Interactive Enrichment of Tabular Data with SemTUI*

Flavio De Paoli[1], Roberto Avogadro[2], Marco Ripamonti[1] and Matteo Palmonari[1]

[1] University of Milano-Bicocca, Milan, Italy
[2] SINTEF AS, Oslo, Norway

Abstract

In this work, we demonstrate the usage of SemTUI, an open-source tool to explore and define semantic-based data enrichment operations. The tool is designed to define sequences of data enrichment operations generalizing a *link & extend* paradigm by integrating external services for end-to-end tabular data annotation, data reconciliation, and data extension. These services are operated through a graphical user interface that helps users explore alternatives and revise the results.

1. SemTUI: a Framework to Link & Extend

Tabular data are extensively used in data analysis and AI-driven projects for model training in both industry and science. Such data often necessitate significant preparation [1], including the challenging task of data enrichment, which involves adding more content-related columns from various sources [2, 3, 4]. This process is particularly difficult for users, data scientists, and data engineers who may have limited knowledge of third-party data used for enrichment [5].

Semantics can play a pivotal role in supporting tabular enrichment tasks, from different perspectives: semantic web resources, reconciliation and semantic annotation methods, and APIs. First, tables can be enriched by leveraging semantic web resources such as Knowledge Graphs (KG) and linked data principles [3, 4, 6, 7], improving downstream tasks as shown in previous work [4, 5]. Second, semantic annotation solutions based on Semantic Table Interpretation (STI) [8] can further expand the functionalities of reconciliation methods: they can label columns, define relationships between column pairs, and link cells to entities in KGs (also referred to as entity linking). Third, APIs can facilitate semantic interoperability by returning identifiers of specific entities, such as geocoding services identifying points by their coordinates.

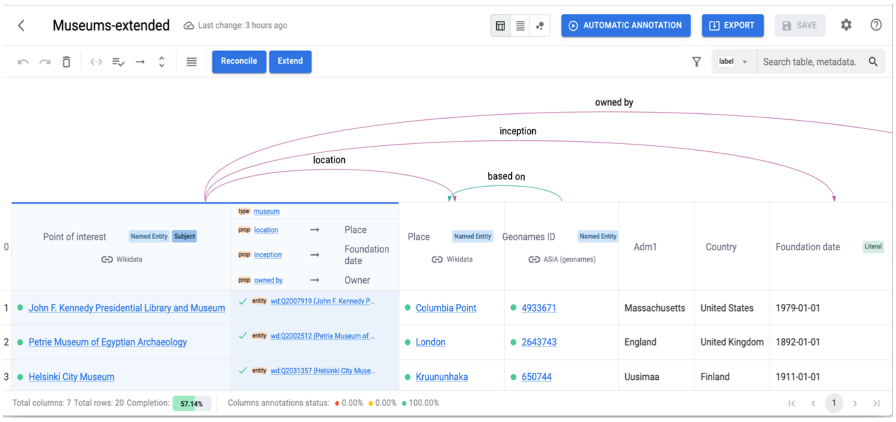
Links to entities from existing data sources are leverages for fetching data from third-party data sources. Considering linking as the process of getting entity identifiers for the values in the table, semantic data enrichment can be interpreted as a *link & extend* process, where data are linked and links are used to fetch more data; these two operations, possibly involving different data sources, can be combined into a sequence of enrichment steps to form data pipelines.

Human interaction also plays an important role in data enrichment, especially in the initial exploratory phase, where users need to understand how to enrich and with which data, inspect the result of enrichment and gain insight into the quality of enrichment operations characterized by intrinsic uncertainty, such as linking/reconciliation steps[1].

Posters, Demos, and Industry Tracks at ISWC 2024, November 13–15, 2024, Baltimore, USA

flavio.depaoli@unimib.it (F. De Paoli); roberto.avogadro@sintef.no (R. Avogadro); marcoripa96@gmail.com (M. Ripamonti); matteo.palmonari@unimib.it (M. Palmonari)

[1] More details about the role of semantics for data enrichment are discussed in a tutorial presented at ESWC2024, whose material is available online at https://enrichmydata.github.io/eswc2024-tutorial/



0	Point of interest Named Entity Subject Wikidata	type museum prop location → Place prop inception → Foundation date prop owned by → Owner	Place Named Entity Wikidata	Geonames ID Named Entity ASIA (geonames)	Adm1	Country	Foundation date Literal
1	● John F. Kennedy Presidential Library and Museum	✓ entity wd:Q2007919 (John F. Kennedy P...	● Columbia Point	● 4933671	Massachusetts	United States	1979-01-01
2	● Petrie Museum of Egyptian Archaeology	✓ entity wd:Q2002512 (Petrie Museum of ...	● London	● 2643743	England	United Kingdom	1892-01-01
3	● Helsinki City Museum	✓ entity wd:Q2031357 (Helsinki City Muse...	● Kruununhaka	● 650744	Uusimaa	Finland	1911-01-01

