WDPlus: Leveraging Wikidata to Link and Extend Tabular Data

Daniel Garijo
dgarijo@isi.edu
Information Sciences Institute, University of Southern California
Los Angeles, California

Pedro Szekely
pszekely@isi.edu
Information Sciences Institute, University of Southern California
Los Angeles, California

ABSTRACT

Scientific observations and other open data are usually made available online in a tabular manner as CSVs and spreadsheets. However, users of these data face three main challenges when attempting to use these products: finding which datasets are related to a topic of interest; determining which existing information can be used to extend a given dataset; and how to share their integrated dataset results with the rest of the community. In this paper we present WDPlus, a framework designed to address these challenges by leveraging Wikidata. WDPlus allows searching for heterogeneous datasets, facilitates completing tabular data using Wikidata and proposes a mechanism to extend Wikidata in a decentralized manner.

KEYWORDS

Knowledge Graphs, Entity Linking, Wikidata, RDF

1 INTRODUCTION

Today, data about any domain can be found on the web in data repositories, web APIs and millions of spreadsheets and CSV files. These data comes in a myriad of formats, layouts, terminology and cleanliness that make them difficult to integrate together.

Users of these data face three main challenges. The first one is finding datasets related to a feature or topic of interest. For example, climate scientists often look for years of observational data from authoritative sources when estimating the climate of a region. The second challenge is how to complete a given dataset with existing knowledge: machine learning applications are data hungry and require as many data points and features as possible to improve their predictions, which often requires integrating data from different sources. The final challenge is sharing integrated results: once several datasets have been merged together, how to make them available to the rest of the community?

Knowledge graphs have become the preferred technology to address these challenges. Large organizations, including search engine providers, shopping giants and finance institutions are investing in large knowledge graphs to integrate and retrieve heterogeneous data. However, data integration pipelines are usually created manually, require significant expertise, and are seldom available to the general public. Similarly, linking to existing datasets in the the Linked Open Data Cloud[1] usually requires the expertise of a knowledge engineer to properly identify the appropriate target instances to link to in other datasets.

Recent initiatives such as Data.world,[2] Google data search [2] and DataCommons[3] aim to facilitate linking, searching and accessing some of the contents of these knowledge graphs. However, much work remains to automatically link, extend and integrate tabular data into existing open knowledge graphs. In this paper we propose WDPlus, a framework designed to address these challenges by leveraging Wikidata [5],[4] an open crowdsourced knowledge graph with over 60 million entities, over 700 million statements describing those entities and a thriving community of curators.

2 WDPLUS FRAMEWORK

WDPlus is a framework designed to explore and build large multi-domain knowledge graphs using the wealth of structured data available on the web. Our approach shifts the burden of semantic linking away from data publishers to communities that have an incentive to do so in a collaborative manner.

An overview of WDPlus can be seen in Figure 2. At its core, WDPlus relies on Wikidata [5]. A series of extensions to Wikidata, i.e., the *Wikidata satellites* (represented as red circles in the figure) surround the core enabling different organizations to create multi-domain knowledge graphs tailored to their needs. A metadata index stores table metadata to facilitate retrieving related datasets and other candidates with potential to expand or become a Wikidata satellite.

New data from the web may be processed through our proposed WDPlus toolkit, designed to automatically create prototype table models that can be extended with existing Wikidata knowledge and be refined into full Wikidata extensions. We provide more information about our toolkit, table metadata index and satellites below.

2.1 The WDPlus Toolkit

In order to process and link tables and spreadsheets, we created a toolkit with the following capabilities:

- **Entity Linking**: One of the main challenges when linking a dataset to existing knowledge graphs is to find whether the entities described in the dataset are already defined in a target knowledge graph. In order to facilitate this task, we have created a *CSV Wikifier* (inspired by [1]), to link tabular data to existing Wikidata entities and disambiguate them based on their most common class. Early results on the ISWC 2019 cell-entity annotation challenge[5] rank our

