# Addressing Knowledge Integration with a Frame-driven Approach

Luigi Asprino

DISI, University of Bologna
luigi.asprino@unibo.it

**Abstract.** Given a knowledge-based system running virtually forever able to acquire and automatically store new open-domain knowledge, one of the challenges is to evolve by continuously integrating new knowledge. This needs to be done while handling conflicts, redundancies and linking existing knowledge to the incoming one. We refer to this task with the name *Knowledge integration*. In this paper we define the problem by discussing its challenges, we propose an approach for tackling the problem, and, we suggest a methodology for the evaluation of results.

## 1 Introduction

Let us imagine a knowledge-based system running virtually forever and able to acquire and automatically store new open-domain knowledge. An example of this kind of systems could be a robot equipped with dialoguing capabilities that enriches its KB with information acquired from interactions with humans and environmental sensors. The actions that the robot chooses to perform might be influenced by the current state of its knowledge base, e.g. an assistive robot learns the preferences of its user by talking to her/him and consequently it performs the actions s/he prefers. The topic of a human-robot dialogue may spread from the user's personal memories to news. In order to interact with a user, a robot's KB must contain the information relevant for the discourse. A possible solution might be to equip the robot with general-purpose background knowledge such as DBpedia or other commonsense KBs. However, despite constantly growing, crowd-sourced KBs are not tailored for representing personal information. Furthermore, considering the need of dealing with potentially any domain, it is unrealistic to pre-define the necessary knowledge. A robot needs a knowledge manager able to evolve the robot's KB with the inputs acquired during its lifetime.

The main challenge in this scenario is to automatically evolve the KB by continuously integrating new open-domain knowledge while handling conflicts, redundancies and linking the existing knowledge with the incoming one. The term knowledge embodies information at both the intensional (e.g. a TBox) and the extensional level (e.g. an Abox). Therefore, the knowledge base should evolve both by including new assertions and by evolving the conceptualization. The problem can be stated as follows.

*Given two (or more) information sources (either structured or unstructured), knowledge integration is the problem of automatically building a knowledge base by extracting the entities from the sources and finding the relations that hold among them.*

The relation has to state: (i) The type of relationship between the entities (e.g. equivalence, subsumption, a domain-specific relation etc.); (ii) When it emerged; (iii) The context where it emerged, e.g. the provenance of the knowledge graphs.

In order to introduce the challenges of the thesis, a preliminary clarification on the terminology is needed. We distinguish a *world entity*, i.e. anything (real, possible or imaginary) in the real world, from an *entity*, i.e. concepts, relations and individuals represented in a knowledge base. Knowledge Integration arises a lot of challenging issues: How different information sources represent the same world entity? Is it possible to automatically find correspondences among same[1] entities coming from different sources? If two entities in the sources represent two different world entities, which is the relationship between them? If two entities contradict each other, is it possible to detect and possibly solve the conflict?

Integrating information from different sources is crucial in today's real-world applications, consequently several communities (e.g. working on databases, web semantics, linguistics etc.) have faced this problem. It has been widely studied from several points of view and many interesting solutions have been proposed. Data Integration, Ontology Matching, Ontology Evolution, Knowledge Fusion, Entity Linking, Co-Reference Resolution, Word Sense Disambiguation are all problems focusing on a particular kind of integration. None of these provides individually a comprehensive solution for the problem. However, it could be addressed by extending and orchestrating techniques developed for specific tasks.

Regardless the way of representing information, that can be delivered either in structured (e.g. relational database) or in unstructured (e.g. plain text) format, human beings express information by reflecting the conceptualization they have in mind. A conceptualization of a stereotyped situation is called *Frame* [4]. Frames are data-structures that can represents facts, like participating to a marriage or being in a certain location. Frames have been mainly being used in linguistics [2] for building lexical databases (e.g. FrameNet); in Natural Language Processing for tasks such as Knowledge Extraction and Semantic Role Labeling; and for Exploration of Encyclopedic Information.

The evidences of frames that emerge from all kinds of knowledge sources suggest to use this cognitive model as a driver for knowledge integration. In my doctoral work I am inquiring the possibility of using this "*remembered framework*" (as defined by Minsky [4]) in the knowledge integration task. In this vision frames constitutes the background knowledge that will serve to ground and merge the incoming knowledge from the various sources.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 presents the approach investigated in this PhD thesis work. Finally, section 4 provides concluding remarks and the research plan.

