A Framework for Quality Assessment of Semantic Annotations of Tabular Data

Roberto Avogadro[1], Marco Cremaschi[1](✉), Ernesto Jiménez-Ruiz[2,3], and Anisa Rula[4]

[1] University of Milano - Bicocca, Milano, Italy
{roberto.avogadro,marco.cremaschi}@unimib.it
[2] City, University of London, London, UK
ernesto.jimenez-ruiz@city.ac.uk
[3] University of Oslo, Oslo, Norway
[4] University of Brescia, Brescia, Italy
anisa.rula@unibs.it

Abstract. Much information is conveyed within tables, which can be semantically annotated by humans or (semi)automatic approaches. Nevertheless, many applications cannot take full advantage of semantic annotations because of the low quality. A few methodologies exist for the quality assessment of semantic annotation of tabular data, but they do not automatically assess the quality as a multidimensional concept through different quality dimensions. The quality dimensions are implemented in STILTool 2, a web application to automate the quality assessment of the annotations. The evaluation is carried out by comparing the quality of semantic annotations with gold standards. The work presented here has been applied to at least three use cases. The results show that our approach can give us hints about the quality issues and how to address them.

Keywords: Data quality · Semantic annotations · Tabular data · Semantic table interpretation

Resource type	Software Framework
Website	https://bitbucket.org/disco_unimib/stiltool/
Permanent URL	http://doi.org/10.5281/zenodo.4704645.

1 Introduction

Much information is conveyed within tables. A prominent example is the large set of relational databases or tabular data present on the Web. To size the spread of tabular data, 2.5M tables have been identified within the Common Crawl repository [12]. The current snapshot of Wikipedia contains more than 3.23M tables from more than 520k Wikipedia articles [7]. The tables may contain high-value

A. Hotho et al. (Eds.): ISWC 2021, LNCS 12922, pp. 528–545, 2021.
https://doi.org/10.1007/978-3-030-88361-4_31

data, but they can be challenging to understand both for humans and machines due to the lack of contextual information or metadata. In order to solve this problem, several techniques have been proposed in the state-of-the-art, whose aim is the semantic annotation of tabular data using information extracted from a Knowledge Graph (KG) (*e.g.*, DBpedia[1]). However, modelling and constructing semantically annotated datasets poses different quality issues due to: *(i)* the automatic procedures which are often error-prone; *(ii)* the autonomous information providers who are not aware of the final usage of the dataset; *(iii)* the schema-last approach which allows to first publish the data and optionally creates the schema. These may create several concerns with regard to the quality of the annotations.

There already exist some approaches which are focused on the quality assessment of the datasets [4,19]. Besides conceptual and theoretical considerations, several tools and methodologies for practical assessment are proposed [4,18]. However, most of these approaches are focused on the quality assessment of datasets and not on the quality assessment of the process used to transform tabular data to their semantic representation. Instead, a few approaches are proposed for the quality assessment of the mappings generated by the mapping languages such as R2RML [5,6,11,14,16]. As explained by the authors in [5], the root cause of the low quality of datasets is often due to the problems encountered during the mapping phase, such as inconsistencies with the KG schema. Inspired by the approaches proposed for the quality assessment of mapping languages, we think that an approach proposed for the quality assessment of the annotation process would be of benefit for the consumption of the semantic annotations.

To better understand the quality issues in a semantic annotation process but, at the same time, their root causes, we provide an open-source framework within the STILTool system [1], named STILTool 2. First, we need to measure and assess the quality of the steps belonging to the semantic annotation process through several quality dimensions. There are different possible ways to assess semantic table annotations, either employing a gold standard or not. As explained in [15] the assessment through gold standards may present advantages (*e.g.*, highly reliable results) and disadvantages (*e.g.*, costly to produce). While other frameworks such as Luzzu [4] implement only metrics that do not use a gold standard, our framework STILTool 2 has the advantage that its architectural design choices allow the implementation of metrics that require or not a gold standard. Second, we aim to guide the users to understand the real causes of the detected quality issues. STILTool 2 is not only able to assess the quality metrics on semantic annotations similarly to Luzzu, but it also provides hints on the possible quality issues in the process of semantic annotation. The insights gained from such assessment are useful to inform users about particular problems and help identify which stage of the annotation process must be improved.

In this work, we make the following contributions:

- we provide a methodology that can be used to characterise the levels of quality for a semantically annotated dataset;

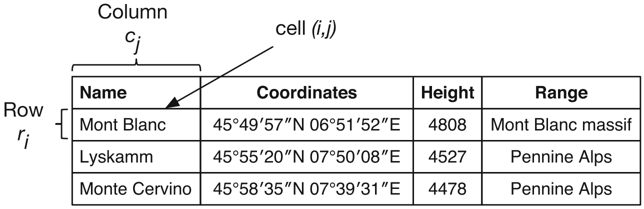
[1] https://wiki.dbpedia.org/.

- we introduce our (open-source) quality assessment framework to be adopted by the SemTab 2021 challenge [9,10];
- we evaluate our approach empirically;
- we briefly present three use cases where STILTool 2 can be used.

The rest of the paper is organised as follows: an overview of the semantic annotation steps is given in Sect. 2. The approach for the assessment of quality metrics for each step of the semantic annotation is detailed in Sect. 3. Details of the architectural and implementation choices are discussed in Sect. 4. Evaluation is provided in Sect. 5. Related work on the assessment of quality metrics is discussed in Sect. 6. Finally, we conclude and suggest planned extensions of our framework in Sect. 7.