[1] https://lod-cloud.net/

[2] https://data.world/

[3] https://datacommons.org/

[4] http://wikidata.org/

[5] https://bit.ly/2NefkLq

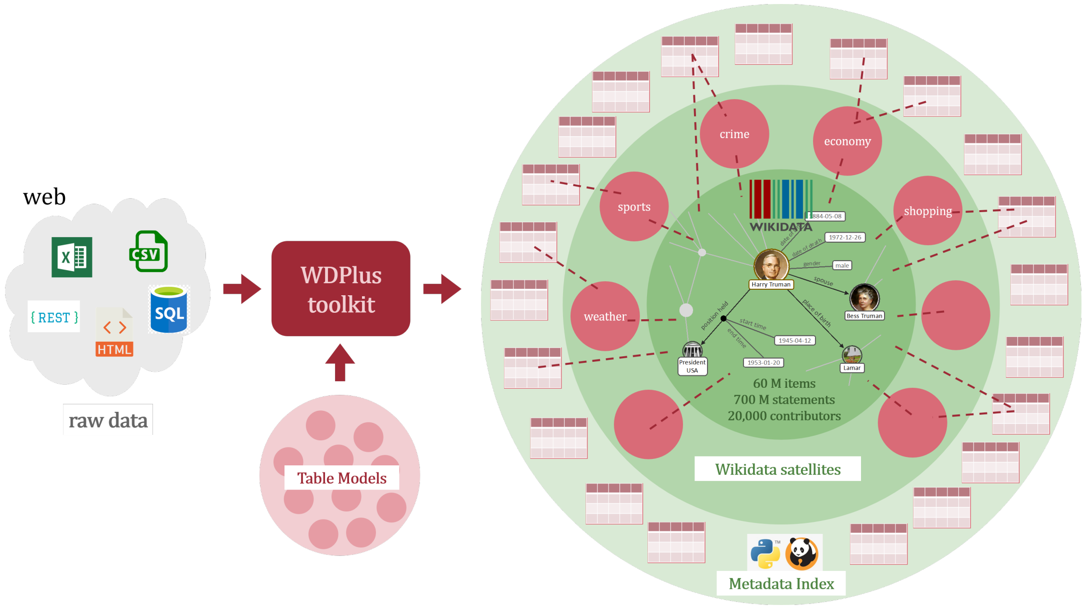


Figure 1: Overview of the WDPlus framework

system within the top performers, with over 0.9 precision and over 0.85 F-1 score.

- **Interactive Table Understanding**: Our toolkit includes a GUI and defines a mapping language [4] to assimilate data from the web as entities and statements compatible with Wikidata. Figure 2 shows a snapshot of the application, which depicts the table to map on the left side of the figure, and the currently specified mapping on the right side. The toolkit allows capturing all the qualifiers associated with a statement, representing provenance, location, timeliness, units and roles of each assertion. These are represented using different colors in the figure (the statement values are represented in green, qualifiers in red and the source of the statements in blue). Table models used for converting tabular data are saved and linked as part of the WDPlus framework.
- **RDF Generation**: Given a table model and a target table, our toolkit includes the means to generate RDF triples that follow the Wikidata data model. These triples can then be browsed and loaded in a Wikidata satellite. WDPlus also mints new triples when new entities or properties do not exist in Wikidata.

The WDPlus toolkit is available online with a MIT License.[6]

2.2 Creating a Metadata Index of Tabular Data

We have created an index to store table metadata, where each record is an instance of the Wikidata *Dataset* class (Q1172284). The rationale for the metadata index is to contain datasets that may not be necessarily materialized as a knowledge graph, but that would be interesting resources to link and extend other datasets. Our metadata schema relies on Wikidata and Schema.org [3], using terms such as title, data download, license, website source, variables, etc. Given a table of interest, WDPlus executes an automated entity linking process using our wikifier and creates an entry in the metadata index. Each entry includes a record of the main distinct Wikidata entities identified in the entity linking process and their labels, which serve to inform the search of related datasets.

