



# Towards Customizable Chart Visualizations of Tabular Data Using Knowledge Graphs

Vitalis Wiens<sup>1,2(✉)</sup>, Markus Stocker<sup>1</sup>, and Sören Auer<sup>1,2</sup>

<sup>1</sup> TIB Leibniz Information Centre for Science and Technology, Hanover, Germany  
`markus.stocker@tib.eu`

<sup>2</sup> L3S Research Center, Leibniz University of Hannover, Hanover, Germany  
`{wiens,auer}@l3s.de`

**Abstract.** Scientific articles are typically published as PDF documents, thus rendering the extraction and analysis of results a cumbersome, error-prone, and often manual effort. New initiatives, such as ORKG, focus on transforming the content and results of scientific articles into structured, machine-readable representations using Semantic Web technologies. In this article, we focus on tabular data of scientific articles, which provide an organized and compressed representation of information. However, chart visualizations can additionally facilitate their comprehension. We present an approach that employs a human-in-the-loop paradigm during the data acquisition phase to define additional semantics for tabular data. The additional semantics guide the creation of chart visualizations for meaningful representations of tabular data. Our approach organizes tabular data into different information groups which are analyzed for the selection of suitable visualizations. The set of suitable visualizations serves as a user-driven selection of visual representations. Additionally, customization for visual representations provides the means for facilitating the understanding and sense-making of information.

**Keywords:** Scholarly communication · Knowledge graphs · Customizable visualizations · Information visualization

## 1 Introduction

Scholarly communication has not changed in its core during the last centuries. Research articles are typically distributed as PDF documents, and the amount of publications increases continuously every year [8]. As a consequence, searching, understanding, and organizing information becomes a burden. Finding and reviewing the literature is tying up cognitive capacity [1], and consumes time which consequently reduces the time available for original research.

The purpose of scientific articles is to inform and share findings. As a means for scholarly communication, the information is presented in documents using text, figures, and tables. While the descriptive text provides detailed insights,

figures and tables serve as a visual, structured, and compressed representation of information. However, this information is buried in PDF representations [10].

The current developments in scholarly communication exploit Semantic Web technologies. These advancements transform the scholarly communication from document-based to knowledge-based information systems employing structured, interlinked, and semantically rich knowledge graphs [1]. In contrast to other Digital Library applications that organize primarily bibliographic metadata, the Open Research Knowledge Graph [7] (ORKG)<sup>1</sup> captures the content of research articles (e.g., research problem, materials, methods, and results).

Generally, the view on the information in scientific articles becomes static and frozen following publication. Thus, further analysis of presented information continues to be a manual effort for readers. Knowledge-based representations provide machine-readable access to information, which serves as input for various applications, including those addressing its presentation to humans. Therefore, it is beneficial to extract and transform the information of scientific articles into structured and machine-readable representations. However, due to its design for machine-interoperability and processing of information, the cognitive load for humans increases with growing size and complexity of such data structures. Visualizations serve a purpose of addressing specific information needs for the data at hand and human's ability to understand complex data through visual representations, "*a picture is worth a thousand words*" [13]. Following the information seeking mantra (overview, zooming/filtering, and details on demand) [15], we argue that user-driven approach for the generation of visualizations and their customization can further facilitate the sense-making of information.

In this article, we focus on the results of scientific articles in the form of tables. Tables provide an *organized* and compressed depiction of information. Various works, such as the recent work of Vu et al. [16], address the transformation of tabular data into knowledge-based representations. In contrast, the objective of our approach is to extract such information and provide customizable and meaningful chart visualizations of tabular data from knowledge graphs. In particular, we address the following challenges:

- i) What minimal information structure is required in a knowledge graph to obtain visual representations of tabular data.
- ii) How to analyze this structured information for visualization generation.