## 2 Related work

The closest problem to knowledge integration is *Data Integration (DI)*. From a theoretical perspective [3], data integration can be seen as the problem of combining the data

[1] "*Same*" means "that refers to same world entity"

residing at different sources, and providing the user with a unified view (also called *global schema*) of these data. The main task of a data integration system is providing the semantic mapping between the sources and the global schema. Pure data integration solutions cannot be employed for many reasons. In DI, the global schema is modeled on top of the sources instead of automatically emerging from them. Furthermore, global schema and semantic mapping are provided at the scratch line and do not automatically evolve over the time. Finally, despite some semi-automatic approach to schema matching, developing the mapping is still manual work.

*Ontology Matching (OM)* is the problem of finding correspondences between semantically related entities of ontologies. It provides a solution to the semantic heterogeneity problem in order to allow the semantic interoperability of the data expressed with respect to the matched ontologies. Shvaiko and Euzenat [6] surveyed the state of the art and proposed the future challenges for this field. Most of the automatic matching techniques rely on shallow text and structural similarity of the entities of the ontologies. These techniques provide an excellent starting point and could be extended by introducing a deeper analysis of the entity semantics. However, matching is not enough, to create an integrated KB it is also needed to find other kinds of relations that link the existing knowledge with the new one.

*Ontology Evolution (OE)* aims at maintaining an ontology up to date with respect to changes in the domain that it models. A recent survey by Zablith *et al.* [7] provides a complete overview of the tasks involved in Ontology Evolution process. Knowledge integration is related to the first three stages of the OE process, i.e.: (i) *detecting the need of evolution*; (ii) *suggesting ontology changes*; (iii) *validating ontology changes*.

Recently, Mongioví *et al.* [5] proposed a novel solution to semantic reconciliation. They reduced the problem to a graph alignment problem. The main contribution of this work is introducing the notion of "global optimization" in the entity matching task. In doing so this approach takes advantage of considering the semantics of the entities within the context of the entire knowledge base (i.e. the entities are no longer compared individually). However, the limit of the approach is that it is highly tailored to knowledge graphs extracted by means of FRED.

## 3 Proposed approach

The high-level approach that is investigated in this PhD thesis work is summarized in the following two steps. (i) *Entity expansion.* In order to perform a semantic comparison of two entities there is the need of representing their meaning by resorting to a same background knowledge. A background knowledge (such as WordNet, FrameNet or Framester) provides a set of individuals (i.e. concepts, properties etc.) with a well-defined semantics which can be used to describe the semantics of the incoming entities. In general, the semantics of an entity can be represented by a semantic network that relates individuals of the background knowledge. Each source's entity is "replaced" by a richer graph representing its meaning in terms of the background knowledge. (ii) *Comparison*. A graph-based comparison of the semantic networks representing the meaning of the entities is performed. The aim of this task is to detect the relationship between two semantic networks representing two entities coming from the information sources.

This methodology has been implemented for addressing the problem of semantic heterogeneity among ontologies [1]. Most of the current ontology matching solutions present two main limits: (i) they only partially exploit the natural language descriptions of ontology entities and lexical resources as background knowledge; (ii) they are mostly unable to find correspondences between entities specified through different logical types (e.g. mapping properties to classes). We introduced a novel approach aimed at finding complex correspondences between ontology entities according to the intensional meaning of their models, hence abstracting from their logical types.

In order to deduce the intensional meaning of ontology entities we proposed to analyze the natural language annotations associated with them. In fact, annotations provide humans with insights of the intensional meaning the designer wants to represent with a certain entity. The main idea of this approach is that words used in annotations *evoke* frames that are representative of the intensional meaning of the entity. Evoked frames can be used to describe the semantics of the ontology entity. In other words, the *expansion step* of the above methodology has been implemented by using frames as background knowledge to describe entities' semantics. The frame-based representation of the meaning of ontology entities allows us to treat ontology entities as *multigrade predicates* hence abstracting from their logical type.

The proposed approach is currently being evaluated. We are evaluating the resulting alignments in a both *direct* and *indirect* way. The benchmarks used for assessing ontology matching systems[2] are not able to evaluate the capability of finding complex correspondences among ontology entities with different logical types. In order to accomplish this purpose we are extending the existing benchmarks for ontology matching. On the other hand, we are using the proposed approach in a question answering system for selecting relevant resources answering a given question. The frame occurrences in a question together with the frame-ontology alignment help in formulate the query over the linked data, hence identifying resources that answer the given question.