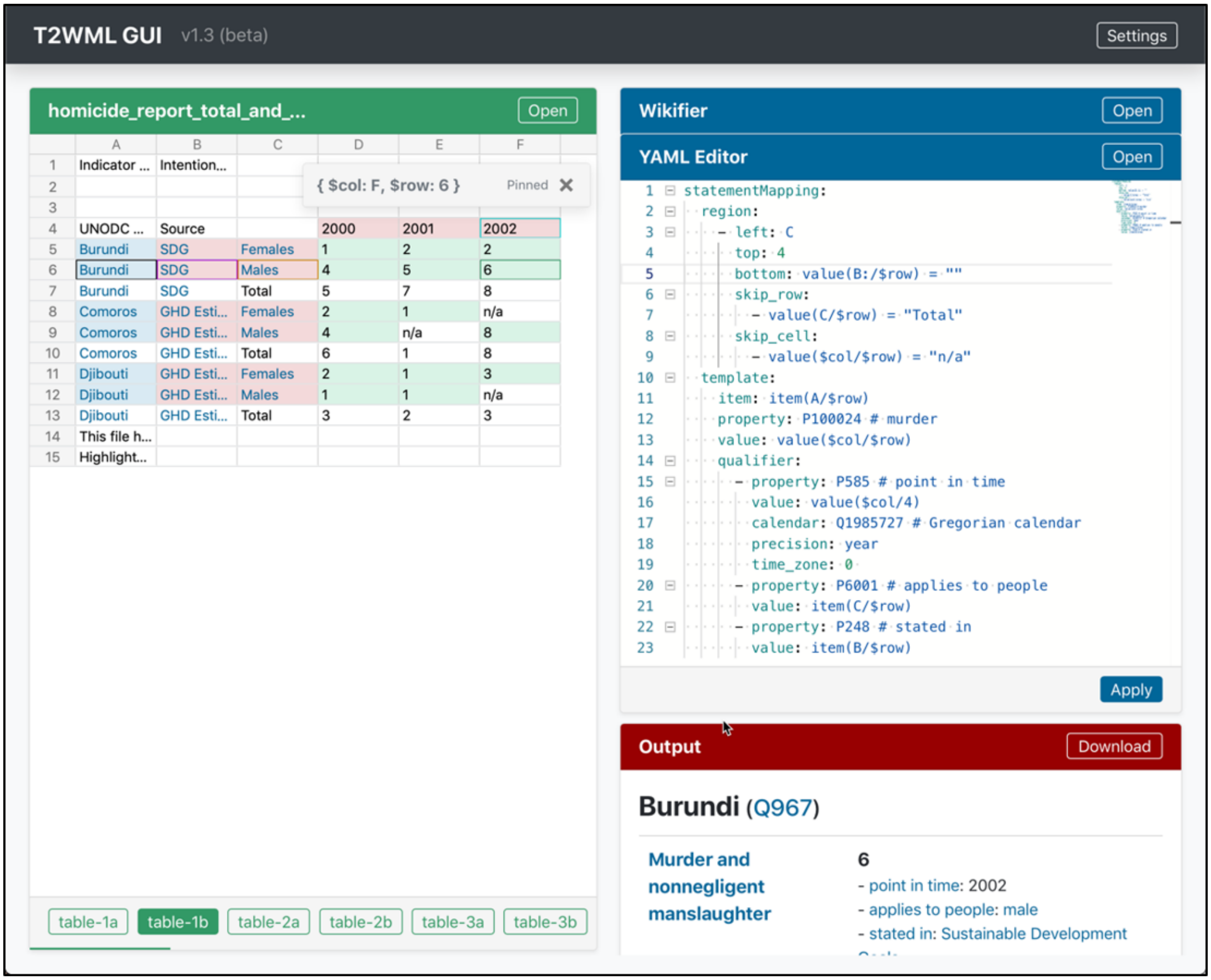
The metadata index connects heterogeneous tables to Wikidata, and therefore we use it for data augmentation. For example, a user with a table containing demographic information by country may be interested in adding climate observations to find correlations between population and temperature. Thus, given a dataset to augment (e.g., CSV with city name and population) and a search term (e.g., temperature), the metadata index returns a ranked selection of datasets that may complement the target dataset with additional metadata. Once a user selects a result dataset, WDPlus automatically adds a new column in the dataset to complete, filling in information for every row from Wikidata or materializing the search results. Missing values are currently not imputed. The API for the metadata index can be found online.[7]

2.3 Augmenting Wikidata with Satellites

Users may want to contribute tabular datasets to WDPlus as an extension of Wikidata entities. This process can be accomplished in two steps. The first one consists on an interactive table understanding step, where the user is presented the GUI shown in Figure 2 with entity linking candidates for all cells included in the target table. The GUI then helps users define a table model, indicating how to map a target table to the Wikidata data model. We collect all qualifiers for each assertion (e.g., source, point in time and space, units, etc.) as these are crucial for ensuring data quality and trust.