Our approach employs a human-in-the-loop technique to transform tabular data into knowledge graph representations with additional semantics. These additional semantics serve as the foundation for obtaining views of the knowledge graph that feed into various data visualization. Using the additional semantics, our approach recreates tables from knowledge graphs and enables the analysis of their content for the creation of customizable chart visualizations.

The remainder of this article is structured as follows. Section 2 summarizes related work, and Sect. 3 describes the proposed approach. Section 4 discusses the

---

<sup>1</sup> <https://orkg.org>.

limitations and implications for additional use cases. Finally, Sect. 5 concludes with an outlook on future work.

## 2 Related Work

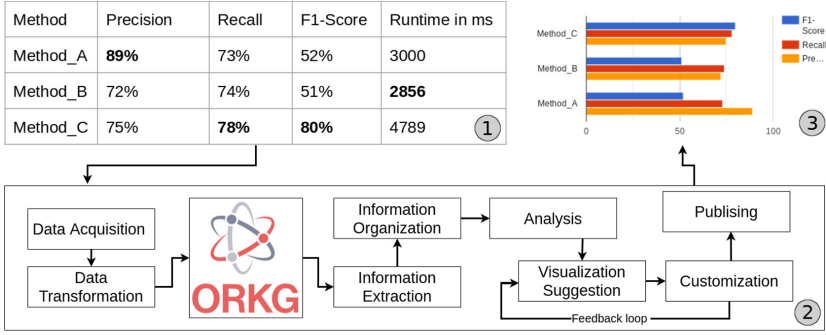
The related work can be categorised into two groups: **a)** transformation of tables into knowledge graph representations; **b)** visualization of knowledge graphs. Addressing the former, the recent work of Vu et al. [16] represents the transformation process in the form of a mapping language (D-REPR). Heterogeneous datasets, such as tables in CSV or JSON formats, with different layouts are described in a model that defines components for the transformation into RDF. These components describe the dataset resource, its attributes and how data alignment is realized. A semantic model component describes how the data is transformed into RDF. Other approaches, such as XLWrap [9], focus on the transformation of spreadsheets into RDF. R2RML [3] is a W3C recommendation that addresses the mapping of relational databases to RDF. However, relational databases can be seen as tables, and therefore, R2RML techniques are also applied to transform tabular data into Semantic Web representations such as RDF. Due to the flexible nature of tables, the challenge of transforming tables into Semantic Web representations typically results in transformation models that are specifically tailored for individual datasets. Similarly, our approach is currently tailored for the representation of row-based-entries for one dimensional values.

Several definitions of knowledge graphs and its features exist; however, we lack a unified definition [5]. Ehrlinger and Wöß [5] argue additionally that “an ontology does not differ from a knowledge base”, meaning that visualization methods for ontologies are also applicable for the visualization of the structure of knowledge graphs. According to a recent survey [4], most methods and tools visualize the content of ontologies using two-dimensional graph-based representations in the form of node-link diagrams.

Approaches, such as RelFinder [6] or the Neo4j graph visualization [11] address the visualization of knowledge graphs based on their structure (i.e., nodes and links). While node-link diagrams are well suited to represent the data structure of knowledge graphs, in some contexts, such as the visualization of tables, the structural representation will not facilitate the comprehension of information. Knowledge graphs have different structures and also contain additional information that does not serve the purpose for information interpretation (e.g., URIs or class assertions). Therefore, in order to generate suitable visualizations, the context and the semantics of the retrieved entries from a knowledge graph need to be incorporated and processed properly for the reconstruction of a table.

The Wikidata Query Service<sup>2</sup> is an application that is closely related to our approach. The system leverages SPARQL and presents results using different visualization methods. It provides a selection of visual representations (e.g.,

<sup>2</sup> <https://query.wikidata.org/>.



**Fig. 1. Overview:** (1) A table for artificial results of Precision, Recall, F1-Score, and Runtime. (2) Processing pipeline. (3) Resulting visual representation.

