Conceptual components and Ontology Design Patterns extraction across different ontologies

Valentina Anita Carriero

University of Bologna, Bologna, Italy
valentina.carriero3@unibo.it

Abstract. Understanding large ontologies, with diverse semantics and modelling practices, is still an issue, and has an impact on many ontology engineering tasks. While existing methods summarise ontologies by extracting the most important nodes or subgraphs, a complete overview of an ontology, and a comparison between multiple ontologies, are not supported. Based on the hypothesis that ontologies are designed as compositions of patterns, this paper presents a research proposal for developing a method able to extract conceptual components from multiple ontologies and the observed ontology design patterns implementing them.

Keywords: ontology design patterns, ontology understanding, empirical knowledge engineering

1 Problem statement

Due to the open nature of the Web of Data, Knowledge Graphs (KGs) use multiple and heterogeneous schemas corresponding to diverse conceptualisations, which can diverge in expressiveness, granularity, coverage, intended meaning, naming conventions, level of axiomatisation. Understanding large ontologies, and being able to compare different ontologies, is still an issue, and is a crucial preliminary step for performing ontology engineering tasks such as ontology reuse, ontology matching and ontology evaluation. A survey on ontology selection and reuse I conducted [6] shows that solutions for ontology reuse are often adopted on a case-by-case basis, hindering the definition of shared practices.
Ontology summarisation [9] aims at making ontologies more understandable by creating summaries that extract the key concepts and relations of an ontology. However, for many tasks such as ontology reuse there is a need to compare different ontologies at the same time. Moreover, an overall comprehension of an ontology goes beyond its key concepts, and should involve all the *facts* an ontology can represent. I call these complex structures, expressing a relational meaning (e.g. membership, location), *conceptual components*. A conceptual component (CC) is the *intensional* counterpart of OWL implementations (ontology design patterns, ODPs) in actual ontologies, and it groups possibly different ODPs implementing the same component across different ontologies. Two designers may create different ODPs to implement the same CC (e.g. because of

different modelling styles): e.g. the *location* component (*being located at a place*) can be implemented in one or more ontologies as a binary (e.g. hasLocation) and a n-ary relation (e.g. Location) between 3 arguments (Time, Object, Place), or can be specialised as the location of a specific object (e.g. a building).
ODPs, being reusable template solutions to recurrent modelling problems [13], have been proposed as a tool for supporting ontology engineering and reuse: however, they are often used unintentionally and, even if intentional, their use is rarely made explicit e.g. through annotations [3]. While contributing to the development of ontologies on cultural heritage [5, 8], I experienced lack of tools e.g. able to recommend candidate ODPs addressing a specific modelling requirement. I also experienced the manual and time-consuming process of creating well-documented ODPs [4, 7].
Based on these premises, this research project aims at providing methods and tools for supporting ontology understanding, and indirectly other ontology engineering tasks, by (i) identifying modelling problems (CCs) common to multiple ontologies and implemented with specific modelling solutions (ODPs), (ii) analysing these solutions and comparing them with documented ODPs, (iii) automatically annotating them to make them recognisable.

2 Importance

Automatically extracting CCs and their corresponding ODPs from KGs can provide a basis for novel approaches to support ontology engineering tasks.
Ontology selection. An ontology designer may need to reuse existing ontologies for modelling her data. The patterns implemented in an ontology reflect the modelling problems that it addresses, rather than the collections of concepts it contains. Therefore, a modularised, pattern-based and topic-centered description of ontologies should support a better understanding, hence selection of them.
Ontology visualisation. Ontology visualisation is an important tool for working with ontologies. Most of the existing frameworks show node-link views with a focus on class hierarchies [11]. Large KGs limit the usability of these tools. The patterns identified in KGs can be exploited for visualising a KG as a network of conceptual components, making its inspection easier.
ODP-based ontology engineering. Currently, ontology patterns are developed manually and collected in repositories such as ODP Portal. Mostly based on a top-down approach, these repositories unavoidably lack completeness [16] and do not provide information on the actual use of ODPs. Identifying ODPs in KGs can provide actual examples of how catalogues' ODPs are implemented in ontologies, thus documenting their use, and can significantly enrich ODP repositories with new patterns, emerged as a result of empirical analysis.
ODP-based ontology matching. Detecting ODPs implemented in multiple KGs can help designers and ontology matching (OM) tools to generate alignments between ontology fragments, i.e. groups of relations, thus contributing in supporting pattern-based automatic OM procedures and KG linking.