[6] https://github.com/usc-isi-i2/t2wml

[7] http://dsbox02.isi.edu:9000/apidocs/



{ $col: F, $row: 6 } Pinned ✕

	A	B	C	D	E	F
1	Indicator ...	Intention...				
2						
3						
4	UNODC ...	Source		2000	2001	2002
5	Burundi	SDG	Females	1	2	2
6	Burundi	SDG	Males	4	5	6
7	Burundi	SDG	Total	5	7	8
8	Comoros	GHD Esti...	Females	2	1	n/a
9	Comoros	GHD Esti...	Males	4	n/a	8
10	Comoros	GHD Esti...	Total	6	1	8
11	Djibouti	GHD Esti...	Females	2	1	3
12	Djibouti	GHD Esti...	Males	1	1	n/a
13	Djibouti	GHD Esti...	Total	3	2	3
14	This file h...					
15	Highlight...					

```yaml
statementMapping:
  region:
    - left: C
      top: 4
      bottom: value(B:/$row) = ""
      skip_row:
        - value(C/$row) = "Total"
      skip_cell:
        - value($col/$row) = "n/a"
  template:
    item: item(A/$row)
    property: P100024 # murder
    value: value($col/$row)
    qualifier:
      - property: P585 # point in time
        value: value($col/4)
        calendar: Q1985727 # Gregorian calendar
        precision: year
        time_zone: 0
      - property: P6001 # applies to people
        value: item(C/$row)
      - property: P248 # stated in
        value: item(B/$row)
```

Burundi (Q967)

Murder and nonnegligent manslaughter 6
- point in time: 2002
- applies to people: male
- stated in: Sustainable Development Goals

Figure 2: Interactive table understanding interface to link datasets to Wikidata records.

The second step generates RDF triples from the table, storing them in a Wikidata satellite using the WDPlus toolkit. WDPlus defines a special range of URIs for each satellite, which are separated in graphs. These graphs may be stored in separate triplestores, allowing to grow Wikidata in a decentralized manner while keeping all related pointers in the same metadata index. Wikidata satellites may define new entities and properties that are not part of Wikidata. However, our early entity linking process minimizes the creation of duplicate entities, ensuring that all satellites are linked together.

A key aspect of WDPlus is that we store all curated table models in our metadata index for each transformed dataset. This helps keeping the provenance of all results in a satellite and may inform the transformation of other tabular data with very similar structure. For example, in the US, demographic data is usually provided for each county as a CSV file. A single table model for one county can be used to transform the county CSVs of the whole country.

3 CONCLUSIONS

This paper introduces WDPlus, a framework that leverages Wikidata to link, search and augment tabular data. While the framework is still under development, our WDPlus prototype integrates heterogeneous data from crime, economic and education domains, illustrating the feasibility of our approach.

REFERENCES

[1] Brank, J., Leban, G., and Grobelnik, M. Annotating documents with relevant wikipedia concepts. In *Proceedings of the Slovenian Conference on Data Mining and Data Warehouses (SiKDD 2017)* (2017).

[2] Brickley, D., Burgess, M., and Noy, N. Google dataset search: Building a search engine for datasets in an open web ecosystem. In *The World Wide Web Conference* (New York, NY, USA, 2019), WWW '19, ACM, pp. 1365–1375.

[3] Guha, R. V., Brickley, D., and Macbeth, S. Schema.org: Evolution of structured data on the web. *Commun. ACM 59*, 2 (Jan. 2016), 44–51.

[4] Szekely, P., Garijo, D., Pujara, J., Bhatia, D., and Wu, J. T2wml: A cell-based language to map tables into wikidata records. In *To appear in Proceedings of the 2019 International Semantic Web Conference* (2019).

[5] Vrandečić, D., and Krötzsch, M. Wikidata: A free collaborative knowledgebase. *Commun. ACM 57*, 10 (Sept. 2014), 78–85.