

Curating Tabular Datasets using Knowledge Graphs

Azanzi Jiomekong^{1,*}, Hippolyte Tapamo¹, Sanju Tiwari², Allard Oelen³ and Sören Auer³

¹Department of Computer Science, University of Yaounde I, Yaounde, Cameroon

²Universidad Autonoma de Tamaulipas, Mexico, India

³TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany

Abstract

Tabular datasets are composed of tables. These tables are used to structure and organize data. However, considering tables individually may make it difficult to identify some information that can be highlighted by linking with other information. To solve this problem, tabular datasets are curated. This curation consists of creating/updating tabular datasets and annotating them using Knowledge Graphs. In this paper, we present a generic workflow that is generally used during the curation process. The example of the creation of Open Research Knowledge Graph comparisons tables is presented to illustrate.

Keywords

Tabular Data, Tabular Dataset curation, Semantic annotation, Knowledge Graph, Open Research Knowledge Graph

1. Introduction

Tables are one of the most used data structures to organize data by software developers, data scientists, business people, etc. Every day, these people have to handle tables that have been extracted from structured, semi-structured and unstructured sources in order to furnish information for decision making. For instance, in a recent work, we extracted tables from scientific papers for several purposes such as Knowledge Graph construction and building of Food Composition Tables datasets [1, 2].


Everyday, data scientists use statistical tools such as RStudio to analyze tabular data that have been extracted from databases of sales, pricing, food composition, etc. and furnish relevant information to decision makers and business people. However, considering tables individually may make it difficult to identify some information that can be identified by linking with other information [3]. That is why tabular data curation can be helpful. Our research on the curation of TSOTSATable dataset [1, 4], Open Research Knowledge Graph [2, 5, 6] and the development of Semantic Tables Annotation systems [7, 8] allowed us to define a generic workflow for the curation of tabular datasets. Thus, the main contribution of this paper is this generic workflow (presented by the Fig 1) that describes the curation process so that it can be helpful to other researchers working on tabular dataset curation. The creation of ORKG comparison tables is

Joint Proceeding of Second International Workshop on Semantic Reasoning and Representation in IoT (SWIoT-2023) and Third International Workshop on Multilingual Semantic Web (MSW-2023), November 13–15, 2023, University of Zaragoza, Zaragoza, Spain

*Fidel Jiomekong <jiofidelus@gmail.com>



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

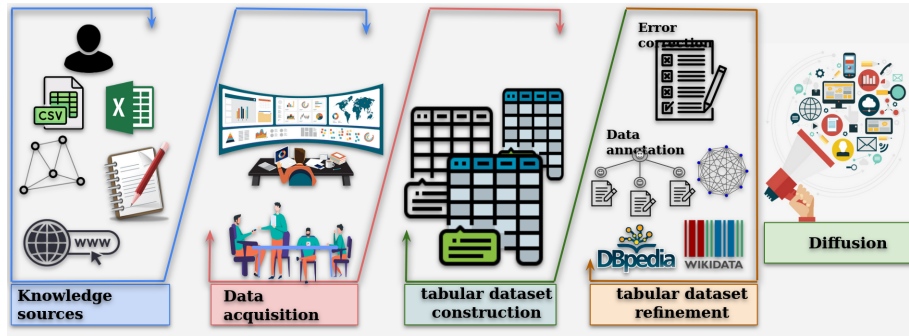


Figure 1: The tabular dataset curation process

provided as a use case. This workflow consists of the identification of data sources from which data will be extracted (see Section 2), data acquisition and organization (see Section 3), and tabular dataset refinement (see Section 4). Given that many automatic systems are generally proposed by researchers during the curation process, we present in Section 5 how these systems are evaluated. The conclusion of this work is presented in Section 7.