3 Related Work

This research requires tackling different topics that have been investigated in the literature.

ODPs. Pattern-based ontology engineering methodologies, such as [2], strongly focus on ODP selection and reuse, but tool support for this task is still insufficient, despite CoModIDE, a recent Protégé plugin for supporting pattern-based design. To facilitate ODP reuse and documentation, [14] proposes a simple language for annotating ODPs in ontologies (OPLa). Although a richer and more robust version of OPLa is desirable [3], it can provide a basis for automatic ODP annotations in ontologies.

ODPs detection. These methods focus on identifying parts of ontologies reusing s-o-t-a ODPs. [15] finds small evidences of ODP reuse in biomedical ontologies starting from ODPs published in catalogues and tries to find their implementations in ontologies, e.g. checking for import declarations and lexical similarities. The limit of this top-down method is that it ignores emerging patterns.

ODPs discovery consists in exploring ontologies to find frequent repeating structures. [17] uses clustering techniques to detect syntactic regularities, i.e. repetitive structures of axioms within an ontology. The result is a set of axiom generalisations for each cluster, only limited to the logical description of an ontology. [16] proposes a tree-mining method to discover possible recurring axiom patterns as frequent subtrees, where association analysis is used to mine co-occurring axiom patterns, which may indicate emerging ODPs. This method is unable to automatically assess whether a mined pattern is a fragment of a known ODP, and does not take into account cyclic patterns and inferences.

Knowledge discovery aims to detect hidden patterns and regularities in large datasets. [19] proposes a method for automatically providing explanations to data patterns using the background knowledge from the web: patterns at the instance level could confirm patterns at the schema level, and background knowledge could be used to explain such patterns.

Ontology selection and understanding. Users can browse terms from different ontologies on catalogues of ontologies (e.g. BioPortal) and semantic search engines (e.g. Schemapedia). Comparison of ontologies and ODP-based filtering of the results are not supported. Most ontology summarisation approaches [9] look for the most important nodes or subgraphs using centrality measures: they do not adopt an ODP-based approach and do not support comparison between multiple ontologies.

Ontology partitioning. Modularisation approaches (e.g. [1, 12]) split an ontology in non-overlapping modules, that combined together form the original ontology. They mainly focus on logical modularisation on one ontology at a time, and no additional insight about the modules is provided.

Complex ontology matching. Complex ontology alignments [18] can overcome the lack of expressiveness of simple (1:1) alignments. Different approaches for complex OM have emerged in the literature; however, it is still regarded as a challenge, and only alignments between two ontologies at a time are considered.

4 Research questions

My research project aims at addressing the following research questions, driven by the respective hypotheses.

RQ1: What are the Conceptual Components, and the ODPs implementing them, used for modelling knowledge graphs on the web?

- **RQ1.1:** Is the boundary of an ODP identifiable?
- **RQ1.2:** How to assess that fragments from different ontologies address the same modelling problem?

RQ2: Considering the ODPs observed on the web and the ODPs defined in online catalogues of patterns: do they match?

RQ3: How to automatically annotate ODPs used in a KG?

(RQ1) H1: A combination of designated heuristics can support the identification of the boundary of implemented ODPs

(RQ1) H2: Densely connected subgraphs detected in ontologies, grouped based on the terminology used in the literal data describing their ontology entities, may indicate modelling problems addressed by different OWL implementations.

(RQ2) H1: Matching observed patterns with catalogues' ones gives a measure of how much, and how, predefined patterns are reused in practice.

(RQ3) H1: A designated language can annotate ODPs observed in KGs: their components, attributes, relations.

