An Ontological Approach to Knowledge Building by Data Integration

Salvatore Flavio Pileggi[1(✉)], Hayden Crain[1], and Sadok Ben Yahia[2]

[1] School of Information, Systems and Modelling, University of Technology Sydney, Ultimo, Australia
SalvatoreFlavio.Pileggi@uts.edu.au, Hayden.J.Crain@student.uts.edu.au
[2] Department of Software Science, Tallinn University of Technology, Tallinn, Estonia
sadok.ben@taltech.ee

Abstract. This paper discusses the uncertainty in the automation of knowledge building from heterogeneous raw datasets. Ontologies play a critical role in such a process by providing a well consolidated support to link and semantically integrate datasets via interoperability, as well as semantic enrichment and annotations. By adopting Semantic Web technology, the resulting ecosystem is fully machine consumable. However, while the manual alignment of concepts from different vocabularies is reasonable at a small scale, fully automatic mechanisms are required once the target system scales up, leading to a significant uncertainty.

Keywords: Ontology · Data integration · Semantic interoperability · Semantic Web · Uncertainty · Data engineering · Knowledge engineering

1 Introduction

Data integration, defined as *"the problem of combining data residing at different sources, and providing the user with a unified view of these data"* [21], can be considered a classic research field as could witness the myriad of contributions in literature. Its relevance is determined by the practical implications in the different applications domains.

In this respect, we rely on an ontological approach to support the data integration process. The benefits of ontology in the different application domains are well-known and have been extensively discussed from different perspectives in several contributions. The knowledge building process, as understood in this paper, is not limited to data integration but it also includes semantic enrichment and annotations. By adopting Semantic Web technology, the resulting ecosystem is fully machine consumable. However, while the manual alignment of concepts from different vocabularies is reasonable at a small scale, fully automatic mechanisms are required once the target system scales up, leading to a significant uncertainty.

V. V. Krzhizhanovskaya et al. (Eds.): ICCS 2020, LNCS 12143, pp. 479–493, 2020.
https://doi.org/10.1007/978-3-030-50436-6_35

This paper provides two key contributions:

- the manual knowledge building process is described and implemented by a tool which systematically supports data integration and semantic enrichment.
- the uncertainty introduced by the automation of the process is discussed.

2 Related Work

The scrutiny of data integration adopting an ontological approach sheds light on some key issues, to wit data semantics and uncertainty representation.

- **Ontological approach to Data Integration**: The role of semantic technology in data integration [21] has been deeply explored during the past years. The contributions currently in literature clearly show that semantic technology provides a solid support in terms of data integration and reuse via interoperability [26]. For instance, [6] proposes an ontological approach to federated databases; ontology-based integration strategies have been proposed to a range of real scientific and business issues [15], such as the integration of biomedical [33] and cancer [39] data, and the integration among systems [25]. Last but not least, ontologies are contributing significantly to an effective approach to the integration of Web Resources (normally in XML [2]) and to linked open data [16]. Ontology may be adopted to support different strategies and techniques [38] and result very effective in presence of heterogeneity [12]. For instance, central data integration assumes a global schema to provide access to information [15], while in peer-to-peer data integration there is no global point of control [15].
- **Data Semantics**: Associating formal semantics to data is a well-known problem in the fields of artificial intelligence and database management. Again, ontological structures play a key role [31] and they normally support an effective formalization of the semantics, which becomes a key asset in the context of different applications, for instance to interchange information [1,27] or to improve data quality [22]. In general, the importance of data semantics to support interoperability is gaining more and more attention within different communities, for example within the geo-spatial information [19] and within the medical community [4,20]. Moreover, the analysis of semantic data may support sophisticated data mining techniques [8,10].
- **Uncertainty Representation**: Probability theory and fuzzy logic have been used to represent uncertainty in data integration works [23]. Uncertainty management works also include possibilistic and probabilistic approaches [14]. A probabilistic approach towards ontology matching was utilized in several works, where machine learning was utilized in estimating correspondence similarity measures [14]. To refine the matcher uncertainty and improve the precision of its alignment, Gal [13] proposed a method to compute top-K alignments instead of computing a *best* single alignment, and proposed a heuristic to simultaneously compare/analyze/examine the generated top-K alignments and choose one good alignment among them. The *best* alignment is an alignment that optimizes a target function F between the two schemata.

Typical ontology matching methods commit to the *best* alignment which maximizes the sum (or average) of similarity degrees of pairwise correspondences. To model the ontology matching uncertainty, Marie and Gal [24] proposed to use *similarity matrices* as a measure of certainty. They aim at providing an answer to the question of whether there are *good* and *bad* matchers.