***Evaluation of results.*** We found only one attempt for developing a gold standard for benchmarking the performance of a semantic reconciliation framework. Mongioví *et al.* [5] proposed to develop a ground truth for semantic reconciliation by adapting EECB 1.0, the gold standard for Co-Reference Resolution. The focus of Mongioví *et al.* was on the matching task, but we also need to test the capability of the framework of instancing new semantic relationship and of solving inconsistencies. Therefore the ground truth can be extended (or a new one can be created) by exploiting crowdsourcing techniques, e.g. by designing a "game with a purpose".

The effectiveness of the methodologies and techniques will be further assessed by employing them in a real scenario. For instance, the results of the thesis could be used for managing the knowledge base of an intelligent agent (e.g. an assistive robot, a question answering system etc.) that decides the actions to perform on the basis of the current state of its KB. In this case, the quality of the of the agent behavior can be considered as representative of the effectiveness of the thesis results.

[2] OAEI, http://oaei.ontologymatching.org/

## 4 Research plan and conclusion

This PhD work is dimensioned for three years. The beginning of the first year has been devoted to an intense literature review of the related research areas and to selecting the promising methodologies that could contribute to the development of a solution. The first year has been also devoted to (i) devising the high-level approach for addressing knowledge integration; (ii) and to implementing a novel solution for ontology matching that uses frames as background knowledge for reducing semantic heterogeneity among entities of different ontologies. This solution tackles knowledge integration at *schema level* but in the next future we plan to extend the approach to face the integration at *extensional level*. We hypothesize that the semantics of an individuals can be described through a frame-based specification thus reducing the semantic heterogeneity among individuals of different knowledge graphs.

In the second year we plan to evaluate the approach presented in Section 3 (i) by extending the existing benchmarks for ontology matching; (ii) by employing the frame-based ontology alignment in a question answering system as support for creating queries from the input questions.

This paper presented a summary of the PhD thesis in its early stage. We presented the problem of knowledge integration, its challenges and an approach that will be investigated during the next years. Finally, we suggested some methodologies for assessing the effectiveness of a knowledge integration framework.

**Acknowledgements.** I wish to express my gratitude to my advisors, Prof. Paolo Ciancarini and Dr. Valentina Presutti, for their encouragement throughout my research for this work. I would also like to thank Professor Aldo Gangemi for introducing me to the wonders of the Frame Semantics and for his precious suggestions.

## References

[1] Luigi Asprino, Valentina Presutti, and Aldo Gangemi. “Matching Ontologies using a Frame-driven Approach”. In: *Proc of EKAW - Satellite Events*. (Bologna, Italy). 2016, (to appear).

[2] Charles J. Fillmore. “Frame semantics”. In: *Linguistics in the Morning Calm*. Hanshin Publishing Co., 1982, pages 111–137.

[3] Maurizio Lenzerini. “Data Integration: A Theoretical Perspective”. In: *Proc of PODS 2002*. (Madison, Wisconsin). ACM, 2002, pages 233–246.

[4] Marvin Minsky. *A Framework for Representing Knowledge*. Technical report. Cambridge, MA, USA, 1974. URL: http://hdl.handle.net/1721.1/6089.

[5] Misael Mongiovì, Diego Reforgiato Recupero, Aldo Gangemi, Valentina Presutti, and Sergio Consoli. “Merging open knowledge extracted from text with MERGILO”. In: *Knowledge-Based Systems* 108 (), pages 155–167. DOI: 10.1016/j.knosys.2016.05.014.

[6] Pavel Shvaiko and Jerome Euzenat. “Ontology Matching: State of the Art and Future Challenges”. In: *IEEE Transactions on Knowledge and Data Engineering* 25.1 (2013), pages 158–176.

[7] Fouad Zablith, Grigoris Antoniou, Mathieu d’Aquin, Giorgos Flouris, Haridimos Kondylakis, Enrico Motta, Dimitris Plexousakis, and Marta Sabou. “Ontology evolution: a process-centric survey”. In: *The Knowledge Engineering Review* 30.1 (2015), pages 45–75.