2 Semantic Annotation Tasks

In order to produce the annotation of tabular data, it is necessary to take two elements as input: *(i)* a *well-formed and normalised* relational table T (*i.e.*, a table with headers and simple values, thus excluding nested and figure-like tables), as the one in Fig. 1, and *(ii)* a KG which describes real world entities in the domain of interest (*i.e.*, a set of concepts, datatypes, predicates/properties, instances, and the relations among them), as the example in Fig. 2. The table in Fig. 1 is extracted from T2Dv2 gold standards[2]. The output returned is a semantically annotated table, as shown in Fig. 3.


Column
c_j
cell *(i,j)*
Row
r_i

Name	Coordinates	Height	Range
Mont Blanc	45°49′57″N 06°51′52″E	4808	Mont Blanc massif
Lyskamm	45°55′20″N 07°50′08″E	4527	Pennine Alps
Monte Cervino	45°58′35″N 07°39′31″E	4478	Pennine Alps

Fig. 1. Example of a well-formed relational table T, with labels that are used in this paper.

We can identify three types of annotations of tabular data [9]: *(i)* Column-Type Annotation (CTA), *(ii)* Columns-Property Annotation (CPA) and *(iii)* Cell-Entity Annotation (CEA). These tasks can be performed by humans or by automatic or semi-automatic approaches. The *CTA* expects the prediction of the semantic types (*i.e.*, KG classes or concepts) for every given table column c_j in a table T, *i.e.*, $CTA(T, c_j, KG) = st_1, ..., st_a$. The *CEA* requires the prediction of the entity or entities (*i.e.*, instances) that a cell $(i, j) \in T$ represents, *i.e.*,

[2] http://webdatacommons.org/webtables/goldstandardV2.html table index: 14311244_0_7604843865524657408, 49801939_0_6964113429298874283.

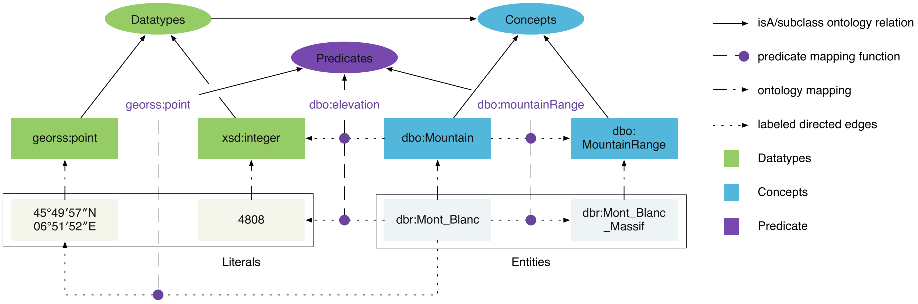
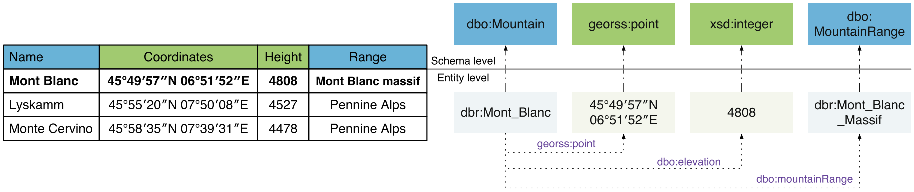


Fig. 2. A sample of Knowledge Graph.



Name	Coordinates	Height	Range
Mont Blanc	**45°49′57″N 06°51′52″E**	**4808**	**Mont Blanc massif**
Lyskamm	45°55′20″N 07°50′08″E	4527	Pennine Alps
Monte Cervino	45°58′35″N 07°39′31″E	4478	Pennine Alps

Fig. 3. Example of an annotated table.

$CEA(T, (i, j), KG) = e_1, ..., e_b$. Finally, the *CPA* expects as output a set of KG properties that represent the relationship between the elements of the input columns c_j and c_k, *i.e.*, $CPA(T, c_j, c_k, KG) = p_1, ..., p_c$. Note that CTA (resp. CEA) focuses on categorical columns (resp. cells) that can be represented with a KG class (resp. KG entity) [10].

To obtain the three types of annotation described above, various processes have been defined in the state-of-the-art, which we can summarise in these steps:

(i) Semantic classification of columns, which considers the content of the cells of each column c_j to mark a column as *Literal column (L-column)* if values in cells are elements of a datatype (*e.g.*, strings, numbers, dates such as 4808, 10/04/1983), or as *Named-Entity column (NE-column)* if values are elements of a concept (*e.g.*, Mountain, Mountain Range such as Mont_Blanc, Mont_Blanc_massif);
(ii) Detection of the subject column (S-column), which has the goal of identifying, among the NE-columns, the column that all the others are referring to (*e.g.*, the Name column in Fig. 3);
(iii) Concept, entity and datatype annotation, which pairs NE-columns with concepts extracted from the KG by first linking cell entities to KG and then inferring the column concept *st* (*e.g.*, the column Name is associated with Mountain in DBpedia[3]), and L-columns with a datatype *dt* in the KG (*e.g.*, the column Coordinates is of type `georss:point`); and

[3] http://dbpedia.org/resource/Mountain.

(iv) Property annotation, which identifies the relations p between the S-column and the other columns (*e.g.*, Name `dbo:elevation` Height).

