAvailability`, as well on relevant individuals used as their values, namely `GOOD_JOB_AVAILABILITY` and `HIGH_ECONOMIC_GROWTH`.

With the notion of *aspect of life* (associated with a person), we arrived at an analogous situation as with 'social phenomenon' (associated with the society). They can be modelled to large degree in parallel: namely, while e.g. the availability of jobs (for people) in general is a social phenomenon, availability of jobs for a particular person is an aspect of life. We modelled the aspects of life with set-valued properties such as `hasPersonJobAvailability`, `hasPersonLivingCost` etc[4]., which mostly have the same set of values as their society-level counterparts (under the assumption that the semantics of the relation is the same whether applied on a particular person or on the 'aggregation' of persons in the society).

The pattern might however become insufficient if we want the ontology to contain relations among social phenomena (or aspects of life) and instances of classes such as `Person` or `Group`, as chaining such relations is the core of the KDD-oriented application. For this purpose, we need to explicitly introduce the `Social_phenomenon` class. Their instances then would naturally be all instances of classes such as `Richness_value` or `Economic_growth_value`. The generic property linking `Society` to `Social_phenomenon` then is `hasSocialPhenomenon`, and all specific properties (such as `hasRichness`) are its subproperties. The situation is analogous for the class `Aspect_of_life` and generic property `hasAspectOfLife`.

An example of OWL code for the extended pattern is at Fig. 2.

[4] Note that in order to distinguish them from properties referring to the society, we adopted the naming convention with `Person` as part of the property name.

```
<owl:Class rdf:ID="Richness_value">
   <rdfs:subClassOf rdf:resource="#Feature_value"/>

<Richness_value rdf:ID="HIGH_RICHNESS">
   <owl:differentFrom>
      <Richness_value rdf:ID="LOW_RICHNESS">
        <owl:differentFrom rdf:resource="#HIGH_RICHNESS"/>
      </Richness_value>
    </owl:differentFrom>
  </Richness_value>

<owl:FunctionalProperty rdf:ID="hasRichness">
  <rdfs:subPropertyOf rdf:resource="#hasSocialPhenomenon"/>
  <rdfs:range rdf:resource="#Richness_value"/>
  <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>
  <rdfs:domain rdf:resource="#Society"/>
</owl:FunctionalProperty>

<owl:Class rdf:ID="Social_Phenomenon">
    <rdfs:subClassOf rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
    <rdfs:subClassOf>
      <owl:Class>
        <owl:oneOf rdf:parseType="Collection">
          <Economic_situation_value rdf:ID="BAD_ECONOMIC_SITUATION"/>
          <Enterpreneurship_quality_value rdf:ID="BAD_ENTERPRENEURSHIP"/>
          <Housing_situation_value rdf:ID="BAD_HOUSING_SITUATION"/>
          <Economic_situation_value rdf:ID="GOOD_ECONOMIC_SITUATION"/>
          <Enterpreneurship_quality_value rdf:ID="GOOD_ENTERPRENEURSHIP"/>
...
        </owl:oneOf>
      </owl:Class>
    </rdfs:subClassOf>
 </owl:Class>

<owl:ObjectProperty rdf:ID="hasSocialPhenomenon">
  <rdfs:range rdf:resource="#Social_Phenomenon"/>
  <rdfs:domain rdf:resource="#Society"/>
</owl:ObjectProperty>
```

Fig. 2. OWL code for the 'richness' feature and associated entities

3.3 Properties with Heterogeneous Domain/Range

Some of the relationships that link especially the classes `Group`, `Person` and/or `Social_phenomenon`[5], such as 'supports', 'controls', 'decides about', 'informs about' or 'fights against', have somewhat vague semantics, which slightly varies depending on the nature of their arguments. For example, 'supporting' a group (e.g. enterpreneurs) or a phenomenon (e.g. security of citizens) means that the subject aims to increase (or not decrease etc.) the *welfare* of the group or the *validity* of the phenomenon (for the given society), respectively. The relationships might even have a *second-order* flavour. For example, 'controlling' a group probably means that the subject (to some degree) controls in which relationships the object group may participate; this of course goes beyond the expressive power of OWL. We decided that such relationships are worth modelling *as single property* with somewhat *heterogeneous* domain/range (rather than refining each to many specific ones), at least for our 'lightweight' (in terms of reasoning) application, since this makes the ontology more comprehensible for a human. However, other applications might require a different approach.

