A Template-Based Approach for Annotating Long-Tail Datasets

Daniel Garijo, Ke-Thia Yao, Amandeep Singh, and Pedro Szekely*

Information Sciences Institute,
University of Southern California
{dgarijo, kyao, amandeep, szekely}@isi.edu

Abstract. An increasing amount of data is shared on the Web through heterogeneous spreadsheets and CSV files. In order to homogenize and query these data, the scientific community has developed Extract, Transform and Load (ETL) tools and services that help making these files machine readable in Knowledge Graphs (KGs). However, tabular data may be complex; and the level of expertise required by existing ETL tools makes it difficult for users to describe their own data. In this paper we propose a simple annotation schema to guide users when transforming complex tables into KGs. We have implemented our approach by extending T2WML, a table annotation tool designed to help users annotate their data and upload the results to a public KG. We have evaluated our effort with six non-expert users, obtaining promising preliminary results.

Keywords: Dataset annotation · Metadata · Knowledge Graph.

1 Introduction

An increasing amount of data is shared on the Web by multiple organizations using Excel and CSV formats. Content creators usually prefer to use tabular data because it is simple to generate, manipulate and visualize by humans; and there is a significant number of tools to help explore and edit the contents of spreadsheets. These data need to be properly understood by others, and hence documentation (e.g., variables captured, provenance, usage notes, etc.) is usually included in auxiliary files or the spreadsheets themselves. As a result, many of these spreadsheets have comments, clarifications, notes and references to other files explaining how to interpret the information contained in them.

In order to convert tabular data to a machine readable format, the Semantic Web community has created Extract, Transform and Load (ETL) tools (e.g., [4]) and mapping languages (e.g., [1, 5]) that help translating spreadsheets into Knowledge Graphs. However, these tools and languages require significant expertise when transforming heterogeneous tabular data with comments, incomplete values or columns that are interrelated to each other, making it difficult for domain experts to integrate their own datasets with existing KGs.

In this paper we describe an approach to help non-experts transform their data into a structured representation through dataset annotations. Our contributions include 1) a dataset annotation schema that helps generating templates for translating datasets into KGs; 2) an extension of the T2WML dataset annotation tool [6] to accommodate the proposed schema; and 3) an approach to upload annotated datasets to a registry once the dataset annotation is complete.

In order to assess our approach, we conducted a preliminary evaluation with 6 users unfamiliar with Knowledge Representation or Semantic Web technologies, who were able to describe and integrate their annotated datasets as a KG.

2 Challenges in Long-Tail Dataset Annotation

We focus on those datasets that are not straightforward to map into a structured representation. Consider for example Table 1, which depicts the food prices in different regions of Ethiopia at different points in time. The table has a time series for the price value of different items at different dates, a repeated column with the item being described (ignore), the item category and different information about the region where that item was produced. The dataset has also some missing values and labels marked as "unknown", which we may want to skip. This dataset is representative of many open datasets with statistical/time series information, and presents some interesting challenges:

- **The main subject of the annotation is not clear**: The table describes the price of an item in a location at a particular time. One possibility would be to assert that the subject of the triple is the item (e.g., Sorghum), having the price column as the object; and the rest of the columns as qualifiers. Alternatively, we could use the country (or the administrative name) as main subject, as it is relevant to create aggregates. Finally, we could also generate a blank node or URI to link together the contents of all columns.
- **Repeated columns and incomplete cell values**: Spreadsheets contain empty values, cell values (or columns) that need to be ignored and comments (specially at the beginning and end) that complicate processing the data.
- **Distinguishing variables from qualifiers**: In some cases, it may be difficult to distinguish whether a column is the object associated to a subject or whether it is qualifying other values. For example, if Table 1 contained a "quality" column, it could be interpreted as a new variable, or as a qualifier indicating the quality of the information source.

Other problems that frequently occur include complex headers that sometimes join the meaning of two columns (e.g., values and units, location and country, etc.); comments in certain parts of the file; or critical missing information, which is externally provided to the file. For example, there are cases where the year in which the file was produced is part of the title of the CSV instead of a column with a constant name.

All these challenges make the automated annotation of datasets a challenging problem. We need an approach for incorporating user feedback from content

Table 1. Table 1: Example of a dataset with food prices in Ethiopia

date	ignore	item	name	category	price	curr	country	admname
7/15/2000	Sorghum	Sorghum	Wholesale	cereals/tubers	238	ETB	Ethiopia	Addis
7/15/2001	Rice	Rice	Retail	cereals/tubers	19	ETB	Ethiopia	Afar
7/15/2002	Rice	Rice	Retail	cereals/tubers	18	ETB	Ethiopia	unkown
7/15/2003	Sorghum	Sorghum	Retail	cereals/tubers		ETB	Ethiopia	Amhara

creators or domain experts that are familiar with these datasets, but do not necessarily know Semantic Web technologies or mapping languages.