2. Identification of data sources

Given the large quantity and the diversity of the data sources, it is essential to identify the data sources that can be used to build tabular datasets. We organized these sources into three dimensions:

- **Humans sources:** Humans are the principal source of knowledge. All the other sources of knowledge are created and updated by humans. Thus, humans can be a valuable source of information. In many cases (bank, sales, etc.) information are directly acquired from people and saved in tables in databases. Thereafter, these tables can be put in CSV format for data analysis purposes. On the other hand, many survey studies use forms to collect data from people and store them in tables. These tables are compiled thereafter and knowledge are extracted from them.
- **Structured sources:** Structured sources such as databases or Knowledge Graphs can be used to build tabular datasets. Existing works show that tabular dataset can be built using Knowledge Graph such as Wikidata and DBpedia [3, 4, 9, 10, 11]. The structure of these data sources make it easy to build automatic tools for obtaining a multitude of tables quickly.
- **Semi-structured sources:** Concerning semi-structured sources, we noted that information can be extracted from tables stored in pdf files in order to build tabular datasets. Jiomekong et al. [1, 12] proposed to identify and extract food composition tables from scientific papers and use these information to build tabular datasets.
- **Unstructured sources:** Unstructured sources such as text can also be used to build tabular datasets. However, it has been reported that knowledge are deeply hidden in the full body text of scientific papers, making it difficult to build automatic tools for their

extraction [6]. Thus, computer-assisted approaches are used [5]. For instance, Open Research Knowledge Graph¹ [6] contains more than 5,000 comparison tables, manually curated by the crowd.

3. Data acquisition and organization

Once the data sources are identified, knowledge should be acquired and used further to build the tabular dataset. Depending on the source of information and the curators, one distinguishes manual acquisition, automatic acquisition and semi-automatic acquisition of data. Manual acquisition consists of acquiring data from a human resource or a data source and building tables with them. Automatic acquisition consists of developing automatic algorithms for extracting information from knowledge sources [13]. In the following points, we present knowledge acquisition using the three dimensions of data sources:

- **Human sources:** The acquisition of information from human sources is always manual because people from which information is coming from should provide these information by talking or writing. Thereafter, the curator(s) will organize the data acquired into tables.
- **Structured sources:** The organization of data from this quality of data sources make it easy to build automation tools for data acquisition. In effect, many Database Management Systems offer features for automatic extraction of tables from the databases and their conversion into CSV format by using a simple query. On the other hand, existing works show how SPARQL endpoint can be used to query knowledge graphs such as Wikidata, DBpedia and use the results of the query to build tabular datasets [10].
- **Semi-structured sources:** Web scraping tools are generally used for information extraction from web pages. Thereafter, simple algorithms can be used to organize these information into tables and build tabular datasets. The structured organization of metadata in scientific papers makes it easy to build automatic tools for metadata extraction and table extraction from review articles [2]. Thus, Open Research Knowledge Graph exploits this structured organization to automatically extract metadata from scientific papers and annotate papers with them.
- **Unstructured sources:** Information extraction from unstructured sources such as the full body text of scientific papers is the most difficult. In this particular case, computer-assisted tools may be used for acquiring scientific knowledge, organizing these scientific knowledge and building tables with them [5].

4. Tabular data refinement

The dataset obtained after the knowledge acquisition step can be seen as a set of isolated tables. However, the structure organization and the content of these tables can make it difficult to achieve the tasks of annotation [3]. On the other hand it has been reported that tabular datasets contain errors such as misspelling, typos, etc. and many problems inherent to the Knowledge Graph or encountered during the matching process [14]. Tabular dataset refinement aims at

¹<https://orkg.org/>

solving these problems and to complete the dataset with semantic annotation. In the following paragraphs, we present firstly the problems that can be found in tabular datasets. Thereafter, we present how the dataset can be completed with semantic annotation.

4.1. Refinement problems

The refinement problems are the problems that may be encountered during the tabular dataset refinement. We are currently documenting these problems using Open Research Knowledge Graph [14]. These problems can be categorized using the following dimensions:

- **Structural problems:** Structural problems consist of: (1) formatting problems such as merged cells, empty lines before the header, mismatched number of headers, missing header, cells with different types of data, etc. and (2) text formatting problems such as date format or number format, insertion of new lines and special characters in text, optical character errors such as the replacement of '0' and 'o', '1' and 'I', etc.
- **Misspelling problems:** These problems come when the content of cells contain words that are wrongly written. For instance, when there is a missing letter in a word, writing a single letter in a word when double letters are to be written and vice versa, etc. These problems may lead to confusion and wrong interpretation of information.
- **Erroneous numbers:** These errors are due to the errors when reporting a number. For instance, putting the age 120 to a person who is 12 years old.
- **Semantic problems:** The semantic problems come when the mention in the table is syntactically different from the label of the entity in the Knowledge Graph due to the use of acronyms, aliases, etc. For example, to name "Cameroon" (English spelling), we can have "Cameroun" (French spelling) or "Kamerun" (German spelling) in the table. On the other hand, the same entity in the table can correspond to many entities in the KG, leading to the problem of ambiguity.
- **Other problems:** Other problems consists of (1) NIL-mentions: the entity in the table does not have a correspondence in the KG, (2) File size: in many cases, files are too large, caused by too many records, too many columns and too many rows in tables. Thus, the computer used to import this kind of data should have enough memory, (3) Data heterogeneity: date values expressed in different formats in the same table, (4) Irrelevant tables: this problem arises when the curation consists of domain tables. For instance, during the curation of the TSOTSA Table dataset, we found many tables that were not related to the domain of Food Science and Nutrition [12].