To represent the inherent uncertainty of the automatic schema matching, Magnani and Montesi [23] used the notion of *probabilistic uncertain semantic relationship* (pUSR), which is a pairwise correspondence defined as a tuple (E_1, E_2, R, P) where E_1 and E_2 are two elements/entities, R is a set of relationship types (equivalence, subsumption, disjointness, overlap, instantiation, *etc.*), and P is a probability distribution over R. The pUSRs form an uncertain alignment.

Dong *et al.* [9] proposed a system that models the uncertainty about the correctness of alignments by assigning each possible alignment a probability. The probabilities associated with all alignments sum up to 1. The authors define a *probabilistic schema mapping* (alignment) as a set of correspondences between a source schema and a target schema, where each uncertain mapping/alignment has an associated probability that reflects the likelihood that it is correct.

Po and Sorrentino [30] quantify uncertainties as probabilities. They define the notion of *probabilistic relationship* as a couple $(\langle t_i, t_j, R\rangle, P)$ where $\langle t_i, t_j, R\rangle$ is a relationship between t_i and t_j of the type R, and P is the probability (confidence) value (in the normalized interval [0–1]) associated to this relationship. Within the range [0–1], they can distinguish between *strong* relationships and *uncertain* relationships (*i.e.*, relationships with a low probability value). Uncertain relationships could be seen as candidate relationships that need further confirmation by a human expert.

There are several pairs of entities in different ontologies that are related to each other but not necessarily with one of the typical well-defined relationships. However, these correspondences vary in their degree of relatedness. This information is difficult to formalize. Therefore, Zhang *et al.* [40] proposed a new type of relation called *Relevance*. The latter represents relationships between entities that are not covered by a strict relation such as equivalence, subsumption or disjointness, *etc.* In this context, we think that the *relevance* relation is very similar to the *overlap* relation. The authors also presented the notion of *fuzzy ontology alignment*, that uses fuzzy set theory to model the inherent uncertainty in the alignment correspondences.

3 Knowledge Building by Data Integration

The knowledge building process is ideally composed of two sequential steps that we refer to as *physical* and *logical integration*:

- **Physical Integration: The Virtual Table Model.** As the name suggests, the physical integration aims to convert data in an interoperable format that ultimately defines the target data space. By adopting Semantic Web technology, physical integration is required only if the target dataset is not already available in a semantic format (e.g. RDF or OWL). The Virtual Table model (Fig. 1) is a simple and intuitive approach to data integration that assumes the target dataset described as one or more tables according to the classical relational model. An external dataset may be mapped into a virtual table and automatically converted in OWL. Data may be automatically retrieved from a relational table [29] or inserted manually by users through the copy&paste functions provided by the user interface as in the tool described later on in the paper.

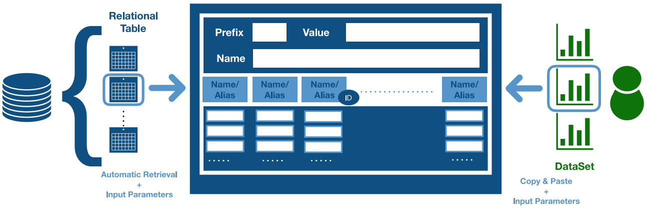


Fig. 1. Virtual Table model.

- **Logical Integration: Semantic Alignment, Internal and External Linking.** Logical integration assumes a given data set already imported within the data space and consists in the consolidation and enrichment of data semantics by specifying additional relationships, such as semantic equivalences, internal and external links. Once a data set has been imported within the semantic data space, it may need to be logically linked to other data and semantically enriched. We structure our knowledge building process by including three different kind of semantic enrichment (Fig. 2): *internal linking*, *metadata association* and *external linking*.

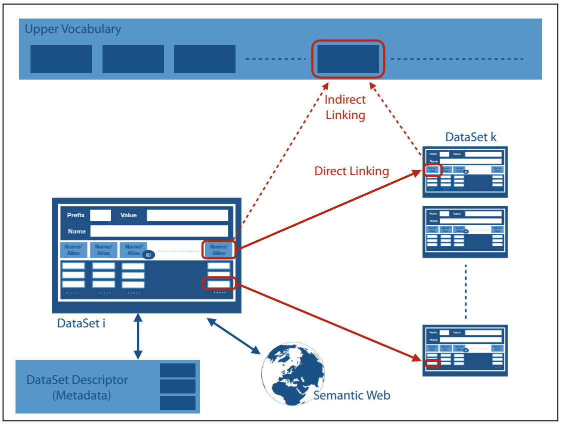
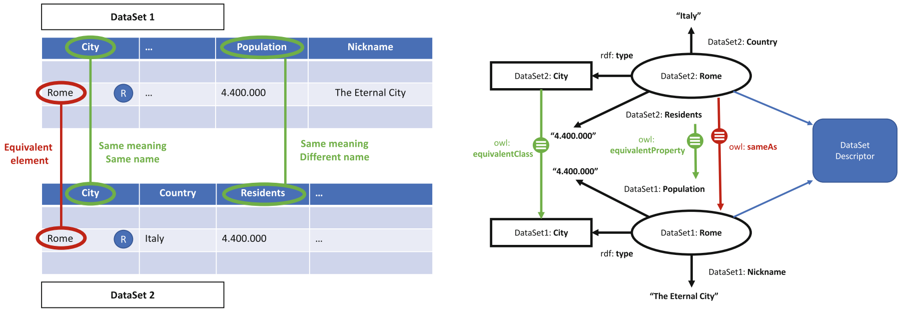


Fig. 2. Semantic linking and enrichment.