3 Using Annotation Templates to Structure Datasets

Our approach has three main elements: an annotation schema, which we use to create mapping templates (Section 3.1); an extension of the T2WML tool to use the proposed vocabulary when converting datasets into KGs (Section 3.2); and an approach to integrate the mapped results with a reference KG (Section 3.3).

3.1 A Schema to Describe Variable Metadata

We have created a simple annotation schema[1] by adding a set of headers to the start of spreadsheet as shown in Table 2. The schema was designed to capture basic metadata and to be easy to understand by content creators unfamiliar with Semantic Web technologies. Therefore we capture 1) the **dataset identifier** to be used when referring to the dataset; 2) the **role** of each column, i.e., whether it is a variable, a unit or a qualifier (location, time or other); 3) the **type** of each column, i.e., whether the column should be the main subject, the format used to represent dates, whether the variables to annotate are a number or a string, etc.; the 4) **column description** in case users need to clarify any of the columns to the persons reusing the data; 5) the **variable name** represented in a column, as in some cases the headers used are difficult to understand; 7) the **variable unit**; and 8) the **header** where the original dataset headers start.

An example of our schema is represented in Table 2 by annotating Table 1. As shown in the example, it is not necessary to complete all headers, in case the information is not known or missing.

3.2 Extending the Table to Wikidata Mapping Language Tool

We have implemented our approach by extending the Table to Wikidata Mapping Language Tool (T2WML) [6]. T2WML is designed to 1) map data in arbitrary data layouts used in Excel and CSV files without the need of complex preprocessing steps to transform tables into a canonical "Database" representation; 2) Enable users who are not familiar with RDF to map spreadsheets and CSV files to KGs; and 3) Integrate mapping and entity linking so that the resulting output is linked to a reference KG.

[1] https://t2wml-annotation.readthedocs.io/en/latest/

Table 2. Example of a dataset using our proposed schema

dataset	Eth-FoodPrices						
role	time	qualifier	qualifier	variable	unit	location	location
type	%m/%d/%Y	string	string	number		country	main_subject
desc.		Name of the crop		Price in Ethiopia			
name		Crop name		food price			
unit							
header	**date**	**item**	**category**	**price**	**curr.**	**country**	**admname**
	7/15/2000	Sorghum	cereals and tubers	238	ETB	Ethiopia	Addis Ababa
	7/15/2001	Rice	cereals and tubers	19	ETB	Ethiopia	Afar

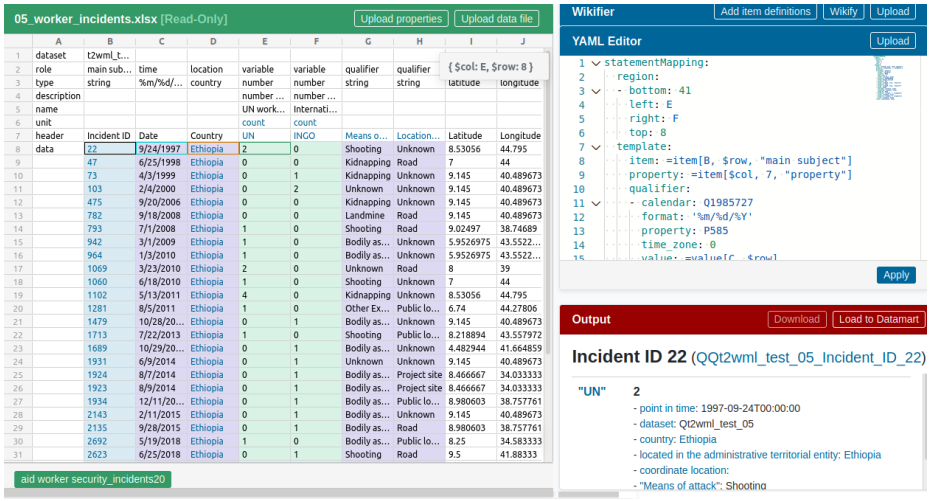

Fig. 1. T2WML screenshot with the annotation schema and mapping template (right). Users can click on the CSV cells to previsualize their results on the bottom right.

T2WML is designed for the Wikidata data model [7]. The main building block in this model is a *statement*, which consists of a subject, a predicate, an object, qualifiers and references. The subject, predicate and object part mirror their RDF counter parts. The qualifiers are predicate/object pairs that provide context information about a subject/predicate/object triple. For example, an employment relation between a person and an organization can be qualified to record the period of time when the person was employed at that organization.

Figure 1 shows how the T2WML extension would process a dataset similar as the one shown in Table 2. T2WML recognizes the different headers annotated in the spreadsheet to generate a template YAML following the T2WML mapping language [6]. Mapped results can be previsualized on the bottom right of the screen, under "Output". This way, users can see how the automatically proposed mappings will process the dataset and edit them accordingly in case of need.