3 Quality Assessment of the Annotation Tasks

Data quality is commonly conceived as a multi-dimensional construct [19] with a popular notion of "fitness for use" and can be measured along many abstract concepts named quality dimensions such as accuracy and completeness. The assessment of quality dimensions is based on quality metrics, where the metric is a heuristic that is designed to fit a specific assessment dimension. In this Section, we provide quality metrics and their relations with the annotation steps, which should help to detect possible quality issues in the semantic annotations.

Table 1 summarizes the relationship between the quality metrics (in the rows) and annotation steps (in the columns). In this version of STILTool 2, we provide only metrics for which a gold standard is required. Therefore, all the metrics proposed in Table 7 are considered to be new.

In the following, we propose a methodology composed of three phases where each phase correspond to three different levels of granularity that are: *(i)* a single annotation step in isolation, *(ii)* the combination of two annotation steps at instance level (*e.g.*, CEA and CPA), and *(iii)* the combination of two annotation steps at schema level (*e.g.*, CTA and CPA). For each step, there is a set of metrics applied for capturing the quality issues. Metrics can be further aggregated to produce a single quality score. To each metric we assign a weight according to its importance with respect to the annotation steps. For simplicity, we assign a default weight of 1.0 to all metrics.

Table 1. Relationship between quality metrics and the semantic annotation steps.

		Annotation steps			
Metric	Abbr	Concept annotation	Entity annotation	Datatype annotation	Property Annotation
Concept and datatype completeness	CM1	Y		Y	
Property completeness	CM3				Y
Entity completeness	CM2		Y		
Entity candidate coverage	EC		Y		
Type specificity	TS	Y			
Link completeness	LC		Y	Y	Y
Link accuracy	AC		Y	Y	Y
Abstract link completeness	ALC	Y			Y
Abstract link accuracy	ALA	Y			Y

3.1 Phase I: Quality Assessment of the Single Annotation Step

In this first phase, we focus on assessing the quality in terms of completeness and consistency of the single annotation steps.

Completeness Dimension refers to the degree to which all required information is present in a particular dataset [19].

Concept and Datatype Completeness returns the number of the non missing concepts and datatypes in the semantic annotation with respect to the gold standard. The two annotation steps which can generate issues related to this quality metric are: concept and datatype annotation.

Property Completeness returns the number of the non missing properties in the semantic annotation with respect to the gold standard. The annotation step which can generate issues related to this quality metric is: property annotation. In the example of Fig. 3 the table is annotated with concepts: `dbo:Mountain`, `dbo:MountainRange`; properties: `georss:point, dbo:elevation, dbo:mountainRange`; and datatypes: `georss:point, xsd:integer`. Suppose that the values of the coordinates column are not present in the KG but location names are such as `Haute-Savoie` which in turn is not present in the table. Therefore, it is not possible to annotate the property for the column *Coordinates* since its values are not available in the KG. As such, the metric, completeness of properties will identify two properties out of three.

Entity Completeness returns the number of the non missing entities in the semantic annotation with respect to the gold standard. The annotation step which can generate issues related to this quality metric is: entity annotation. In the example of Fig. 3 the table is annotated with entities such as `dbr:Mont_Blanc` in the NE-columns. Suppose that for disambiguation reasons, the `Lyskamm` mountain cannot find an entity in the KG. Therefore, it is not possible to indicate an entity for that value, as such, the metric completeness of entities will identify two entities out of three.

Entity Candidate Coverage returns the number of correct candidate entities with respect to the gold standard. The annotation step which can generate issues related to this quality metric is: entity annotation. For example, consider the table in Fig. 3, retrieving all the entities candidates for the cells belonging to NE-columns *Name* and *Range* columns. Suppose that the candidates of the cells "Mont Blanc", "Lyskamm" and "Pennine Alps" contain the correct entities from the candidate list obtained by our approach. In this case, the metric will return a coverage of 60%, meaning that only three cells out of five obtained the correct entity in the list of the candidates returned. This metric is also an indication of the upper threshold of the precision of our approach, *i.e.*, whatever we do in the next steps of the selection of the entity, we will never get a precision higher than the coverage score.

Consistency Dimension means that a knowledge base is free of (logical/formal) contradictions with respect to particular knowledge representation and inference mechanisms [19].

Type Specificity returns the number of "specific/generic" types with respect to the gold standard. The annotation step which can generate issues related to this quality metric is: concept annotation. In particular, this metric will not only identify a boolean of correct and wrong concepts but will identify *good* concepts too. These concepts are in a subclass or superclass relationship with the correct concept (also referred to as perfect concept), that is, they are *descendent* and *ancestor* concepts, respectively. For example consider Fig. 3 and suppose our approach annotates the column *Name* with `dbo:NaturalPlace` and *Range* with `dbo:MountainRange`. In this case, we will have one ancestor annotation and one perfect annotation.

3.2 Phase II: Quality Assessment of the Combined Annotation Steps at Instance Level

In this second phase, we focus on assessing the quality in terms of interlinking completeness and accuracy of the combined annotation tasks of CEA and CPA.

Interlinking Dimension refers to the degree to which entities are linked to each other within a data source or among two or more data sources [19]. We are interested to measure the completeness and the accuracy of links (*i.e.*, RDF triples) because the combination of the elements in the triples such as pairs of two entities or, an entity and its property, may provide us additional insights about the coverage or accuracy.