Internal linking is an ontology alignment process among the different datasets which are considered part of the data space. That is a key process to enable the effective integration at an user level of heterogeneous datasets. For instance, two attributes belonging to two different datasets may have the same meaning. Semantic technology provides simple and effective mechanisms to establish semantic equivalences among classes, instances, relationships and attributes. As discussed in the following section, these mechanisms may be used in a relatively easy way, if they properly supported by user-friendly interfaces. Semantic correspondences among ontology elements may be established directly or indirectly (Fig. 2). Direct linking, namely semantic equivalences established directly from a dataset to another, is simple from a management perspective but may result not too much effective in complex environments, i.e. within collaborative systems, or, more in general, when the scale of the system in terms of number of linked datasets becomes relevant. On the other side, indirect linking established through upper vocabularies is well-known and consolidated techniques that may result in a much more effective approach. However, it introduces an additional cost from a management perspective. The semantic infrastructure allows generic linking within the semantic space or externally. So a dataset or an element belonging to a dataset may be related with other concepts to define or extend the semantics associated. For example, a given dataset may be related to a number of keywords, to a research project or to a scientific paper by adopting the PERSWADE-CORE vocabulary [28].

A simple example of data integration involving two datasets is represented in Fig. 3. As shown, both target datasets address information related to cities. Figure 3a represents dataset in their original format, while Fig. 3b depicts the integrated space as a knowledge graph. The column city is considered like a *Web Resource* that in this case is also the primary key for both tables. Although the two datasets present some redundancies, they provide, in general, different information about cities. In this case, the integration process will enable the two original datasets within the semantic data space assuring semantic consistency among the different fields and concepts. Indeed, from a semantic perspective, even this simple use case proposes a number of potential issues that have to be addressed in order to guarantee a correct and effective integration. As shown in the figure, there are several semantic equivalences among the two datasets to be represented. They include attributes (columns in the virtual table model) that have the same name and the same meaning within their original context, as well as attributes that have different names but the same meaning. For instance, the attributes “Population” and “Residents” refer to the same concept, namely the number of people currently living in a given city. Additionally, equivalent resources have to be semantically related. In the example, “Rome” appears in both tables. This syntactic equivalence is integrated by a semantic one to properly address the reference to the city of Rome. OWL provides simple mechanisms to define equivalences among concepts.

Overall, a simplification of the scenario previously discussed can be represented by the knowledge graph in Fig. 3b that adopts OWL 2 structures. More concretely, the equivalence among classes is enforced by the OWL rule *OWL:equivalentClass*, as well as the equivalences among properties is specified by *OWL:equivalentProperty*. Similarly, an OWL statement including *OWL:sameAs* applies to instances of classes.



(a) Representation as tables

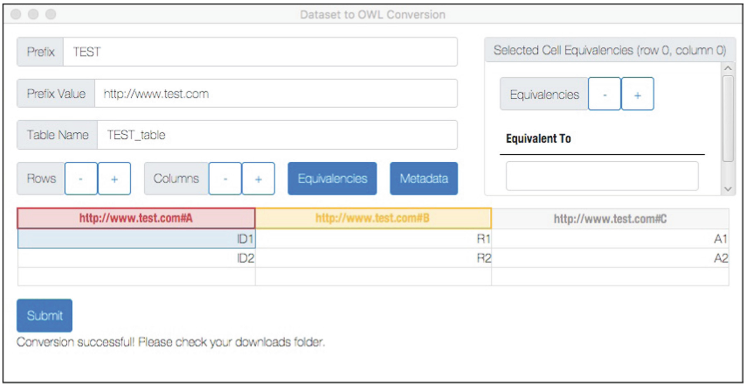
(b) Integrated data space

Fig. 3. An example of integration of two datasets.

