Klint: Assisting Integration of Heterogeneous Knowledge*

Jacobo Rouces
Aalborg University, Denmark
jrg@es.aau.dk

Gerard de Melo
Tsinghua University, China
gdm@demelo.org

Katja Hose
Aalborg University, Denmark
khose@cs.aau.dk

Abstract

An increasing number of structured knowledge bases have become available on the Web, enabling many new forms of analyses and applications. However, the fact that the data is being published by different parties with different vocabularies and ontologies means that there is a high level of heterogeneity and no common schema. This paper presents Klint, a web-based system that automatically creates mappings to transform knowledge as provided by the sources into data that conforms to a large unified schema. The user can review and edit the mappings with a streamlined interface. In this way, Klint allows for human-level accuracy with minimum human effort.

1 Introduction

The Web of Data includes a rich and increasing amount of structured knowledge bases and has enabled many new applications and forms of analyses. Such knowledge bases usually offer their data in a format based on subject-predicate-object triples, such as RDF. Yet, knowledge bases model information in different ways, so querying them jointly is a daunting task. The reason is that, in order to capture all relevant knowledge, a structured query will have to consist of a disjunction of all possible semantic patterns occurring in the myriad of heterogeneous vocabularies used in the data.

Automatic data integration would solve this challenge, but is often an AI-hard problem, especially since many applications require knowledge with a precision of 90% or higher. Moreover, existing work in this area has mostly focused on connecting entries via binary properties such as `owl:sameAs`. However, such properties can only connect individual identifiers but not easily capture mappings between more complex patterns of triples that represent the same information but are structurally different [Rouces *et al.*, 2015].

In this paper, we present Klint (Knowledge integrator), a Web-based system enabling semi-automatic schema integration. Given one or more existing RDF ontologies, Klint generates tentative integration rules for these ontologies representing mappings into a unified schema. For this unified schema, Klint relies on FrameBase [Rouces *et al.*, 2015], a wide-coverage, highly expressive and extensible schema that can be used to represent and integrate [Rouces *et al.*, 2016] a wide range of knowledge from many sources in a homogeneous and seamless way.[1] Simultaneously, Klint offers an agile and simple interface that enables the user to inspect and adapt the tentative integration rules, achieving the desired balance between precision and scalability.

2 Assisted Schema Integration

Klint allows a user to integrate one or more entire knowledge bases (KBs) into FrameBase with minimum effort. Input KBs can be loaded from an RDF file or a SPARQL endpoint. Other structured data formats can also be used after pre-processing with a suitable RDF converter[2].

Integration Heuristics. Klint automatically creates complex integration rules for each element in the source schema, using integration algorithms based on linguistic annotations in FrameBase [Rouces *et al.*, 2016], extended with support vector machine learning from a labeled training set.

Interface. Each integration rule is represented as a graph in the right pane of Klint's graphical interface (Figure 1). Users can navigate across different integration rules with the buttons at the top bar, making modifications in a given graph if necessary. **Variable nodes** (presented in red) represent universally quantified variables over entities. They bind the pattern from the source KB to the integrated FrameBase pattern. The remaining nodes are classified according to the entity type they represent.

- **Source nodes** (green) represent resources from the source KB and connect two variable nodes.
- **FrameBase nodes** (blue) represent FrameBase resources and also connect variable nodes. They provide

*The research leading to these results has received funding from the European Union Seventh Framework Programme under grant agreement No. FP7-SEC-2012-312651. Additional funding was received from National Basic Research Program of China Grants 2011CBA00300, 2011CBA00301, NSFC Grants 61033001, 61361136003, 61550110504, as well as from the Danish Council for Independent Research (DFF) under grant agreement No. DFF-4093-00301.

[1] See http://www.framebase.org/.