Link Completeness returns the number of the non missing triples in the semantic annotation with respect to the gold standard. The annotation steps which can generate issues related to this quality metric are: entity and property annotation. Referring to Fig. 3, we only have one subject column and the others are either Literal or NE-columns, therefore, the total number of possible triples generated by this table of dimension 3×3 (without considering the subject) is nine. Suppose that our approach generates eight out of nine triples, thus the metric will return 89% of completeness.

Link Accuracy returns the number of correct triples in the semantic annotation with respect to the gold standard. The annotation steps which can generate issues related to this quality metric are: entity and property annotation. While completeness focus on the number of missing triples returned, this metric assesses if all the three elements (subject, property and object) of the triple are correct. Suppose a triple returned from the annotation in Fig. 3 where only the subject is correct <`dbr:Mont_Blanc,dbo:mountainRange,dbr:Mont_Blanc_Massif`> thus, the triple is considered not accurate which will be penalized by assigning a score of zero. While the metrics of completeness and accuracy in **Phase I** indicate the single elements of this triple to be correct, the link accuracy metric captures the errors due to the combination of the elements in a triple.

3.3 Phase III: Quality Assessment of the Combined Annotation Steps at Schema Level

In the third phase we focus on assessing the quality in terms of interlinking completeness and accuracy of the combined annotation tasks of CTA and CPA.

Types Interlinking Dimension refers to the degree to which types are linked to each other through a property. Interlinking aspects can be influenced by the combination of types and property annotation tasks. For example, if two columns are to be annotated with the types A and B in CTA and with the property R in CPA, this combined annotation can be represented as an abstract triple <`A,R,B`>. We are interested to measure the completeness and the accuracy of links which refer to (abstract) RDF triples.

Abstract Link Completeness returns the number of the non missing (abstract) triples in the semantic annotation with respect to the gold standard. As shown in Table 1, the annotation steps which can generate issues related to this quality metric are: concept and property annotation. This metric is similarly calculated as the link completeness metric in **Phase II** where each entity has at maximum one type assigned.

Abstract Link Accuracy returns the numbers of correct (abstract) triples in the semantic annotation with respect to the gold standard. As shown in Table 1, the annotation steps which can generate issues related to this quality metric are concept and property annotation. For example, if we consider the (abstract) triple <`dbo:NaturalPlace, dbo:locatedInArea, dbo:MountainRange`> generated by the annotation, the metric will identify it as not correct with respect to the gold standard, although the elements separately can be correct (*e.g.*, `dbo:MountainRange` and `dbo:NaturalPlace` are both ancestors)

4 System Overview and Implementation

Figure 4 shows the general architecture of STILTool 2[4]. The tool is developed with the Django framework[5] in Python, and exploits a MongoDB[6] database as data repository. Three main layers can be identified. Within the *view*, three main components have been implemented. The first component allows to view the list, and manage, the gold standards. The second component allows the management of semantic annotations. The third component of the view allows the visualisation of the loaded tables. For each table, the tool visualises the analysis related to the evaluation metrics (*e.g.*, Accuracy, Recall, F measure) and the quality dimensions described in the previous sections. In the second level, the *controller*, the methods (creation, reading, updating and deletion) for managing the gold standards and the semantic annotations have been implemented. Two components, on the other hand, allow the calculation of quality and evaluation metrics. The controller also allows the query of external KGs (*i.e.*, DBpedia and Wikidata) necessary to calculate quality metrics. In the last level, the *model*, it is possible to identify the representations in the form of an object (ORM) of the entities present in the database.

[4] https://stiltool.disco.unimib.it/.
[5] https://www.djangoproject.com/.
[6] https://www.mongodb.com/.

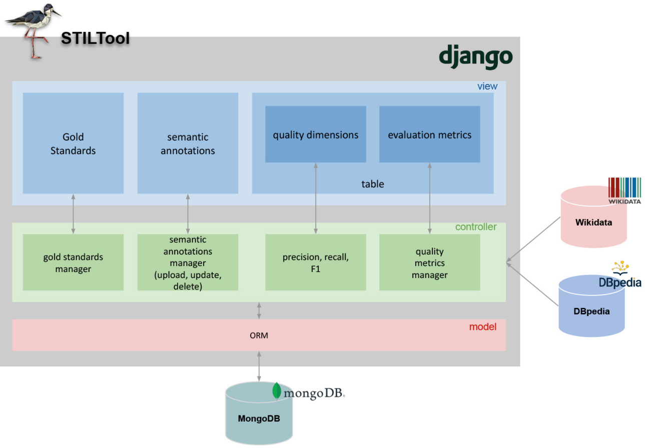


Fig. 4. Architecture of STILTool 2.

The tool is available through a Git repository[7]. The tool has been encapsulated in a Docker container, with an image on Docker Hub[8], to facilitate the deployment and scalability by replication using HAProxy[9]. HAProxy an open-source software that provides a load balancer and proxy server for TCP and HTTP-based applications that spreads requests across multiple servers. It is written in C and has a reputation for being fast and efficient (in terms of processor and memory usage).

The management of messages is performed by using Task Queues (*i.e.*, Celery Workers[10]). Figure 5 shows two screenshots of the application. The first (left) displays information on metrics, while the second (right) displays statistics on the most common errors.