Table, Tree, and Timeline) for the resulting data. While the Wikidata Query Service provides a generic solution for the customizable visualization of knowledge graphs, we present an approach that incorporates additional semantics and guides the visualization generation process that is designed for the visual representation of tabular data in the form of customizable charts.

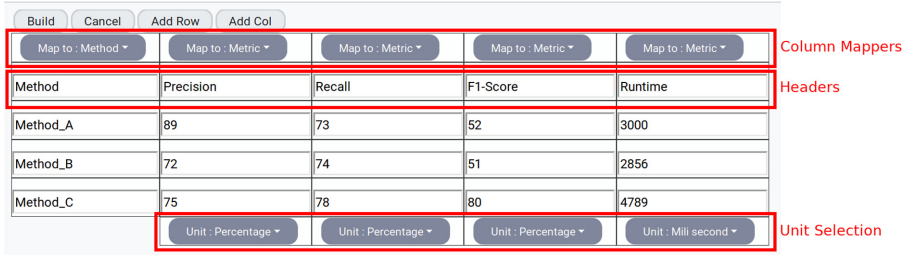
### 3 Approach

Our approach is motivated and aligned with the objectives of the Open Research Knowledge Graph (ORKG) [7], i.e., the structured representation of contributions in scientific articles and the facilitation of information perception and its sense-making. However, our approach addresses the customizable visualization for tabular data that originates from knowledge graphs. As a running example, we use an imaginary table summarizing the performance of different methods, which is common in Computer Science articles (see Fig. 1).

#### 3.1 Data Acquisition and Transformation

At first, the data acquisition phase transforms the table into a knowledge graph representation and ensures the correct assignment of additional semantics using a human-in-the-loop approach. Knowledge graph structures typically reflect a triple-based representation  $\langle s \ p \ o \rangle$ , where the subject  $s$  and the object  $o$  are interlinked by the predicate  $p$ . Our approach augments tabular data with additional semantics during the data acquisition phase, preserving the *context* which allows more efficiently to create further analysis and visualizations from this structured data. Our transformation model builds upon the following heuristics:

- i) The cell entries of the first column provide the subjects; in our example, these are the methods. Thus, cell values of a row are bound to the method. Related to this, our transformation model is also row-based.



**Fig. 2.** Widget for the tabular data transformation process eases the data input process and appends additional semantics to cell values.

- ii) Other columns provide values for measurements of a metric. Thus, our transformation model adds to the cell value two additional attributes, namely the metric and the unit of the cell value. The header values of the columns determine the metric, while a human-in-the-loop approach assigns the units for the corresponding columns.

As illustrated in Fig. 2, a simple tabular input widget eases the process for the user to enter the data and also ensures the correct assignment of additional semantics for the table.

While, in general, the particular value is of interest, it is also necessary to incorporate the *context*. The numerical value “89” is just a data point lacking any meaning. Adding metric and unit to this value captures more *context*. This context enables to describe the cell value as: The value “89” describes *Precision*, it has the unit percentage, and it refers to a method (Method\_A).

### 3.2 Information Extraction and Organization

The reconstruction of a table requires the information about the transformation model and its structural representation. This information is obtained from the data acquisition phase. However, due to the unknown order of returned triples, the ordering of rows and columns can change. Nevertheless, we obtain a reconstructed table with sufficient *context* for our example. Furthermore, the reconstructed table becomes interactive through corresponding implementations, e.g., sorting the columns ascending or descending based on their values. As illustrated in Fig. 3, this straight forth and back transformations provide already interactions with tabular data and another view on the information.