4.2. Tabular dataset annotation using Knowledge Graphs

Annotating tabular datasets consists of assigning semantic tags from a Knowledge Graph to the elements in the tables. Two types of annotations can be considered [15]:

- **Structural annotation:** This consists of completing the tabular dataset with structural information such as table headers, subject column, etc.
- **Semantic annotation:** This consists of mapping the elements in the tabular dataset to the entities in the KG. Recently, SemTab² introduced new terminologies for tabular data

²<https://sem-tab-challenge.github.io/2023/>

annotation, splitting the table annotation into three sub-tasks which are: (1) Cell Entity Annotation (CEA): consisting of matching the content of a cell in the table to an entity in the KG, (2) Column Type Annotation (CTA): consisting of matching the the column type to a class in the KG, (3) Column Property Annotation (CPA): consisting of assigning a KG property to the relationship between two columns, and Table Topic Detection (TTD): assigning a KG type to a table.

5. Evaluation of curation systems

The evaluation of tabular dataset curation can be structured into two dimensions: the data acquisition dimension and the tabular data refinement dimension:

- **Data acquisition evaluation:** Given that automatic methods are generally used to acquire data from data sources, these methods should be evaluated to determine their performance on the quantity of knowledge that is extracted compared to the one that was supposed to be extracted.
- **Dataset refinement evaluation:** The methods for automatic refinement take a KG and a tabular dataset as input and provide as output annotated tables. These methods should also be evaluated to determine how accurate they are to solve the curation tasks. We are currently documenting the different evaluation metrics used for this purpose [16].

Whatever the evaluation, they use gold standard and evaluation metrics. Human opinion can also be demanded during a retrospective evaluation. Human evaluation is good because experts are able to determine if everything is extracted and determine if the right annotation was assigned to a tabular data element. However, given the size of many datasets, human annotation is tedious. Thus, one relies on automatic annotation. The main evaluation metrics used are Accuracy, Recall, Precision, F-score, Average Recall, Average Precision, Average F-score [3]. In addition to the system evaluation, the computational performance of the system can be evaluated to determine the resource consumption during its execution. For instance, runtime measurements are used to measure the amount of time that the annotation algorithm can take to perform its task, memory consumption measures the amount of memory that the annotation algorithm consumes throughout its execution.

6. Use case: creation of ORKG comparisons tables

ORKG is an open research infrastructure designed to acquire, publish and process structured scholarly knowledge published in the scholarly literature [2, 6]. It uses ORKG comparison tables to compare research contributions dealing with the same research problem. Currently, ORKG contains more than 1300 comparisons tables ³.

To create ORKG comparison tables, the first step consists of providing the semantic description of papers used. From these papers, key-insights are identified, extracted and organized into ORKG research contributions. The latter address research problems and are described using key