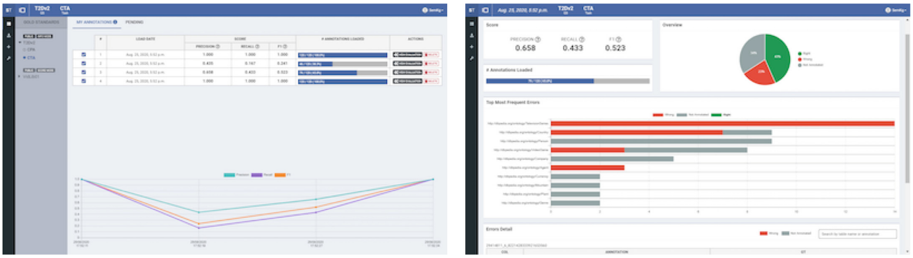


Fig. 5. Screenshots of the STILTool 2.

[7] https://bitbucket.org/disco_unimib/stiltool/.
[8] https://hub.docker.com/repository/docker/cremarco/stiltool.
[9] http://www.haproxy.org/.
[10] https://docs.celeryproject.org/en/latest/userguide/workers.html.

5 Evaluation and Use Cases

The main aim of STILTool 2 is to assess semantically annotated datasets included in different real-world use cases. To shed light on the state of the semantically annotated datasets, we consider the datasets from the SemTab 2020 challenge [10][11]. Specifically, the real-world datasets involved in the challenge represent the multiple kinds of dirty data one finds in practice. We have also selected for the same datasets different annotations proposed by the tools[10] that participated in the challenge.

In the following, we present our experimental setup, including the datasets and their annotations. After that, we give an overview of the quality assessment of the different annotations and provide some insights from the results. With the above considerations in mind, we aim to answer the following questions:

- What are the results of the quality metrics for each annotation provided by a different tool?
- Can we say something about the errors related to the quality assessment result?
- How is the quality evaluation influenced by the KG used?

5.1 Gold Standards

Several approaches on the tabular data annotation have been proposed over the past years. To validate these approaches, several gold standards have also been proposed. Among these, it is possible to mention T2Dv2[12], LimayeAll [13], Limaye200 [21] and Zhang2020 [20]. Furthermore, in the last period, semantic annotation has received an ever-increasing interest within the scientific community. This interest is also shown by the birth of some international challenges, such as "SemTab"[13], already in its second version. The target KG in 2019 was DBpedia [9], while in 2020 was Wikidata [10]. A new gold standard, Tough Tables (2T) [2], was also introduced during SemTab 2020 Round4. In the context of the SemTab 2020 challenge, the table corpora are significantly large with thousands of tables and cells to annotate (*cf.* Table 3).

The approaches of the tabular data annotation only consider one gold standard at a time, meaning that a new gold standard can be uploaded, and the same table can be evaluated on different gold standards separately. In the current gold standards, the tables are annotated using the elements (*i.e.*, entities, classes, properties) coming from the same KG. However, STILTool is agnostic to the use of one or more Knowledge Graphs (KGs) (Table 2).

5.2 Results

We evaluate the proposed approach using the above annotated datasets by different tools. We carry out two set of experiments. Table 4 analyses the semantic

[11] https://www.cs.ox.ac.uk/isg/challenges/sem-tab/2020/index.html.
[12] http://webdatacommons.org/webtables/goldstandardV2.html.
[13] http://www.cs.ox.ac.uk/isg/challenges/sem-tab/.

Table 2. Characteristics of the Gold Standards.

	Table	Columns				Rows				Columns			Concepts	Pred.
		Total	Min	Max	Avg	Total	Min	Max	Avg	S	NE	L		
T2Dv2	234	1157	1	13	4	27966	5	585	119	231	-	-	39	154
Limaye200	200	919	2	11	4	4036	3	102	20	200	504	216	84	-
SemTab2019	14966	75429	1	38	92	515302	1	1533	631	14966	22883	52546	22176	17084
SemTab2020	131648	534892	1	8	23	1401463	2	15477	62	131468	156595	378297	191069	402636

Table 3. Overview of the SemTab 2020 table corpus in each round.

	Round1	Round2	Round3	Round4	
				Standard	Tough Table
Tables	34K	12K	63K	22K	180
Cells to annotate	985K	283K	768K	951K	105K
Unique cells to annotate	264K	138K	378K	516K	23K
Average cell length	20	21	20	14	11

annotation tool Mantistable on two different datasets: SemTab2019 Round4 on DBpedia and SemTab 2020 Round4 on Wikidata - Standard (*i.e.*, without Tough Table). The three metrics considered in the table refer to schema, property and entity completeness (*cf.* Table 1), respectively, with respect to the annotation tasks and the gold standard provided in the SemTab challenge. Mantistable indicates a high quality when DBpedia is used while the quality decreases for the cases of Wikidata which may be explained by the fact that the DBpedia dataset is smaller and less complex than Wikidata and as such the research of correct candidate entities and their disambiguation is easier.

Table 4. Overview of the metrics obtained by Mantistable in Round4 of Semtab 2019 and SemTab 2020.