3.1 A Tool for Supervised Data Integration

Our implementation supports most part of the knowledge building process as previously presented and discussed. It is based and relies on intuitive user inputs rather than on strong skills in ontology and Semantic Web technology. However, it assumes the understanding of basic concepts, i.e. the difference between an object and an attribute. The primary goal of the tool is to support the systematic conversion of a given dataset into an independent and self-contained ontology in OWL. The user interface (Fig. 4 allows to directly import a relational table. Regardless of the method used to import data (based on copy&paste in this case), the user is asked to characterize the table each column according to one of the following options:

- *ID*. It is normally equivalent to the primary key in the relational model. However, it is assumed to be an unique data field. Therefore, keys composed by multiple fields cannot be directly used and need to be encoded previously.
- *Resource*. By using this option, associated data is considered like an object, namely a Web resource in Semantic Web technology. A Web resource has an unique identifier and can be further characterized.
- *Attribute*. It's a normal data field, e.g. a text or a number.



Prefix: TEST

Prefix Value: http://www.test.com

Table Name: TEST_table

Selected Cell Equivalencies (row 0, column 0)

Equivalent To

http://www.test.com#A	http://www.test.com#B	http://www.test.com#C
ID1	R1	A1
ID2	R2	A2

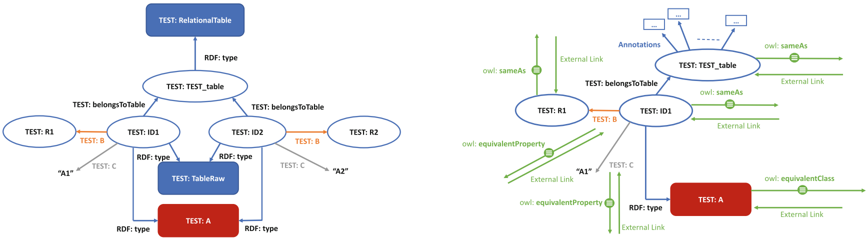
Conversion successful! Please check your downloads folder.

Fig. 4. User interface.

Through the provided interface, users may specifies meta data for the imported table, such as *source*. *license*, *description* and *publisher*. Last but not least, relatively friendly alienation among concepts is supported.

3.2 Semantic Representation

The output of the example proposed in Fig. 4 is represented as a knowledge graph in Fig. 5a. As shown, the IDs (as previously defined) is associated with a new class (*A* in this case) and the instances of ID (*ID1* and *ID2*) are also stated as member of the internal class *TableRaw*. This last concept identifies rows in the virtual table *TEST_table*, which is stated as a member of the class *RelationalTable*. Resources (*B* in the example) are converted in OWL Object properties, while attributes (*C* in the example) are converted in OWL data properties. The resulting schema may be semantically enriched trough concept alignment and external linking (Fig. 5b).



(a) Semantic representation (b) Semantic enrichment

Fig. 5. Semantic representation of the integrated dataspace.

4 Uncertainty in Non-supervised Ontology Alignment

Uncertain schema correspondence is often generated by (semi-)automatic unsupervised tools and not verified by domain experts. Even in manual or semi-automatic tools, the users may not understand enough the domain and, thus, provide incorrect or imprecise correspondences. In some domains, it is not clear what the correspondences should be [9]. Schema elements (entities) can be ambiguous in their semantics [30] because entities are close (*i.e.*, related to each other) but neither synonyms (*i.e.*, completely similar) nor dissimilar (*i.e.*, completely different) [5,17]. Therefore, matching systems turn out to be uncertain, since it is not accurate to declare whether two entities are equivalent or not [40]. In the ontology domain, ontological entities do not always correspond to single physical entities, they rather share a certain amount of mutual information [40]. Indeed, real-world ontologies generally have linguistic, structural and semantic ambiguities, resulting from their heterogeneous domain conceptualizations [3]. Eventually, ambiguity and heterogeneity in ontology models/representations are carried in the process of matching and integrating ontologies [3]. Finally, *Uncertain query* is commonly associated with multiple structured queries generated by the system as candidate queries reflecting uncertainty about which is the real intent of the user.

Klir and Yuan [18] defined two basic types of uncertainty: (*i*) *Fuzziness* which is the lack of definite or sharp distinctions; and (*ii*) *Ambiguity* which is the existence of one-to-many correspondences that may introduce a disagreement in choosing among several correspondences.

There are two choices to remove (or at least reduce) the alignment uncertainty in schema matching processes: either with the support of a user (manually) or by using a threshold. According to the former approach, *aka user feedback*, users can manually select matching and non-matching correspondences from the alignment, *i.e.* in semi-automatic matching process when the system requests help [7]. The latter approach is based on a *threshold* that can be established in a semi-automatic manner (*i.e.* using user feedback cycles) or in an automatic manner (i.e. using learning approaches) in order to minimize the introduction of false correspondences. A matcher filters/discards correspondences having a confidence value that does not reach a certain threshold, assuming that correspondences with low confidence/similarity measures are less adequate than those with high similarity measures. However, separating correct from incorrect correspondences in an alignment is a hard task [14]. To find the optimal/best threshold, many trials should be made by varying/tuning the confidence value threshold [30]. In addition, different thresholds can be assigned to different applications. For example, a recommendation system may have relatively low thresholds since false positives are tolerated, while a scientific application may have high thresholds [40]. As a rule of thumb, the information loss, caused by the removal of uncertainty, leads to a worsening of the alignment quality [30]. In fact, any selection of a threshold often yields false negatives and/or false positives. Therefore, the exact alignment cannot be found by setting a threshold [13].