Method	Precision	Recall	F1-Score	Runtime in ms
Method_A	<b>89%</b>	73%	52%	3000
Method_B	72%	74%	51%	<b>2856</b>
Method_C	75%	<b>78%</b>	<b>80%</b>	4789

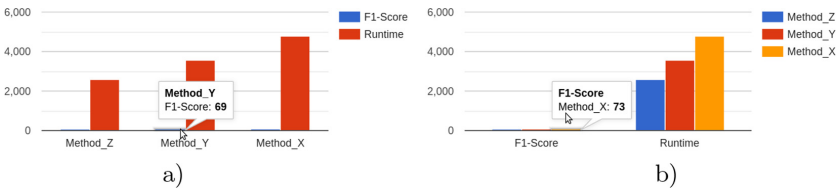
Table View					
	Method	Runtime	F1-Score	Recall	Precision
1	Method_C	4,789	80	78	75
2	Method_A	3,000	52	73	89
3	Method_B	2,856	51	74	72

**Fig. 3.** Illustration of the original table and the reconstructed table from a knowledge graph. *Note:* The ordering of the columns is not preserved.

The reconstructed table serves as input data for chart visualizations. However, we argue that the *context* is viable for the creation of suitable chart visualizations. In this article, we define the context of a cell value as follows:

**Definition 1.**  $Context(value(i, j)) = (RowLabel(0, j), Unit(i), Metric(i))$   
Where  $i \geq 1$ , is the column index and  $j$  the row index.

The *RowLabel* refers to the entries from the first column that are used as subject anchors in the knowledge graph representation. The *Unit* is provided by the user, and the *Metric* is obtained from the header values of the corresponding column. Data units are a crucial factor in creating meaningful chart visualizations. We argue that metrics with the same units provide *reasonable* candidates for grouping information and avoid false interpretations when visualized in the same chart, i.e., significant differences in data ranges shift the attention focus to the visual elements that have a higher presence in the chart, see Fig. 4.



**Fig. 4.** Column chart visualization indicating the possible false first impression through unrelated units and large differences in the data ranges.

The semantics of *Units* provide the means to create information groups by clustering columns, i.e., the extraction of sub-tables through the matching of compatible units. These groups reflect information that relates (or co-relates) to a certain extend. The semantics of *Metrics* provide the means to guide the selection of suitable chart visualization types. In particular, it is the definition of compatible chart types for individual metrics.

**Units:** The additional semantics of *Units* provide means to align the cell values to a uniform representation for a particular unit. These semantics serve as alignment definitions between them. For example, percentage and per-mil are easily brought into correspondence using an alignment factor of 10, or milliseconds are transformed to seconds using an alignment factor of 1000. The semantics for unit alignment enable the approach to detect compatible units and bring them into correspondence for clustering related (or co-related) information.

**Metrics:** The semantics of metrics provide additional criteria for building information groups (i.e., the subdivision of sub-tables). As mentioned before, units provide *reasonable* candidates for clustering related (or co-related) information into groups. However, identical units are used in different metrics. For example, percentage can refer to performance measurements in information retrieval

tasks or statistical distributions. The definition of compatible metrics refines the grouping of related information and determines which columns serve as input.

Metrics provide additional value validation mechanisms. In particular, they define a data range. For example, the metric *Precision* has a range of  $[0, \dots, 100]$ , or *Runtime* cannot be expressed as negative values. This value range restrictions define a validation mechanism for transformation models that populate knowledge graphs with tabular data. However, the value range restrictions for the myriad of measurement factors need to be defined individually for each metric.

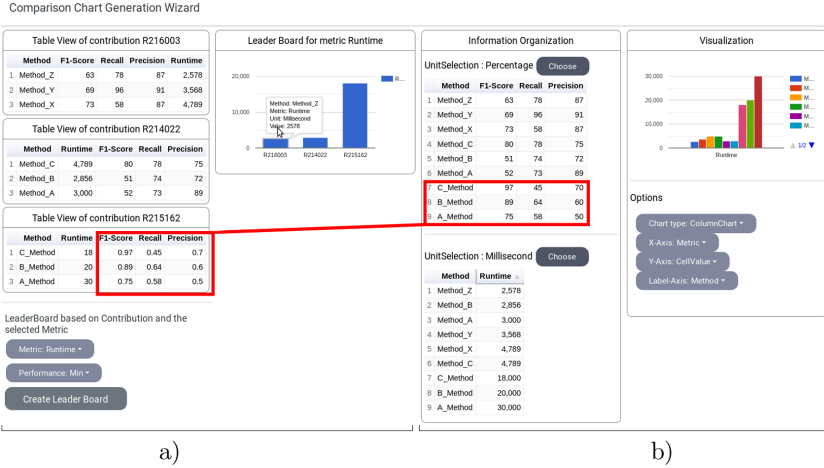
### 3.3 Customizable Visualization Generation