[2] http://www.w3.org/wiki/ConverterToRdf

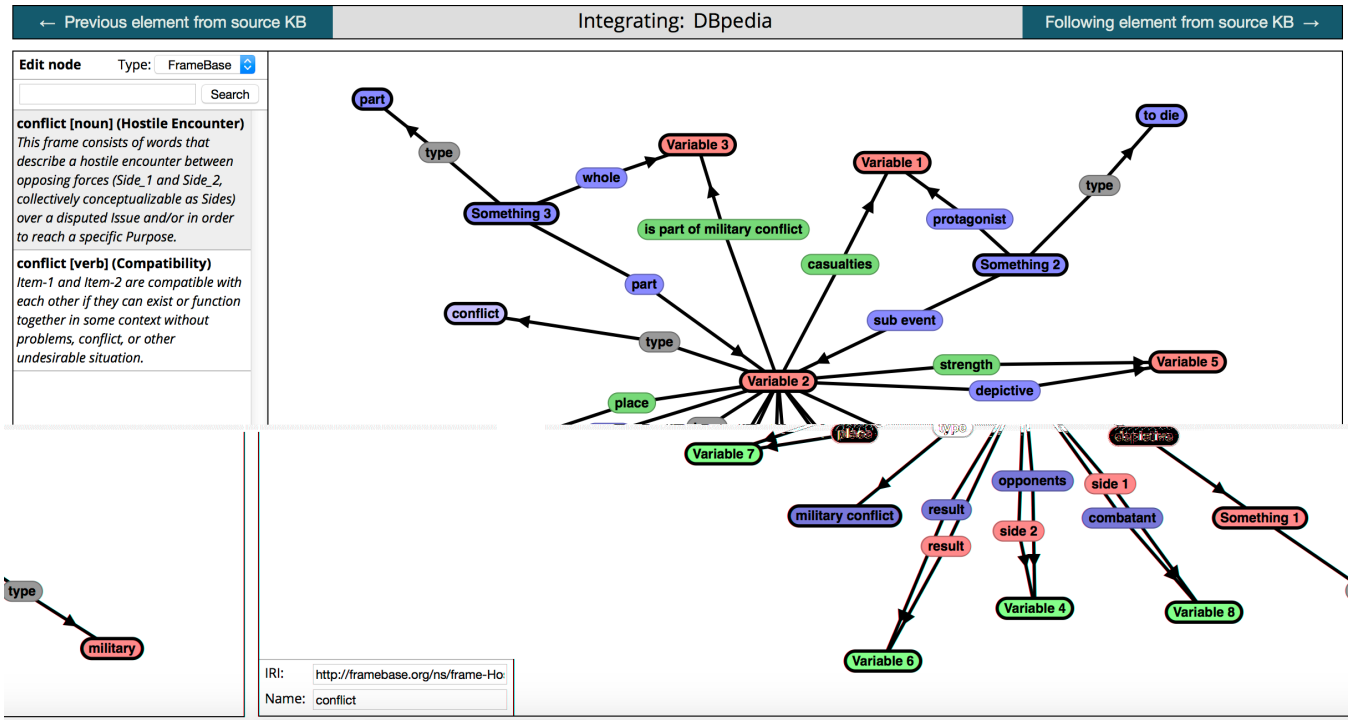

Figure 1: Example of Klint's interface: integrating elements from DBpedia – Klint used the contextual and lexical information from the source elements to suggest two candidate values for the integrated type (selected node, "conflict"), for which the correct assigned value, `Hostile_encounter-conflict.n` was the first suggestion. The FrameBase properties were auto-inserted and some with high lexical overlap were automatically integrated as well. The complex structures that invoke some additional frames were created using the direct search function.

the *translation* of the source pattern to FrameBase.

- **Auxiliary nodes** (gray) represent resources from third-party KBs, usually representing common idioms or very specific entities.

The nodes are connected via directed edges representing triples. Since an RDF triple involves three resources, each triple is represented by two successive edges, one from the subject to the predicate and another from the predicate to the object.

Both edges and nodes can be added, deleted, and edited. When a node is selected, the node is highlighted and the left panel is activated, where the user can change its name and unique identifier (in RDF, this is an Internationalized Resource Identifier).

Automatic Suggestions. When selecting a FrameBase node from an automatically created integration rule, Klint provides the user with an ordered list of alternative suggestions for its value in the left pane. If users still do not find an appropriate choice in the list, they can use the search box to conduct a custom search. This search re-uses the algorithm of the integration engine [Rouces *et al.*, 2016], but allows free input.

When a new FrameBase node is created, its identifier is originally unspecified. In this case, custom search can also be used, but if the node is connected to others, Klint will also try to use the integration engine to suggest possible values automatically.

When an element from the candidate list on the left is chosen, and this element is a class, then associated FrameBase predicates are added as well, connected via subject-predicate edges. Users can complete the ones they consider relevant by connecting the predicate nodes with object nodes via new edges.

3 Conclusion

In this paper, we have presented Klint, a web-based framework that allows users to supervise the automatic integration of heterogeneous knowledge bases by providing a user-friendly graph-based interface that enables reviewing and curating complex integration rules produced by state-of-the-art integration algorithms.

References

[Rouces *et al.*, 2015] Jacobo Rouces, Gerard de Melo, and Katja Hose. FrameBase: Representing N-ary Relations using Semantic Frames. In *ESWC'15*, 2015.

[Rouces *et al.*, 2016] Jacobo Rouces, Gerard de Melo, and Katja Hose. Complex Schema Mapping and Linking Data: Beyond Binary Predicates. In *LDOW'16*, 2016.