³<https://www.orkg.org/orkg/stats>

Properties	ADOG - Annotating Data with Ontologies and Graphs description - 2019	MTab: Matching Tabular Data to Knowledge Graph using Probability Models description - 2019	Manistable: an Automatic Approach for the Semantic Table Interpretation description - 2019	Entity Linking to Knowledge Graphs to Infer Column Types and Properties description - 2019	CSV2KG: Transforming Tabular Data into Semantic Knowledge description - 2019	DAGOBAH: An End-to-End Context-Free Tabular Data Semantic Annotation System description embedding - 2019
datasets	Automatically Generated	Automatically Generated	Automatically Generated	Automatically Generated	Automatically Generated	Automatically Generated
evaluation	Average Hierarchical Score Average Perfec Score Precision Recall	Average Hierarchical Score F1 Score	Average Hierarchical Score Average Perfec Score F1 Score Precision Recall	Average Hierarchical Score Average Perfec Score F1 Score Precision Recall	Average Hierarchical Score Average Perfec Score F1 Score Precision Recall	Average Hierarchical Score Average Perfec Score F1 Score Precision
experimental tool	ArangoDB Elasticsearch	DBpedia endpoint DBpedia lookup Fast Text fCy Spacy Wikidata lookup	Django Docker MongoDB Python programming language Web interface	Elasticsearch Wikidata API	DBpedia lookup API DBpedia spotlight API Python programming language	DBpedia API DWT extractor Elasticsearch Wikidata API Wikidata crisis search engine Wikimedia API
future work	Ameliorate the Column Property Annotation Integrate multi-source systems	Classify table types before matching Correctly recognize table headers Improve the completeness and correctness of Knowledge Graph Perform holistic matching	Allow easier third party integration via remote API calls Change the workflow to better implement the methods for the CEA and CPA tasks	Study sensitivity of the approaches to characteristics of the datasets	Take into account other data sources or embedded values of occurring triples	Combine Wikimedia and Wikidata in a joint embedding space Implement a full vectorial approach Testing other clustering algorithms Use Faostext-like embedding
has links	https://github.com/danleopolvaira/awc-annotation-challenge	https://github.com/Phucy/MTAB	https://bitbucket.org/disco_unim/manistable-tool.py	https://github.com/IBCHServices/CSV2KG	https://github.com/IBCHServices/CSV2KG	
has process	Ontology Graph - Native links - Discovered relatedness edges		Data Preparation - Column Analysis - Entity Linking - Predicate Annotation - Concept and Datatype Annotation	Candidate generation - Candidate selection Candidate generation - Feature selection - Candidate selection	Cell Lookup - Infer columns - Infer properties - Header annotation - Infer other cells - Infer columns	Tables pre-processing - Embedding enrichment and Lookup - Candidates Clustering and Scoring
keywords	DBpedia Knowledge Graph Ontologies			Knowledge Graph Semantic Web Table Understanding	Entity Recognition Property Recognition Semantic Annotation	DAGOBAH Embedding Entity Linking

Figure 2: An example of comparison table

insights including materials, methods, implementation, results, etc. To this end, key-insights extracted are matched to ORKG entities. When these entities do not exist, new ones are created. Thereafter, several research contributions can be compared by creating an ORKG comparison table. This comparison table can be published with a DOI and exported in different formats. It can be improved by other researchers by correcting errors/mistakes or updated with more research contributions.

The Fig. 2 presents an excerpt of a table comparing several systems for tabular data annotation⁴. The left panel presents the different properties used for the comparison. The right panel contains in the header, scientific papers that are compared and in the cells key-insights that were extracted from these papers. From this table, you can find for instance that the dataset used for evaluating these systems is automatically generated; several evaluation metrics such as average hierarchical score, F1, recall and precision are used; several tools such as Wikidata and DBpedia endpoint, DBpedia and Wikidata API, elastic search, etc. are used during experimentation. It should be noted that each element of the table is identified using an URI.

To create this comparison table, we used as data source the SemTab paper repository⁵. This repository contains all the papers published by SemTab@ISWC. Given the unstructured nature of these papers, we manually extract scientific knowledge from them and use ORKG as a computer assistant tool to organize these knowledge. Once extracted and saved in ORKG, we used the ORKG comparison table wizard to create this table. This table was shared with authors for its evaluation.

7. Summary and conclusion

This paper presents a workflow for curating tabular datasets using Knowledge Graphs. This workflow consists of the identification of knowledge sources, data acquisition and organization, and tabular dataset refinement. During the curation process several problems such as structural, semantic, misspelling, may be encountered. Once these problems are solved the tabular data elements are matched to the KG entities and classes. Given that automatic systems are generally used for these tasks, several evaluation metrics such as recall, precision, F-measure, are used to evaluate these systems. We illustrated with the case of the creation of an ORKG comparison table, comparing several systems for tabular datasets annotation.

Given that Large Language Models such as ChatGPT, Llama 2, etc. are making new waves in the field of natural language processing and artificial intelligence, we are currently exploring its capabilities for the curation of tabular datasets.