Generally speaking, the uncertainty generated during the matching process is lost or transformed into exact one (defuzzification) [23]. Therefore, there is a concrete need to incorporate uncertain/inaccurate correspondences and handle uncertainty in alignments [14], due to the inherent risk of losing relevant information [23].

Uncertainty Management. From a literature review, we have identified two different levels to deal with uncertainty management in schema matching: some solutions try to quantify the uncertainty of an entire alignment when there are many alignments produced for the same matching case; others try to represent and quantify the uncertainty of the correspondences of a given alignment.

Management of Ontology Alignment Uncertainty. A semantic alignment (*aka mapping*), denoted as $\mathcal{A} = \{c_1, c_2, \ldots, c_n\}$, is a set of semantic correspondences between two or more matched ontologies. It is the result/output of the ontology matching process.

The uncertainty of a matcher should be explicitly reflected in an uncertainty measurement in order to be able to choose good enough alignments [32]. The work in [9] introduced the notion of *probabilistic schema alignments*, namely a set of alignments with a probability attached to each alignment. The purpose of defining *probabilistic* alignments is to answer queries with uncertainty about (semi-)automatically created alignments [32].

Management of Correspondence Uncertainty. In general, given two matched ontologies $\mathcal{O}_1$ and $\mathcal{O}_2$, a semantic correspondence (*aka* a *relation* or a *relationship*) is a 4-tuple $< e_{\mathcal{O}_1}, e_{\mathcal{O}_2}, r, n >$ where $e_{\mathcal{O}_1}$ is an entity belonging to $\mathcal{O}_1$, and $e_{\mathcal{O}_2}$ is an entity belonging to $\mathcal{O}_2$, r is a semantic relation holding (or intended to hold) between $e_{\mathcal{O}_1}$ and $e_{\mathcal{O}_2}$, such as equivalence $(\equiv)$, subsumption $(\sqsubseteq/\sqsupseteq)$, disjointness $(\perp)$, or overlap $(\between)$, and n is a confidence value/measure/probability assigning a degree of trust/reliability/correctness on the identified relation and ranging typically between $[0, 1]$, where 0 represents no similarity and 1 represents full similarity. In the equivalence case, n indicates whether both entities have a high or low similarity measure/degree. The higher the confidence degree, the more likely the relation holds [11]. A matcher would be inclined to put a similarity value of 0 for each entity pair it conceives not to match, and a value higher than 0 (and probably closer to 1) for those correspondences that are conceived to be correct [13]. On the other hand, in the *crisp* correspondences (composing the *crisp* alignments), the confidence values of all correspondences are equal to 1.

Correspondences Generated by a Matcher Aggregation. Some matchers assume that similar entities are more likely to have similar names. Other matchers assume similar entities share similar domains. Other matchers assume that similar entities are more likely to have similar neighbors (parents, children, and siblings). And others assume that similar entities are more likely to have similar instances [24]. In order to combine principles by which different schema matchers judge the similarity between entities, the combined matcher aggregates the outcome (*i.e.*, the output alignment) of all matchers to produce a single alignment. It automatically computes the overall similarities of correspondences by aggregating the similarity degrees assigned by individual matchers. In the automatic process of matching, it is proven that an ensemble (a combination) of complementary matchers (*e.g.*, string-based, linguistic-based, instance-based, and structural matchers, *etc.*) outperforms the behavior of individual matchers [30] since they compensate for the weaknesses of each other [24]. In recent years, many matching tools use schema matcher ensembles to produce better results. Therefore, the similarity measure of a correspondence is generally the result of aggregating multiple similarity measures [7], and as the number of such similarity measures increase, it becomes increasingly complex to aggregate the results of the individual measures. The generated similarity degrees of correspondences are dependent to the choice of the weights of individual matchers assigned by aggregation algorithms for the similarity combination [40]. Therefore, a similarity degree of a given correspondence represents the "belief" of a matcher in the correctness of that correspondence [13]. However, the real issue in any system that manages uncertainty is whether we have reliable probabilities (degrees of similarity), because unreliable probabilities can lead us to choose erroneous or not good enough correspondences. Obtaining reliable probabilities for uncertainty management systems is one of the most interesting areas for future research [9]. Finally, disregarding semantic similarity degrees of the alignment correspondences may impede the overall integration process [5].