Approach	Round4					
	DBpedia			Wikidata (Standard)		
	CM1	CM2	CM3	CM1	CM2	CM3
Mantistable	0.99	0.998	0.331	0.579	0.702	0.685

Table 5 shows the approaches assessed according to the completeness metrics. In this case, the two KG used are SemTab 2020 Round4 on Wikidata - Standard and SemTab 2020 Round4 on Wikidata - Tough Table, but since the latter does not cover the CPA, thus we cannot provide CM3. The results shown in the table for the metrics CM1 and CM2 are higher for Round4 - Standard than Round4 - Tough Table. In particular, we notice this huge difference on CM2, which indicates that the entity annotation task performed on Round4 - Tough Table is more difficult to be performed since the dataset itself is complex. We notice that on the best four scores for Round4 - Standard on metric CM1 are by the approaches *SSL, LinkingPark, MTab4Wikidata, bbw* while the worst is from *Kepler-aSI*. In Round4 - Tough Table the best scores on metric CM1 is *SSL, MTab4Wikidata, LexMa, AMALGAM* and the worst continues to be *Kepler-aSI*. The first approach SSL remains constant while some others get worse, and some that were not having high scores in the Round4 - Standard are getting higher scores in Round4 - Tough Table. Overall, we may conclude that some approaches remain almost constant (high/low score) in both Round4 - Standard and Round4 - Tough Table and another group although have a high score on Round4 - Standard get worse either on CM1 or in CM2 in Round4 - Tough Table, *i.e.*, this indication of low quality on instance or schema level will need two different directions of improvements.

Table 5. Overview of the metrics calculated for the different SemTab 2020 approaches in Round4.

Approach	Round4				
	Standard			Tough Table	
	CM1	CM2	CM3	CM1	CM2
AMALGAM	0.993	0.954	-	0.991	0.412
bbw	0.999	0.989	0.999	0.483	0.869
dagobah	0.998	0.998	0.998	0.924	0.379
JenTab	0.998	0.996	0.998	0.876	0.527
Kepler-aSI	0.23	0.016	-	0	0.001
LexMa	-	0.864	-	0.998	0.585
LinkingPark	1.0	1.0	0.993	0.994	0.998
MTab4Wikidata	0.999	1.0	1.0	1.0	0.998
SSL	1.0	1.0	1.0	1.0	0.99

Table 6 shows the quality assessment according to the metrics of Phase II and Phase III of the approach, in particular, Interlinking Completeness and Accuracy Completeness of triples and (abstract) triples. As we may notice, completeness is higher than accuracy which is an indication that while the retrieved entities, properties and types are almost the same as indicated by the gold standard the correctly retrieved entities, properties and types are less.

Table 6. Overview of the metrics calculated on CEA and CTA triples for the different SemTab 2020 approaches in Round4.

Approach	Round4			
	Standard CEA triples		Standard CTA triples	
	LC	AC	ALC	ALA
Mantistable	0.698	0.685	0.491	0.475
bbw	0.975	0.941	0.996	0.912
dagobah	0.993	0.966	0.998	0.908
JenTab	0.989	0.949	0.992	0.792
LinkingPark	0.997	0.939	0.957	0.799
MTab4Wikidata	0.996	0.982	0.997	0.924
SSL	0.885	0.808	0.998	0.889

Table 7 shows the results for the Type Specificity (TS) metric provided by the different approaches. It considers how many times the perfect annotation has been identified. In case when the perfect annotation is not retrieved, it looks for the first ancestor or first descendent; otherwise, the type is classified as an error. The results of this metric show that in most cases the problem is not the most specific or generic type but most of the approaches get the wrong types. These cases are due to a wrong identification of the type or the type was not found.

Table 7. Overview of the metrics calculated on type specificity (TS) for the different SemTab 2020 approaches in Round4.

Approach	Round4							
	Standard				Tough Table			
	perfect	ancestor	descendent	error	perfect	ancestor	descendent	error
Mantistable	0.56	0.003	0.005	0.425	0.304	0	0.113	0.583
AMALGAM	0.833	0.023	0.009	0.135	0.515	0.004	0.15	0.331
bbw	0.966	0.018	0.002	0.014	0.289	0.072	0.078	0.561
dagobah	0.944	0.039	0.001	0.016	0.511	0.228	0.043	0.219
JenTab	0.894	0.043	0.004	0.059	0.502	0.08	0.041	0.378
Kepler-aSI	0.147	0.007	0.006	0.84	0	0	0	1.0
LinkingPark	0.913	0.055	0.005	0.027	0.58	0.093	0.067	0.261
MTab4Wikidata	0.963	0.018	0.012	0.008	0.617	0.033	0.146	0.204
SSL	0.927	0.024	0.002	0.047	0.27	0.043	0.102	0.585

Table 8 shows the Entity Candidate Coverage metric obtained only by the Mantistable approach because the data about the candidate entities were not available for the other approaches. As shown from the results, the Round4 - Tough Table has a coverage value of 0.748 because of its complexity, while

Round4 - Standard has almost a total coverage. This value in Round4 - Tough Table indicates that the next steps of the STI annotation process will not improve the results. Therefore, this metric serve as an upper limit and thus will influence our decision on proceeding or not with the subsequent steps of the STI process *i.e.*, we learn a priori that if we run all the other steps we will get an equal or even a worse score. Thus this metric may save time and resources.

Table 8. Entity Coverage metrics calculated on Mantistable approach.

Approach	Round4	
	Standard	Tough Table
	EC	EC
Mantistable	0.989	0.748

Flexibility. Our approach is flexible since it evaluates different types of metrics according to cells, columns or rows.

Correctness of Metrics. In order to test the correctness of implemented metrics, we have implemented unit tests and in cases of small datasets we have checked the result obtained by our approach manually.

5.3 Use Cases

The proposed quality assessment framework may be used in many use cases. These includes the following three scenarios:

Comparison and Evaluation of Semantic Table Interpretation (STI) Approaches. The framework can be used for comparing different STI approaches. The functionalities of the previous version of STILTool have been defined as part of the SemTab 2020 challenge. The organizers of this challenge have expressed their intention to adopt STILTool 2 as part of the next challenge, SemTab 2021.