5 Method and Results achieved so far

The intuition behind this research is that an ontology is (un)intentionally developed as a composition of ontology design patterns, intended as modelling solutions observed in existing ontologies, regardless their correctness or quality. I hypothesise that these patterns can be empirically observed based on their density and their vocabulary. I expect (i) the density of their internal connections (i.e. connections between entities of the same ODP) to be higher than the density of the connections between entities from different ODPs; (ii) the combination of the words describing an ODP to evoke the relation it represents (e.g an ODP implementing the location component may possibly include terms such as *place* and *is located at*). Community detection can recognise phenomena such as (i), while text clustering allows me to exploit the vocabulary of ODPs in order to group them based on their similarity, which may indicate they are addressing the same modelling problem.

CH ontologies corpus. As an empirical basis, I built a corpus of schemas modelling Cultural Heritage (CH). The choice of selecting ontologies from this domain is motivated by (i) my previous experience in developing CH ontologies, and (ii) the general complexity of concepts and diversity of terminology

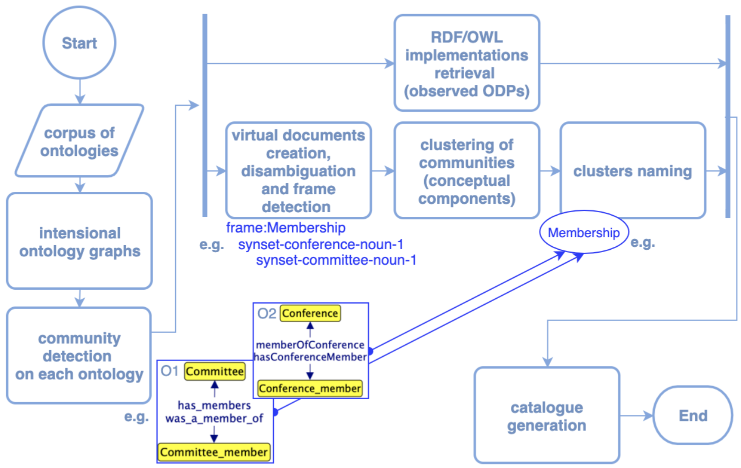

Fig. 1: Method for conceptual components extraction.

of this domain. Indeed, there are many different domains that are related to the already vast domain of CH (e.g. geology for archaeological properties), heterogeneous types of cultural heritage (musical, architectural, etc.), and different cultural institutions have often their own classifications and terminologies. 43 ontologies have been selected from the literature, using a catalogue of general purpose ontologies, i.e. LOV, and publishing an online survey.

Method. For extracting CCs and ODPs from multiple ontologies, I adopt an empirical approach that combines community detection, word sense disambiguation, frame detection, clustering techniques (see Figure 1). The same method can be exploited in order to group ODPs from actual ontologies and catalogues' ones.

Intensional ontology graphs. In order to transform an ontology into a graph that can be processed by community detection algorithms (undirected and unlabelled graphs) while preserving the formalisation of the ontology conceptualisation, I formally represent it as an *intensional graph*, i.e. a graph aiming at encoding the *intensional* level of an ontology. An edge $:p$ is generated between every two nodes that are domain and range of a property $:p$. As for property restrictions on classes, an edge $:p$ links the class local to the restriction and the class in the restriction expression. Finally, for each edge $:p$ between two nodes $:n1$ and $:n2$, two unlabelled edges are generated: the first between $:n1$ and a new node $:n1\text{-}p\text{-}n2$, the second between $:n1\text{-}p\text{-}n2$ and $:n2$, in order to preserve the context of use of $:p$.