The analysis of the additional semantics performs the most of the heavy lifting. However, the dimensions of the table also pose restrictions on the selection of suitable chart visualizations. For example, spider-charts require at least 3 dimensions in order to span an area for a value. While this criteria is met when the number of rows is adequate (e.g., visualizing *Precision* with the corresponding methods as axial dimension), this representation becomes invalid if the axis mapping is flipped and the dimensional criteria is not met (e.g., only *Precision* serves as the axial dimension). This simple example indicates that the selection for axis mapping is also crucial for the visualization suggestion. As illustrated in Fig. 1, this refers to the feedback loop for the visualization suggestion.

## 4 Discussion

Our approach builds upon the semantics and the structure of the tabular data representation in a knowledge graph. Thus, it is currently limited to the chosen transformation model. Furthermore, the approach addresses the one dimensional representation of columns and rows. In our approach, the first column of the table refers to unsorted entries. However, when dealing with order dependent entries, such as time series or physical distances, the position on the axis (sorting) is significant for the information comprehension. Currently, our approach does not address order dependent entries in the first column.

The approach has been described in the context of tabular data visualizations within a single paper. However, tables are frequently used in scientific articles of various type. Incorporating additional semantics enables new opportunities for analysis of information across papers, too. In particular, through the additional semantics of units and metrics the information distributed across several tables (in different articles) can be organized for further analysis. Figure 5 show-cases the visualization generation of tables across different articles.



**Fig. 5.** Prototype for chart visualization using the comparison feature of ORKG: a) The individual tables, selection options for leader-board generation and a leader-board visualization; b) Information organization for merged tables and the resulting column chart. The value representation transformation is indicated in red. (Color figure online)

## 5 Conclusion

In this article, we have presented an approach for customizable chart visualizations of tabular data using knowledge graphs. The approach builds on additional semantics that are added during the data acquisition process. Using these semantics, tables are reconstructed and organized in information groups, i.e., sub-tables based on metrics and units. The semantics of *Metrics* select suitable visualization from a large space of all chart types. Customizations are enabled through chart type selection and axis mappings. Using the paper comparison feature of ORKG [12], the approach realizes advanced use cases, such as the visualization of information distributed among tables in multiple articles and leader-boards.

The *context* plays an important role in extracting tabular data from knowledge graphs and the creation of visual representations. Our approach creates the context using the a-priori known data structure and its additional semantics. Future work will address the extension for the definition of additional semantics related to order dependent entries for the first column. The semantics of *Metrics* define the interplay among them and which chart visualizations are suitable. Thus, future work will address the many definitions of metrics. Additionally, we plan to investigate the alignment to existing vocabularies related to units [14] and the RDF Data Cube Vocabulary [2] in order to increase the flexibility and robustness of the approach. Furthermore, we argue that pattern matching and sub-graph identification will enable the realization of semi-automated generation for context items that guide the information organization and the analysis, enabling the chart visualization of non-tabular data from knowledge graphs.



In conclusion, we argue that the approach introducing additional semantics and further rules will foster the creation of suitable and custom visual representations for tabular data using knowledge graphs and that it facilitates comprehension through different perspectives on the information in tables.

**Acknowledgments.** This work is co-funded by the European Research Council project ScienceGRAPH (Grant agreement #819536). Additionally, we would like to thank our colleagues Mohamad Yaser Jaradeh and Kheir Eddine for valuable discussions and suggestions.