Integration and Quality Assessment of Product Data. In this scenario it is required to integrate product data by first annotating them. The semantic annotations are the main driver for the integration of product datasets. One of the key features of the integration process is the data fusion task. Consider two different semantically annotated datasets containing product data and their properties, as well as a set of hierarchies of types connected to entities. The data fusion process produces a third, final dataset, containing consolidated descriptions of the linked product data. This process depends on the quality of the input data, therefore, it requires a mechanism for data quality check. We use STILTool 2 to check the quality of each input dataset against the gold standard. If the two datasets are annotated using two different KGs then STILTool 2 will take as input two different gold standards. To assure the quality of the fusion process we need to have annotations with high quality.

Natural Language Generation of RDF Triples. A considerable amount of data, presented in a structured, tabular form, is available on the Web nowadays. For the informational content of such data to be made accessible and understandable to *all* users, its translation into natural language can be a valid solution. Table summarisation is the process of obtaining a summary of the tabular data in such a way as to describe the complex information it conveys. This summary can be generated concerning the interest and information needs of the user. In this scenario, it is evident the importance of the high quality of annotations. Considering the table in Fig. 1, an incorrect annotation relating to the first cell (Mont Blanc) would completely distort the sentence's meaning; for instance, a sentence relating to Mont Blanc on the moon[14] could be generated, conveying utterly incorrect information. In this scenario, deep learning models, particularly Neural Machine Translation models, are used for sentence generation. In this case, STILTool 2 can be used to measure the quality of the annotated datasets. For example, the evaluation of the WebNlg 2017 dataset which should use triples extracted from DBpedia, allowed us to identify some properties not currently present within this KG.

5.4 Limitations

As described in the previous sections, STILTool allows measuring the quality of a dataset by using gold standards, but data quality is commonly conceived as "fitness for use" for a specific application or real-world cases, meaning it will be subjective. For example, there are cases where the same tabular data can be annotated differently, depending on the user's needs and related design choices (*e.g.*, use of different vocabularies). Gold standards can be used to fit a particular (potentially narrow, but controlled) view of the task by making certain assumptions with a specific purpose in mind. However, to create semantic table annotation approaches that can satisfy real-world needs, it is necessary to consider the output to achieve (*i.e.*, in term of annotations), so we need controlled and predefined scenarios to get specific insights and evaluate the approaches. The desired output can then be described through a gold standard, partial gold standard, or silver standard, which can be used for enabling a fully automated evaluation. The generality is guaranteed since STILTool allows the use of different gold standards, both those defined in the state-of-the-art and those defined by users, to satisfy real-world needs.

Regarding the second limitation, which is related to approaches tailored to a specific type of semantic table annotation, STILTool considers all the annotation tasks described in Sect. 2. In particular, the Columns-Property Annotation task involves identifying a subject column to define the relationships between the subject column and the other columns in the form of properties. However, in the

[14] https://en.wikipedia.org/wiki/Mont_Blanc_(Moon).

current version our tool considers all the steps and there may be some limitations for other semantic annotation tools that address only some of the tasks (*e.g.*, apply CPA without identifying subject columns). In future versions of the tool, it will be possible to evaluate annotations of tables without subject columns or with more than one subject column, to introduce greater generalisation and therefore consider more real-world cases.

6 Related Work

In this work, we propose quality assessment metrics for semantically enriched tabular data as a result of an STI process and its annotations. Different approaches have been proposed to assess the quality of knowledge graphs. The approaches of quality assessment can be distinguished in those applied to the quality of datasets [4,19] and mapping definitions [6] which can be further classified into i) manual, ii) semi−automatic and iii) automatic. In particular, the work in [19] focuses on the definitions and formalisation of quality assessment metrics for knowledge graphs. In a more recent work, [4] proposes the formalisation of the quality metrics from the practical and implementation point of view. In [8], the authors evaluate the quality assessment of crawled datasets containing around 12M RDF triples. The main aim was to discuss common problems found in RDF datasets, and possible solutions. The authors also provide suggestions on how publishers can improve their data, so that consumers can find "high-quality" datasets. However, these approaches do not provide any quality metrics for the transformation process.

A number of works have been published on the quality assessment of RDF mapping languages [5,6,11,14,16]. The existing literature tends to focus on a particular subset of quality metrics. Randles et al. [16] propose a framework to assess and improve the quality of R2RML mapping language. The quality metrics are provided in SHACL which require additional knowledge on writing them. The work in [11] propose an extension of the quality assessment framework, Luzzu [3], which is mainly used for the quality assessment of the RDF datasets by introducing four quality metrics for the quality assessment of mappings. The authors in [14] propose a tool for the quality assessment of mappings. In [6], the authors assess mapping definitions from semistructured data to RDF by proposing an incremental, iterative and uniform validation workflow where violation might arise from incorrect usage of schemas, in addition, they suggest mapping refinements based on the results of these quality assessments. The authors have extended RDFUnit to also cover the validation of mappings against its vocabularies and ontologies. Dimou et al. [5] demonstrate that assessing an RDF dataset requires a considerable measure of time, therefore it cannot be often executed, and when that happens, the violations' root is not detected. On the other hand, assessing the RDF mappings requires essentially less time and the violations' root can be detected. There is (to the best of our knowledge) no study to support the quality assessment of the STI process and its annotations.