Community detection. Community detection splits networks into groups of nodes, such that there is a higher density of edges within groups than between them: since I expect the links between the entities involved in one ODP to be denser than the links between entities from different ODPs, communities could identify possible patterns. A modified version of a s-o-t-a algorithm [10] is run on the intensional graphs: the algorithm recursively splits communities with density higher than the average density of all communities detected at the previous step, since I experimented that this would improve the results. Starting from communities as sets of nodes, the OWL ontology fragments containing these nodes

(classes and properties) are retrieved: for each node, I include the triples asserting its type, domain and range axioms, inverse properties, super- and equivalent classes, restrictions that involve at least one property within the community.
Clustering. If communities are possible ODPs, their vocabulary shall evoke the relational meaning captured by these ODPs: if we cluster the communities according to their vocabularies, we may identify conceptual components that are shared by all of them, potentially. Each community is represented by a *virtual document*: a string concatenating all English labels (or local IDs if no label is present) from its entities, removing repetitions. These documents are disambiguated with UKB and enriched with FrameNet frames[1] (along with their more general frames in the hierarchy) that have a close match with the synsets in the virtual documents, by querying Framester. These virtual documents are given as input to the K-Means clustering algorithm. Each cluster is labeled with the most frequent frame(s) and/or synset(s) in the virtual documents belonging to the cluster, and accompanied by a description, based on the concatenation of all terms representing its communities.
Based on this method, it is possible to build a catalogue of conceptual components and observed ODPs, where each conceptual component links to its associated ODPs within the ontologies: e.g. see the catalogue from the CH corpus.
Experiments. I run the method described in *Method* on both the CH corpus and another corpus (*Conf*) from the Conference evaluation track of the Ontology Alignment Evaluation Initiative (OAEI). Through a manual inspection of the communities, I found that there are recognisable patterns common to many communities, which correlate with the modelling practice adopted for a specific ontology (fragment). Some communities ($\sim$5% of CH and 1% of Conf) can be identified as missing a conceptual unity, because of poor axiomatisation in the source ontology: e.g. communities grouping semantically heterogeneous properties that are not involved in any restrictions. Instead, the majority of the communities have a good level of semantic coherence e.g. representing the acquisition of a cultural property from a previous owner, or the membership in a conference. The clusters identified from both corpora represent different conceptual components, at different levels of abstraction: components as *event*, *categorization*, *intentionally act* are present in both corpora, while other components specific to the domain emerge (e.g. *performing arts* in CH, *award* in Conf). In both corpora, some clusters could be either split or merged.
Even if my goal is not to produce ontology alignments, I evaluated these initial results against the ontology matching (OM) task, based on the hypothesis that, given a pair of similar entities to be aligned, they should belong to either the same cluster or two related clusters. The results of my method are compared with alignments asserted within the ontologies of the corpus; alignments generated by an OM tool with good performances; gold standard alignments on the Conf corpus in the OAEI 2020 conference track. This experiment shows that clusters and their relations may be used to improve the performance of alignment algorithms, significantly reducing the dimension of the task.

[1] Abstract conceptualisations of objects or events that may be evoked by words.

Annotating ODPs. I am working on an extended version of OPLa annotation language, able to annotate an ontology with its implemented ODPs and CCs.

6 Evaluation

The method will be evaluated on: (i) the communities as possible ontology design patterns, (ii) the clusters of communities as possible conceptual components (RQ1). The evaluation will be both manual and automatic. A first evaluation of the initial experiments, consisting in manually inspecting communities and clusters detected on the CH and Conf corpora and in analysing the results in the context of the ontology matching task, has been already performed.

A corpus of ontologies annotated with the ODPs they implement will be built in order to have a gold standard for comparing our automatically detected ODPs. This activity will test if the extended annotation language is fit for the purpose (RQ3). An experiment will be designed that clusters community from ontology corpora and catalogues' ODPs: based on the terminology, observed and s-o-t-a ODPs addressing the same CC should end up in the same cluster (RQ2). The results produced will go also through an evaluation phase by humans, in order to empirically assess the degree of agreement between automatic and manual detection of ontology patterns and conceptual components, possibly designing crowdsourcing tasks. Even if a user-based evaluation of the catalogue would be valuable, it is not an easy user study to be designed, as it would be evaluating key concepts detection, since it requires involving experts in ontology design based on patterns. Therefore, an *indirect* user-based evaluation will be considered, by e.g.: (i) evaluating a pattern-based visualisation tool that will be based on my method; (ii) using my method as a basis for ontology selection tasks.

These activities will be integrated within the EU H2020 project Polifonia, which will realise an ecosystem of computational methods and tools for the European musical heritage (MH) knowledge on the web: this will provide me with an additional large-scale experimental basis.