Correspondence Ambiguity. An ambiguous alignment [11] is a one-to-many $(1 : n)$, a many-to-one $(n : 1)$, or a many-to-many $(n : n)$ alignment. This means that it contains some ambiguous correspondences [11] (*i.e.*, that match the same entity from one ontology with more than one entity from the other ontology). An ambiguous correspondence is a correspondence in which at least one entity is also involved in other correspondences. Contrary to one-to-one $(1 : 1)$ alignments in which an entity appears in at most one correspondence.

The ambiguous correspondences are generally a source of uncertainty because they can be interpreted in two ways: (*i*) A first point of view considers that only a single ambiguous *equivalence* correspondence (probably the one that has the highest confidence value) truly reflects a synonym/alternative entity, while the remaining ones (having lower confidence values) rather reflect similar, related or overlapping terms, not strictly denoting equivalent entities [37]; (*ii*) A second point of view considers the ambiguous *equivalence* correspondences as actually *subsumption* correspondences, because an entity in one ontology can be decomposed into several entities in another ontology [13]. This happens in case where one ontology is more granular (or general) than the other one [37].

Correspondences in Coherent and Conservative Alignments: Consistency Principle. The *consistency principle* [36] states that the integrated ontology –resulting from the integration of the input ontologies– should be coherent (*i.e.*, all entities of the integrated ontology should be satisfiable), assuming that the input ontologies are also coherent (*i.e.*, the input ontologies also do not contain any unsatisfiable entities). An *unsatisfiable* entity (class or property) is an entity containing a contradiction in its description, for which it is not possible for any instance to meet all the requirements to be a member of that entity. In some applications where the logical reasoning is involved, ensuring coherence is of utmost importance since the integrated ontology must be logically/semantically correct to be really useful, otherwise it may lead to incomplete or unexpected results.

Conservativity Principle. The conservativity principle [34–36] requires that the original description (especially the *is-a* structure/class hierarchy) of an input ontology should not be altered after being integrated. Hence, the introduction of new semantic relations between entities of each matched ontology is not allowed, especially new subsumption relations causing structural changes. The conservativity principle aims that the use of the new integrated ontology –resulting from the integration of the input ontologies– does not affect the original behavior of the applications already functioning with the input ontologies (that were integrated).

Example 1 (Coherence/Conservativity Violation). Suppose that we have a class A in $\mathcal{O}_1$, two disjoint classes (B and C) in $\mathcal{O}_2$, and two correspondences c_1 and c_2 stating that A is a subclass of B and C. Formally,

$$\mathcal{O}_1 = \{A\} \quad \mathcal{O}_2 = \{B \perp C\}$$
$$\mathcal{A} = \{c_1, c_2\} \quad c_1 = < A \sqsubseteq B > \quad c_2 = < A \sqsubseteq C >$$

If a reasoning process is applied on the integrated ontology $\mathcal{O}_3$, then A will be an unsatisfiable class since it will become a subclass of two disjoint classes.

Now if we consider the following two ontologies: $\mathcal{O}_1$ has two classes A and B, and $\mathcal{O}_2$ has two classes A' and B' where B' is a subclass of A'. Formally,

$$\mathcal{O}_1 = \{A, B\} \quad \mathcal{O}_2 = \{B^{'} \sqsubseteq A^{'}\}$$
$$\mathcal{A} = \{c_1, c_2\} \quad c_1 = < A \equiv A^{'} > \quad c_2 = < B \equiv B^{'} >$$

If the ontology matching generates two correspondences c_1 and c_2 stating that A is equivalent to A', and B is equivalent to B', then the original structure of $\mathcal{O}_1$ will change in the integrated ontology $\mathcal{O}_3$ because of the addition of a new subsumption linking A and B.

Whenever an unsatisfiable entity or a conservativity violation is identified in the integrated ontology, then an alignment repair algorithm first identifies the correspondences causing these problems. The identified correspondences may actually be *erroneous* correspondences, but may also be *correct* correspondences introducing violations because of the incompatible conceptualizations of the matched ontologies. A human expert can then be notified and pointed to manually check and specify his/her opinion on these correspondences, to give his/her contribution to the matching process [23]. Otherwise, the alignment repair system can resolve these violations by automatically removing the identified correspondences and generating a repaired (coherent and conservative) output alignment. In a text annotation application, it is not necessary to ensure the coherence of the integrated ontology. However, in other applications, *e.g.*, query answering, logical errors in the integrated ontology may have a critical impact in the query answering process. Similarly, in some cases, the conservativity principle is no longer required, since the integrated ontology will be used by another specific application, *i.e.* not by the applications already using the ontologies that were integrated. Therefore, there is a need to represent/express correspondences causing (*consistency* and *conservativity*) violations in the forthcoming integrated ontology, and model them in the Alignment format [23]. The *Alignment*[1] format, (*aka* the *RDF Alignment format*), is the most consensual ontology alignment format used for representing simple pairwise alignments. In this format, we can not differentiate between a normal correspondence and a *repaired* one (involved in integration violations and identified by alignment repair systems). Therefore, there is a representation problem in the ontology alignment repair area.