7 Conclusions and Future Work

STILTool 2 aims to perform a quality assessment of semantic annotations of tabular data. To the best of our knowledge, our proposal is one of the most comprehensive frameworks to support the evaluation of the semantic annotation process. STILT2 can be used in evaluation and comparison over the different tasks of semantic annotation. The modularity of STILTool 2 allows us to implement and extend with other quality metrics, which can operate in one of the phases as defined in Sect. 3. The framework has been published with an open-source licence in order to be used by the whole community. STILTool 2 will be adopted by the organisers of SemTab 2021 to support the evaluation campaign. SemTab participants will also potentially benefit from the use of the framework.

In the future, we plan to maintain and extend the tool with additional quality metrics such as Correct Object/Datatype Property Values. Another direction is to analyse the root causes by not only visualising the aggregated scores of quality but highlighting the quality issues in the annotations. We also plan to introduce estimated quality metrics that may need a partial gold standard and indicate the quality score for the whole dataset. If a gold standard is not available, our goal is to store versions of different annotations applied on the same dataset to analyse their evolution. One key point in the evolution analysis is the computation of quality metrics between different versions to detect the quality issues [17].

References

1. Cremaschi, M., Siano, A., Avogadro, R., Jimenez-Ruiz, E., Maurino, A.: STILTool: a semantic table interpretation evaLuation tool. In: Harth, A., et al. (eds.) ESWC 2020. LNCS, vol. 12124, pp. 61–66. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-62327-2_11
2. Cutrona, V., Bianchi, F., Jiménez-Ruiz, E., Palmonari, M.: Tough tables: carefully evaluating entity linking for tabular data. In: Pan, J.Z., et al. (eds.) ISWC 2020. LNCS, vol. 12507, pp. 328–343. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-62466-8_21
3. Debattista, J., Auer, S., Lange, C.: Luzzu-a methodology and framework for linked data quality assessment. JDIQ **8**(1) (2016)
4. Debattista, J., Lange, C., Auer, S., Cortis, D.: Evaluating the quality of the LOD cloud: an empirical investigation. SWJ **9**(6), 859–901 (2018)
5. Dimou, A., et al.: DBpedia mappings quality assessment. In: Poster & Demo at ISWC, vol. 1690. CEUR (2016)
6. Dimou, A., et al.: Assessing and refining mappings to RDF to improve dataset quality. In: Arenas, M., et al. (eds.) ISWC 2015. LNCS, vol. 9367, pp. 133–149. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25010-6_8
7. Fetahu, B., Anand, A., Koutraki, M.: TableNet: an approach for determining fine-grained relations for Wikipedia tables. In: WWW 2019, pp. 2736–2742. ACM (2019)
8. Hogan, A., Umbrich, J., Harth, A., Cyganiak, R., Polleres, A., Decker, S.: An empirical survey of linked data conformance. JWS **14**, 14–44 (2012)

9. Jiménez-Ruiz, E., Hassanzadeh, O., Efthymiou, V., Chen, J., Srinivas, K.: SemTab 2019: resources to benchmark tabular data to knowledge graph matching systems. In: Harth, A., et al. (eds.) ESWC 2020. LNCS, vol. 12123, pp. 514–530. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-49461-2_30
10. Jimenéz-Ruiz, E., Hassanzadeh, O., Efthymiou, V., Chen, J., Srinivas, K., Cutrona, V.: Results of SemTab 2020. In: Proceedings of the Semantic Web Challenge on Tabular Data to Knowledge Graph Matching, vol. 2775, pp. 1–8 (2020)
11. Junior, A.C., Debattista, J., O'Sullivan, D.: Assessing the quality of R2RML mappings. In: Joint Proceedings of the 1st Sem4Tra and the 1st AMAR at SEMANTiCS 2019), vol. 2447. CEUR (2019)
12. Lehmberg, O., Ritze, D., Meusel, R., Bizer, C.: A large public corpus of web tables containing time and context metadata. In: WWW 2016, pp. 75–76. ACM (2016)
13. Limaye, G., Sarawagi, S., Chakrabarti, S.: Annotating and searching web tables using entities, types and relationships. VLDB **3**(1–2), 1338–1347 (2010)
14. Moreau, B., Serrano-Alvarado, P.: Assessing the quality of RDF mappings with EvaMap. In: Harth, A., et al. (eds.) ESWC 2020. LNCS, vol. 12124, pp. 164–167. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-62327-2_28
15. Paulheim, H.: Knowledge graph refinement: a survey of approaches and evaluation methods. Semant. Web **8**(3), 489–508 (2017)
16. Randles, A., Junior, A.C., O'Sullivan, D.: Towards a vocabulary for mapping quality assessment. In: The 15th OM at (ISWC 2020), vol. 2788, pp. 241–242 (2020)
17. Rashid, M., Torchiano, M., Rizzo, G., Mihindukulasooriya, N., Corcho, O.: A quality assessment approach for evolving knowledge bases. SWJ **10**(2), 349–383 (2019)
18. Sejdiu, G., Rula, A., Lehmann, J., Jabeen, H.: A scalable framework for quality assessment of RDF datasets. In: Ghidini, C., et al. (eds.) ISWC 2019. LNCS, vol. 11779, pp. 261–276. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-30796-7_17
19. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: a survey. SWJ **7**(1), 63–93 (2016)
20. Zhang, S., Meij, E., Balog, K., Reinanda, R.: Novel entity discovery from web tables. In: WWW 2020, pp. 1298–1308. ACM (2020)
21. Zhang, Z.: Effective and efficient semantic table interpretation using tableminer+. SWJ **8**(6), 921–957 (2017)