7 Discussion and Future Work

Preliminary results of this research proposal are already satisfying and show the potential of the method, but they still need to be improved.

Future work will include different activities. The intensional graphs need to include also all class expressions, in order to maximise the detection of relevant communities and limit loss of information. A refinement of the conceptual components (by merge or split of detected clusters) based on heuristics to be defined will be a next step. I will study a method to match observed patterns, detected in actual ontologies, to online catalogues's patterns, which may have a great impact for supporting interoperability. Moreover, the extended version of OPLa will be necessary for automatically annotating ODPs in ontologies.

Acknowledgements. This research is conducted under the supervision of Dr. Valentina Presutti and is supported by the CNR (E-RIHS infrastructure).

References

[1] Flora Amato et al. “Partitioning of ontologies driven by a structure-based approach”. In: *Proc. of IEEE ICSC*. IEEE. 2015, pp. 320–323.

[2] Eva Blomqvist et al. “Experimenting with eXtreme Design”. In: *Proc. of EKAW*. Vol. 6317. 2010, pp. 120–134.

[3] Eva Blomqvist et al. “Considerations regarding Ontology Design Patterns”. In: *Semantic Web* 7.1 (2016), pp. 1–7.

[4] Valentina Anita Carriero et al. “An Ontology Design Pattern for representing Recurrent Events”. In: *Proc. of WOP*. 2019.

[5] Valentina Anita Carriero et al. “ArCo: The Italian Cultural Heritage Knowledge Graph”. In: *Proc. of ISWC*. 2019, pp. 36–52.

[6] Valentina Anita Carriero et al. “The Landscape of Ontology Reuse Approaches”. In: *Applications and Practices in Ontology Design, Extraction, and Reasoning*. Vol. 49. SSWS. IOS Press, 2020, pp. 21–38.

[7] Valentina Anita Carriero et al. “An Ontology Design Pattern for representing Recurrent Situations”. In: *Advances in Pattern-based Ontology Engineering*. SSWS. IOS Press, 2021, pp. 166–182.

[8] Valentina Anita Carriero et al. “Pattern-based design applied to cultural heritage knowledge graphs”. In: *Semantic Web* 12.2 (2021), pp. 313–357.

[9] Sejla Cebiric et al. “Summarizing semantic graphs: a survey”. In: *VLDB Journal* 28.3 (2019), pp. 295–327.

[10] Aaron Clauset et al. “Finding community structure in very large networks”. In: *Physical review E* 70.6 (2004), pp. 066111-1–6.

[11] Marek Dudáš et al. “Ontology visualization methods and tools: a survey of the state of the art”. In: *The Knowledge Engineering Review* 33 (2018), pp. 1–39.

[12] Soudabeh Ghafourian et al. “Graph-based partitioning of ontology with semantic similarity”. In: *Proc. of ICCKE*. 2013, pp. 80–85.

[13] Pascal Hitzler et al., eds. *Ontology Engineering with Ontology Design Patterns - Foundations and Applications*. Vol. 25. SSWS. IOS Press, 2016.

[14] Pascal Hitzler et al. “Towards a Simple but Useful Ontology Design Pattern Representation Language”. In: *Proc. of WOP*. 2017.

[15] Christian Kindermann et al. “Detecting Influences of Ontology Design Patterns in Biomedical Ontologies”. In: *Proc. of ISWC*. 2019, pp. 311–328.

[16] Agnieszka Ławrynowicz et al. “Discovery of emerging design patterns in ontologies using tree mining”. In: *Semantic Web* 9.4 (2018), pp. 517–544.

[17] Eleni Mikroyannidi et al. “Inspecting regularities in ontology design using clustering”. In: *Proc. of ISWC*. 2011, pp. 438–453.

[18] Elodie Thiéblin et al. “Survey on complex ontology matching”. In: *Semantic Web* 11.4 (2020), pp. 1–39.

[19] Ilaria Tiddi. “Explaining Data Patterns using Knowledge from the Web of Data”. PhD thesis. Open University, Milton Keynes, UK, 2016.