References

- [1] A. Jiomekong, C. Etoga, B. Foko, V. Tsague, M. Folefac, S. Kana, M. M. Sow, G. Camara, A large scale corpus of food composition tables, in: SemTabISWC, 2022.
- [2] A. Oelen, M. Stocker, S. Auer, Creating a scholarly knowledge graph from survey article tables, in: Digital Libraries at Times of Massive Societal Transition, Springer International Publishing, 2020, pp. 373–389. URL: https://doi.org/10.1007/978-3-030-64452-9_35. doi:10.1007/978-3-030-64452-9_35.
- [3] E. Jiménez-Ruiz, O. Hassanzadeh, V. Efthymiou, J. Chen, K. Srinivas, Semtab 2019: Resources to benchmark tabular data to knowledge graph matching systems, in: A. Harth, S. Kirrane, A.-C. Ngonga Ngomo, H. Paulheim, A. Rula, A. L. Gentile, P. Haase, M. Cochez (Eds.), The Semantic Web, Springer International Publishing, Cham, 2020, pp. 514–530.

⁴<https://orkg.org/comparison/R642234/>

⁵<https://www.cs.ox.ac.uk/isg/challenges/sem-tab/>

- [4] A. Jiomekong, U. Melie, Tsotsatable dataset: a dataset of food and its composition, 2023. URL: <https://doi.org/10.5281/zenodo.8169063>. doi:10.5281/zenodo.8169063.
- [5] A. Jiomekong, S. Tiwari, An approach based on open research knowledge graph for knowledge acquisition from scientific papers, SSRN (2023). doi:<http://dx.doi.org/10.2139/ssrn.4333481>.
- [6] S. Auer, A. Oelen, M. Haris, M. Stocker, J. D'Souza, K. Eddine Farfar, L. Vogt, M. Prinz, V. Wiens, M. Y. Jaradeh, Improving access to scientific literature with knowledge graphs, BIBLIOTHEK – Forschung und Praxis (2020). doi:<http://dx.doi.org/10.18452/22049>.
- [7] A. Jiomekong, B. Foko, Towards an approach based on knowledge graph refinement for tabular data to knowledge graph matching, 2022, pp. 111–122. URL: <https://ceur-ws.org/Vol-3320/paper12.pdf>.
- [8] B. Foko, A. Jiomekong, T. Hippolyte, T. Sanju, Exploring naive bayes classifiers for tabular data to knowledge graph matching, 2023.
- [9] M. Hulsebos, Çağatay Demiralp, P. Demiralp, Gittables benchmark - column type detection, 2021. URL: <https://doi.org/10.5281/zenodo.5706316>. doi:10.5281/zenodo.5706316.
- [10] V. Cutrona, F. Bianchi, E. Jiménez-Ruiz, M. Palmonari, Tough tables: Carefully evaluating entity linking for tabular data, in: J. Z. Pan, V. Tamma, C. d'Amato, K. Janowicz, B. Fu, A. Polleres, O. Seneviratne, L. Kagal (Eds.), The Semantic Web – ISWC 2020, Springer International Publishing, Cham, 2020, pp. 328–343.
- [11] N. Abdelmageed, S. Schindler, B. König-Ries, BiodivTab: A Tabular Benchmark based on Biodiversity Research Data, in: SemTab@ISWC, submitted, 2021.
- [12] A. Jiomekong, M. Uriel, T. Hippolyte, C. Gaoussou, Semantic annotation of tsotsatable dataset, 2023.
- [13] A. Jiomekong, G. Camara, M. Tchunte, Extracting ontological knowledge from java source code using hidden markov models, Open Computer Science 9 (2019) 181–199.
- [14] A. Jiomekong, Problems encountered by semantic table annotations during semtab 2019 challenge, 2023. <https://orkg.org/comparison/R600534/>.
- [15] P. Nguyen, I. Yamada, N. Kertkeidkachorn, R. Ichise, H. Takeda, Semtab 2021: Tabular data annotation with mtab tool., in: SemTabISWC, 2021, pp. 92–101.
- [16] A. Jiomekong, Evaluation metrics used during tabular data to knowledge graph matching challenges since 2020, 2023. URL: <https://orkg.org/comparison/R604322/>. doi:10.48366/R604322.