3.3 Uploading Annotated Results to a Public Knowledge Graph

Once users finish annotating a dataset, they can export their results in a structured format like RDF. However, creating a KG with this information still needs significant expertise. Therefore, we have created the USC Datamart, a catalog which includes 1) key dataset metadata (i.e., creator, variables included, etc.) of the datasets uploaded by users; and 2) the contents of those annotated datasets (with variables and their qualifiers like location, date, units, etc.). We have extended T2WML to allow uploading the structured results into the USC Datamart through a dedicated API[2], enabling users to share their results online (see the *Upload to Datamart* button in Figure 1). Each dataset has its own id, which can be updated with new data. This way if a time series consists on a set of spreadsheets with the same structure for different regions, they can all be uploaded using a similar mapping template and the same dataset id.

With the USC Datamart, users may retrieve dataset metadata (e.g., to find out which variables does a dataset include, or the time period they cover) and once they find the desired information they can download it as a table for their own analysis. A usage example of the Datamart API can be seen online.[3]

4 Preliminary Evaluation

In order to assess our approach, we performed a preliminary evaluation with six users. None of these users were familiar with Semantic Web technologies or mapping languages, but three of them had expertise in data science and scripting languages like Python or R. All users received a training in T2WML (one hour) to understand the main capabilities of the tool and the annotation schema.

The goal of the evaluation was to assess if users could understand the proposed schema and use it in T2WML to annotate and upload datasets similar to the one described in Table 1 (with their corresponding challenges). The evaluation included three datasets with different indicators (economic, demographic, production, etc.) in African countries. Each dataset was assigned to two different users. As a result, all users were able to upload their datasets successfully to the USC Datamart, with on the fly corrections for one of the datasets where the temporal information was part of the title of the file, instead of in its contents.

When asked for feedback, users reported that the proposed annotation approach was preferable to creating their own scripts for data cleaning, but they claimed that it can be difficult to 1) align their own terminology to Wikidata and 2) understand the difference between a variable and their corresponding qualifiers. This means that while our approach successfully tackled the first two challenges described in Section 2 (annotating the main subject and incomplete columns), additional work is required to guide users in the annotation process. We are improving tutorials and documentation to address these issues.

[2] https://github.com/usc-isi-i2/datamart-api
[3] https://tinyurl.com/y2lygs5v

5 Related Work

A significant number of tools (e.g., [4, 5]) and mapping languages (e.g., [1, 2]) have been created by the community to help users map their datasets into KGs. In this work we created a schema to help guide users in the dataset annotation process without having to learn the complexity of existing tools or languages.

Other work has focused on automated table understanding to label the structure of tables without having users to annotate datasets themselves (e.g., [3]). This work is very relevant to our own, and we plan to expand our approach in this direction, (giving users the ability to correct the annotations proposed automatically). In this paper we aim to ensure users understood the proposed schema and also to have an end-to-end process (from annotation to upload) incorporated in a single tool (T2WML).

6 Conclusions and Future Work

In this paper we have described our approach for allowing content creators to describe their own datasets to transform them into structured KGs. Our preliminary results show that users are able to understand and use our schema for annotating their datasets easily, enabling them to create and populate an existing KG. Our next step will focus on incorporating table understanding approaches which will make the process easier for users describing their own data.

References

1. Dimou, A., Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., Van de Walle, R.: RML: a generic language for integrated RDF mappings of heterogeneous data. In: Proceedings of the 7th Workshop on Linked Data on the Web. CEUR Workshop Proceedings, vol. 1184 (Apr 2014)
2. Ermilov, I., Auer, S., Stadler, C.: Csv2rdf: User-driven csv to rdf mass conversion framework. In: Proceedings of the ISEM. vol. 13, pp. 04–06 (2013)
3. Ghasemi-Gol, M., Pujara, J., Szekely, P.: Learning cell embeddings for understanding table layouts. Knowledge and Information Systems (Sep 2020)
4. Gupta, S., Szekely, P., Knoblock, C.A., Goel, A., Taheriyan, M., Muslea, M.: Karma: A system for mapping structured sources into the semantic web. In: Extended Semantic Web Conference. pp. 430–434. Springer (2012)
5. Heyvaert, P., De Meester, B., Dimou, A., Verborgh, R.: Declarative Rules for Linked Data Generation at your Fingertips! In: Proceedings of the 15th ESWC: Posters and Demos (2018)
6. Szekely, P., Garijo, D., Bhatia, D., Wu, J., Yao, Y., Pujara, J.: T2WML: Table to wikidata mapping language. In: Proceedings of the 10th International Conference on Knowledge Capture. p. 267–270. K-CAP '19, ACM (2019)
7. Vrandečić, D., Krötzsch, M.: Wikidata: A free collaborative knowledge base. Commun. ACM **57**(10), 78–85 (Sep 2014)