5 Conclusions and Future Work

This paper presented a simple approach for knowledge building from raw datasets by adopting rich data models (ontologies). The tool developed proposes some automatic features to import data, which is mapped on virtual tables. Nevertheless, we need to automate the key mechanism to enforce semantic consistence among the different datasets is supposed. It becomes unrealistic once the scale of the system becomes significant or in presence of heterogeneity. Future work will be oriented to the automation of the whole process by particularizing existing techniques to the specific case of datasets mapped on virtual tables. Furthermore, we will include an additional virtual structure to support multi-dimensional data based on the RDF Data Cube Vocabulary[2].

Acknowledgments. The implementation of the tool described in the paper has been funded by the Faculty of Engineering and IT at the UTS through a Summer Scholarship granted to Hayden Crain. Additionally, we thank Prof. Iwona Miliszewska for encouraging and supporting the research collaboration between the UTS and TalTech.

References

1. Abdul-Ghafour, S., Ghodous, P., Shariat, B., Perna, E.: A common design-features ontology for product data semantics interoperability. In: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, pp. 443–446. IEEE Computer Society (2007)

[1] http://alignapi.gforge.inria.fr/format.html.

[2] The RDF Data Cube Vocabulary, https://www.w3.org/TR/vocab-data-cube/. Accessed: 8/01/ 2019.

2. Amann, B., Beeri, C., Fundulaki, I., Scholl, M.: Ontology-based integration of XML web resources. In: Horrocks, I., Hendler, J. (eds.) ISWC 2002. LNCS, vol. 2342, pp. 117–131. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-48005-6_11
3. Bharambe, U., Durbha, S.S., King, R.L.: Geospatial ontologies matching: an information theoretic approach. In: 2012 IEEE International Geoscience and Remote Sensing Symposium, IGARSS, pp. 2918–2921. IEEE (2012)
4. Bhatt, M., Rahayu, W., Soni, S.P., Wouters, C.: Ontology driven semantic profiling and retrieval in medical information systems. J. Web. Semant. **7**(4), 317–331 (2009)
5. Blasch, E.P., Dorion, É., Valin, P., Bossé, E.: Ontology alignment using relative entropy for semantic uncertainty analysis. In: Proceedings of the IEEE 2010 National Aerospace & Electronics Conference, pp. 140–148. IEEE (2010)
6. Buccella, A., Cechich, A., Rodríguez Brisaboa, N.: An ontology approach to data integration. J. Comput. Sci. Technol. **3**, 62–68 (2003)
7. Cross, V.: Uncertainty in the automation of ontology matching. In: Fourth International Symposium on Uncertainty Modeling and Analysis, 2003. ISUMA 2003, pp. 135–140. IEEE (2003)
8. Delgado, M., SáNchez, D., MartıN-Bautista, M.J., Vila, M.A.: Mining association rules with improved semantics in medical databases. Artif. Intell. Med. **21**(1–3), 241–245 (2001)
9. Dong, X.L., Halevy, A., Yu, C.: Data integration with uncertainty. VLDB J. **18**(2), 469–500 (2009). https://doi.org/10.1007/s00778-008-0119-9
10. Dou, D., Wang, H., Liu, H.: Semantic data mining: a survey of ontology-based approaches. In: Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing, ICSC, pp. 244–251. IEEE (2015)
11. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-38721-0
12. Gagnon, M.: Ontology-based integration of data sources. In: 2007 10th International Conference on Information Fusion, pp. 1–8. IEEE (2007)
13. Gal, A.: Managing uncertainty in schema matching with Top-K schema mappings. In: Spaccapietra, S., Aberer, K., Cudré-Mauroux, P. (eds.) Journal on Data Semantics VI. LNCS, vol. 4090, pp. 90–114. Springer, Heidelberg (2006). https://doi.org/10.1007/11803034_5
14. Gal, A., Shvaiko, P.: Advances in ontology matching. In: Dillon, T.S., Chang, E., Meersman, R., Sycara, K. (eds.) Advances in Web Semantics I. LNCS, vol. 4891, pp. 176–198. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-89784-2_6
15. Gardner, S.P.: Ontologies and semantic data integration. Drug Discov. Today **10**(14), 1001–1007 (2005)
16. Jain, P., Hitzler, P., Sheth, A.P., Verma, K., Yeh, P.Z.: Ontology alignment for linked open data. In: Patel-Schneider, P.F., et al. (eds.) ISWC 2010. LNCS, vol. 6496, pp. 402–417. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-17746-0_26
17. Jan, S., Li, M., Al-Sultany, G., Al-Raweshidy, H.: Ontology alignment using rough sets. In: 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD, vol. 4, pp. 2683–2686. IEEE (2011)
18. Klir, G.J., Yuan, B.: Fuzzy Sets and Fuzzy Logic: Theory and Applications, p. 563. Prentice Hall, Upper Saddle River (1995)
19. Kuhn, W.: Geospatial semantics: why, of what, and how? In: Spaccapietra, S., Zimányi, E. (eds.) Journal on Data Semantics III. LNCS, vol. 3534, pp. 1–24. Springer, Heidelberg (2005). https://doi.org/10.1007/11496168_1