## References

1. Auer, S., Kovtun, V., Prinz, M., Kasprzik, A., Stocker, M., Vidal, M.E.: Towards a knowledge graph for science. In: Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, pp. 1–6 (2018)
2. Cyganiak, R., Reynolds, D.: The RDF data cube vocabulary (2014). <https://www.w3.org/TR/vocab-data-cube/>
3. Das, S., Sundara, S., Cyganiak, R.: R2RML: RDB to RDF mapping language (2012). <https://www.w3.org/TR/r2rml/>
4. Dudáš, M., Lohmann, S., Svátek, V., Pavlov, D.: Ontology visualization methods and tools: a survey of the state of the art. *Knowl. Eng. Rev.* **33** (2018)
5. Ehlringer, L., Wöß, W.: Towards a definition of knowledge graphs. *SEMANTICS (Posters Demos SuCESS)* **48**, 1–4 (2016)
6. Heim, P., Hellmann, S., Lehmann, J., Lohmann, S., Stegemann, T.: RelFinder: revealing relationships in RDF knowledge bases. In: Chua, T.-S., Kompatsiaris, Y., Merialdo, B., Haas, W., Thallinger, G., Bailer, W. (eds.) *SAMT 2009. LNCS*, vol. 5887, pp. 182–187. Springer, Heidelberg (2009). [https://doi.org/10.1007/978-3-642-10543-2\\_21](https://doi.org/10.1007/978-3-642-10543-2_21)
7. Jaradeh, M.Y., et al.: Open research knowledge graph: next generation infrastructure for semantic scholarly knowledge. In: Proceedings of the 10th International Conference on Knowledge Capture, K-CAP 2019, New York, NY, USA, pp. 243–246. Association for Computing Machinery (2019)
8. Johnson, R., Watkinson, A., Mabe, M.: *The STM Report. An Overview of Scientific and Scholarly Publishing*, 5th edn. (2018)
9. Langegger, A., Wöß, W.: XLWrap – querying and integrating arbitrary spreadsheets with SPARQL. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) *ISWC 2009. LNCS*, vol. 5823, pp. 359–374. Springer, Heidelberg (2009). [https://doi.org/10.1007/978-3-642-04930-9\\_23](https://doi.org/10.1007/978-3-642-04930-9_23)
10. Mons, B.: Which gene did you mean? *BMC Bioinform.* **6**, 142 (2005)
11. Neo4j. Neo4j graph visualization. <https://neo4j.com/developer/graph-visualization/>. Accessed Mar 2020
12. Oelen, A., Jaradeh, M.Y., Farfar, K.E., Stocker, M., Auer, S.: Comparing research contributions in a scholarly knowledge graph. In: Proceedings of the Third International Workshop on Capturing Scientific Knowledge Co-located with the 10th International Conference on Knowledge Capture (K-CAP 2019), Marina del Rey, California, 19 November 2019, vol. 2526. *CEUR Workshop Proceedings*, pp. 21–26. CEUR-WS.org (2019)

13. Peña, O., Aguilera, U., López-de Ipiña, D.: Linked open data visualization revisited: a survey. *Semant. Web J.* (2014)
14. Rijgersberg, H., van Assem, M., Top, J.: Ontology of units of measure and related concepts. *Semant. Web* 4(1), 3–13 (2013)
15. Shneiderman, B.: The eyes have it: a task by data type taxonomy for information visualizations. In: *Proceedings of the 1996 IEEE Symposium on Visual Languages*, Boulder, Colorado, USA, 3–6 September 1996, pp. 336–343 (1996)
16. Vu, B., Pujara, J., Knoblock, C.A.: D-REPR: a language for describing and mapping diversely-structured data sources to RDF. In: *Proceedings of the 10th International Conference on Knowledge Capture*, pp. 189–196 (2019)