3.4 Setting the Generality of Domain/Range Constraints

Another tricky issue, which is certainly not unique to our case, is the *level of generality of domain/range constraints*. Many binary relations can be applied on *multiple* but not *all* subclasses of a certain class. For example, the domain of our property `informsAbout` could be either:

- `owl:Thing`, i.e. we do not want to explicitly exclude any kind of entity from being able to inform about something
- the union of a few apparently relevant high-level classes, such as `Person OR Group OR Media`; we can assume that instances of 'most' their conceivable subclasses are able to inform about something; we could also introduce an upper-level class such as `Active_entity` and use it for this purpose
- the union of many finer-grained concepts, e.g. (`Politician OR Journalist OR Political_party OR Administrative_body OR Company OR Media OR Political_country`); we can assume that instances of *all* their conceivable subclasses are able to inform about something.

If the abstraction level of a domain[6] constraint is kept high, there is still the possibility to define zero cardinality of the property just for undesirable subclasses. For example, we might want to declare that many different groups (companies, political parties, families, but also unforeseen ones) can inform about something, but e.g. professional groups can't, because they are not capable of a concerted activity. We can then declare `Group` as the domain of `informsAbout`, and at the same time create the axiom

[5] When speaking about `Person` or `Group` as domain or range of these relationships, we also mean their subclasses, cf. the following subsection.

[6] Due to disbalanced roles of domain and range in OWL (class-centric axiomatisation of ontologies), the following approach cannot be used for range constraints.

$$\texttt{Professional\_group} \sqsubseteq (=0\ \texttt{informsAbout})$$

For the purpose of generating potential explanations of KDD hypotheses, it is more practical to have properties with disjunctions of finer-grained classes in both domain/range, as such disjunctions subsume fewer concepts overall than e.g. a single root class does: the space of possible explanation chains is thus pruned. For this reason, we used the last approach from the list above. The disadvantage of this approach for DL reasoning is that from the validity of relation instance (such as `informsAbout(FINANCIAL_TIMES,HIGH_ECONOMIC_GROWTH)`) we can only derive that `FINANCIAL_TIMES` belongs to an anonymous class represented with such a disjunction; in contrast, with the middle approach from the list above, we could e.g. derive its membership to the *named* class `Active_entity`.

3.5 Miscelaneous Problems and Proposed Solutions

In this subsection we enumerate some other problems we encountered. Most of them are connected to our (broad) domain. We are aware that some of them are so general that they have certainly been subject of prior work in philosophical (or even applied) ontology; we plan to align our observations with such work in the future.

- The notion of `Society` is not easy to grasp. Traditionally, it encompasses a (smaller or larger) group of people and their social relationships. For simplicity, however, we identified it with the respective group, hence made it simply subclass of `Group`. This relies on the assumption that the relationships are to some degree modelled externally to the `Society` concept, i.e. via object properties linking the concepts that are its parts or members.
- Even the decision whether a certain concept will be part *or* member of `Society` was not easy. For example, by intuition, some organisations such as political parties seem to be direct members of society rather than just its parts. However, to preserve the set-theoretic soundness of the ontology, we only considered persons as society members. Further research on parts/wholes patterns[7] is likely to address such issues.
- Particularly tricky is the concept of `Country` (and the same may hold even for `city`). Most existing ontologies only consider it as subclass of `Location`. Countries however also exist in political sense, which should be kept distinct; for example, one cannot expect from a location to *compete* with another location (which we had to cover in our ontology). We thus decided to create two distinct concepts: `Country` and `Political_country`, and linked them with mutually inverse properties `isPoliticalViewOf` and `hasPoliticalViewOf`, respectively. Note that e.g. the CIA Factbook ontology attributes to 'country' a diverse blend of properties, some (probably most) of which would, from our point of view, be related to `Country` and some to `Political_country`. `Political_country`, though sharing many properties with `Group`, is not a subclass of `Group`, but can be e.g. *controlled* by a `Group`.

[7] `http://www.w3.org/2001/sw/BestPractices/OEP/SimplePartWhole`

3.6 Summary of Pattern Usage in the Project

In this subsection we first summarise the used ontology patterns described in the previous subsections, leaving out the concrete domain context. Namely:

1. The published pattern of *value sets* with mutually exclusive values [7] was adapted to our context by extending it with a 'roof' class (needed as domain/range of important domain properties) subsuming all values as its instances, and with a common super-property.
2. A candidate pattern might be one dealing with the choice between a single property with (somewhat) *heterogeneous domain and/or range* and multiple finer-grained properties with homogeneous domain and/or range.
3. A related candidate pattern might be one dealing with preference for *domain/range constraint generality/specificity*, with trade-off between
   - simplicity and possibility to infer membership to a named class on the one hand
   - elimination of unwanted applications of the property on the other hand.