20. Lenz, R., Beyer, M., Kuhn, K.A.: Semantic integration in healthcare networks. Int. J. Med. Inform. **76**(2–3), 201–207 (2007)
21. Lenzerini, M.: Data integration: a theoretical perspective. In: Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, pp. 233–246. ACM (2002)
22. Madnick, S., Zhu, H.: Improving data quality through effective use of data semantics. Data Knowl. Eng. **59**(2), 460–475 (2006)
23. Magnani, M., Montesi, D.: Uncertainty in data integration: current approaches and open problems. In: Proceedings of the First International VLDB Workshop on Management of Uncertain Data, MUD, pp. 18–32 (2007)
24. Marie, A., Gal, A.: Managing uncertainty in schema matcher ensembles. In: Prade, H., Subrahmanian, V.S. (eds.) SUM 2007. LNCS (LNAI), vol. 4772, pp. 60–73. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-75410-7_5
25. Mate, S., et al.: Ontology-based data integration between clinical and research systems. PLoS ONE **10**(1), e0116656 (2015)
26. Noy, N.F.: Semantic integration: a survey of ontology-based approaches. ACM Sigmod Rec. **33**(4), 65–70 (2004)
27. Patil, L., Dutta, D., Sriram, R.: Ontology-based exchange of product data semantics. IEEE Trans. Autom. Sci. Eng. **2**(3), 213–225 (2005)
28. Pileggi, S.F., Voinov, A.: Perswade-core: a core ontology for communicating socio-environmental and sustainability science. IEEE Access **7**, 127177–127188 (2019)
29. Pileggi, S., Hunter, J.: An ontology-based, linked open data framework to support the publishing, re-use and dynamic calculation of urban planning indicators. In: 15th International Conference on Computers in Urban Planning and Urban Management (2017)
30. Po, L., Sorrentino, S.: Automatic generation of probabilistic relationships for improving schema matching. Inf. Syst. **36**(2), 192–208 (2011)
31. Poggi, A., Lembo, D., Calvanese, D., De Giacomo, G., Lenzerini, M., Rosati, R.: Linking data to ontologies. In: Spaccapietra, S. (ed.) Journal on Data Semantics X. LNCS, vol. 4900, pp. 133–173. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-77688-8_5
32. Shvaiko, P., Euzenat, J.: Ten challenges for ontology matching. In: Meersman, R., Tari, Z. (eds.) OTM 2008. LNCS, vol. 5332, pp. 1164–1182. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-88873-4_18
33. Smith, B., et al.: The obo foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol. **25**(11), 1251 (2007)
34. Solimando, A., Jiménez-Ruiz, E., Guerrini, G.: Detecting and correcting conservativity principle violations in ontology-to-ontology mappings. In: Mika, P., et al. (eds.) ISWC 2014. LNCS, vol. 8797, pp. 1–16. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11915-1_1
35. Solimando, A., Jiménez-Ruiz, E., Guerrini, G.: A multi-strategy approach for detecting and correcting conservativity principle violations in ontology alignments. In: Proceedings of the 11th International Workshop on OWL: Experiences and Directions, OWLED 2014, co-located with ISWC, pp. 13–24 (2014)
36. Solimando, A., Jimenez-Ruiz, E., Guerrini, G.: Minimizing conservativity violations in ontology alignments: algorithms and evaluation. Knowl. Inf. Syst. **51**(3), 775–819 (2017)
37. Stoilos, G., Geleta, D., Shamdasani, J., Khodadadi, M.: A novel approach and practical algorithms for ontology integration. In: Vrandečić, D., et al. (eds.) ISWC 2018. LNCS, vol. 11136, pp. 458–476. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00671-6_27

38. Wache, H., et al.: Ontology-based integration of information-a survey of existing approaches. In: OIS@ IJCAI (2001)
39. Zhang, H., et al.: An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival. BMC Med. Inform. Decis. Mak. **18**(2), 41 (2018). https://doi.org/10.1186/s12911-018-0636-4
40. Zhang, Y., Panangadan, A.V., Prasanna, V.K.: UFOM: unified fuzzy ontology matching. In: Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration, IEEE IRI, pp. 787–794. IEEE (2014)