4 General Reflection on Ontology Design Patterns Usage

In this very short section, we attempt to revisit preliminary ideas from our earlier work, taking into account the ongoing work by SWBPD WG as well as own experience from (not only) the presented project. In [8], design patterns were identified as a vehicle that could help balance the three crucial features of semantic web ontologies: *accuracy*, *comprehensibility* and *reason–ability*. Let us elaborate on that a bit:

1. The usage of patterns, in particular if they were *explicitly* declared in ontology meta-data, could significantly improve the *comprehensibility* of ontologies
2. *Sufficient number* of well-exemplified patterns (and their variants) would help to achieve high *accuracy* of ontologies, as the patterns could reveal to the designers useful combinations of language constructs they were unaware of; they could then more faithfully reflect the state of affairs in the given domain
3. *Application-sensitive* design of pattern *variants* would lead to better *reasonability*, as the designers could then better tune the shape of the ontology towards the main intended application types—while keeping the door open for other types, to some degree.

5 Conclusions and Future Work

We described the structure of a newly-developed ontology for the domain of *municipal social reality*. Attention was paid to application of published *ontology design patterns* and to indication for useful new patterns. The ontology was designed for the sake of a particular application in the domain of KDD, namely,

for testing the possibility of ontology-based support of explanation of hypotheses discovered by an *association mining* tool. We believe that the current ontology, however imperfect it is, could be used for a different knowledge discovery application from the same domain (i.e. in connection with social reality poll data).

In the future, we would like to align our bottom-up-built ontology with some adequate *upper-level ontology*; we believe that this could, among other, lead to validation of further published or even formulation of new ontology design patterns. We would also like to pay more attention to expressing (mainly as relation instances) *additional heuristic knowledge* available in our domain, which can be matched with newly discovered knowledge. Such a(not-yet formalised) knowledge base actually arose in connection with the polls in question, and we would like it to provide more concrete prior knowledge to be matched with mined hypotheses (in a similar way as we used clinical causalities and other heuristic rules in our pre-cursor medical project [3]. Finally, we would like to follow up with our earlier effort to expose KDD results on the semantic web [5].

Acknowledgements

The research is partially supported by the grant no. 26/05 of the Internal Grant Agency of University of Economics, Prague. The author is grateful to dr. Miroslav Flek for providing his expertise in municipal sociology and insight into the principles of 'European Region - Prague 2004' poll, and to dr. Alois Surynek for general consultations on key problems in sociology research.

References

1. Boella, G., van der Torre, L.: An Agent-Oriented Ontology of Social Reality. In: Proc. FOIS'04, Torino 2004, Springer LNCS.
2. Buitelaar, P., Franke, J., Grobelnik, M., Paass, G., Svátek, V. (eds.): ECML/PKDD 2004 Workshop on Knowledge Discovery and Ontologies (KDO-04), Pisa 2004.
3. Češpivová, H., Rauch, J., Svátek V., Kejkula M., Tomečková M.: Roles of Medical Ontology in Association Mining CRISP-DM Cycle. In: ECML/PKDD04 Workshop on Knowledge Discovery and Ontologies (KDO'04), Pisa 2004.
4. Gómez-Perez, A., Fernández-Lopez, M., Corcho, O.: Ontological Engineering. Springer 2004.
5. Lín, V., Rauch, J., Svátek, V.: Content-based Retrieval of Analytic Reports. In: Schroeder, M., Wagner, G. (eds.). Rule Markup Languages for Business Rules on the Semantic Web, Sardinia 2002, 219–224.
6. Rauch, J., Šimůnek, M.: Alternative Approach to Mining Association Rules. In: Lin, T. Y., Ohsuga, S. (eds.), The Foundation of Data Mining and Knowledge Discovery (FDM02), IEEE 2002.
7. Rector, A. (ed.): Representing Specified Values in OWL: "value partitions" and "value sets". W3C Working Group Note, 17 May 2005, online at `http://www.w3.org/TR/swbp-specified-values/`.
8. Svátek, V.: Design Patterns for Semantic Web Ontologies: Motivation and Discussion. In: 7^{th} Conf. on Business Information Systems (BIS-04), Poznan, April 2004.
9. Svátek, V., Rauch, J., Flek, M.: Ontology-Based Explanation of Discovered Associations in the Domain of Social Reality. Submitted paper.