Total columns: 7 Total rows: 20 Completion: 57.14% Columns annotations status: ● 0.00% ● 0.00% ● 100.00%

Figure 1: An example of a table extension performed with the GUI of SemTUI: the source table has headers *Point of interest*, *Place*, and *Foundation date*. The table was annotated with Wikidata, using an STI service: properties that state the relationships between columns are shown on the top; linked cells are highlighted (green indicates confidence); the first column is expanded to show metadata and details about the annotations, i.e., the estimated column type (Museum), the outgoing properties; and identifiers and labels of the entities. Places was further reconciled against GeoNames to add the new column *GeoNames ID* and two more columns with data from GeoNames: *Adm1* and *Country*.

In this paper, we present a demonstration of SemTUI [9, 10], a modular framework for interactive enrichment of tabular data that exploits semantics under the three perspectives discussed above. It comes with a graphical user interface (GUI), where users can upload, visualize, enrich, and finally export a table. SemTUI interoperates with multiple services, especially data reconciliation and extension services, including end-to-end table annotation services designed for STI. Some services were developed by us, e.g., Alligator [11][2] for reconciliation and end-to-end annotation, while others are based on third-party APIs integrated into the framework, e.g., GeoNames [3] and Here [4] APIs for geocoding.

An example of enrichment with the SemTUI GUI is described in Figure 1. SemTUI is designed as an open system, where more services can be integrated with limited effort and is released as open source under Apache 2.0 License. Links to GitHub repositories and two demonstration videos are available on the tool documentation page [5].

2. State of the Art and Novelty

The main novelty in SemTUI consists in combining five main features, eventually enabling sequences of data enrichment operations according to the *link and extend* paradigm: 1) specification of data enrichment operations and inspection/editing of the results through a GUI (with a special focus on reconciliation services); 2) support for end-to-end STI; 3) support for an open set of reconciliation and extension services; 4) support for KG-based enrichment 5) generalization of semantic-based enrichment to consider third-party APIs.

[2] The paper presents an improved method for cell entity annotation; for all the other STI tasks, Alligator reuses the algorithms of s-elBat [12]

[3] https://www.geonames.org/

[4] https://www.here.com/docs/

[5] https://i2tunimib.github.io/I2T-docs/resources/

The primary competitors of SemTUI that offer semantic table enrichment functionalities do not encompass all five features. **OpenRefine**[6] served as an inspiration for our work, sharing many similarities in approach despite its primary focus on data cleaning operations. ***Kgextension*** [3] is a library designed to facilitate KG-based enrichment within *scikit-learn*. However, its interaction is mediated through a notebook interface, which is less intuitive compared to a graphical user interface. **MAGIC** [7] provides support for KG-based enrichment, but it is tailored to a specific knowledge graph (KG). **DAGOBAH-UI** [6] is a tool focused on STI annotations and leverages links for enrichment from Wikidata. **ASIA**, a previous tool developed by us and integrated into a broader data transformation framework [13], can be considered a precursor to SemTUI. ASIA had limitations in visual exploration and extensibility for integrating external enrichment services, these issues have been addressed in SemTUI through a redesign of the interaction paradigm and graphical user interface (GUI). While SemTUI supports automatic STI annotation services, it focuses on enrichment features and does not support arbitrary data transformations like ASIA[7]. Additionally, few other tools such as MantisTable [14] exist for visualizing or editing STI annotations; some of these tools are not maintained anymore and some other have been developed in the context of the SemTab challenge [15], an initiative to evaluate STI approaches; however, all these tools offer limited support for data enrichment scenarios similar to *link and extend* processes.

Table 1
Comparison between SemTUI and frameworks that provide semantic enrichment functionalities.

Tool	GUI Data Enrichment	End-to-End STI Support	Extensible Services	KG-Based Enrichment	API Generalization
OpenRefine	yes	no	partially	partially	partially
Kgextension	no	no	yes	yes	partially
MAGIC	yes	no	no	partially	no
DAGOBAH-UI	partially	yes	no	no	no
ASIA	partially	no	yes	yes	no

For space constraints, Table 1 provides a comparative summary of SemTUI and other frameworks that deliver semantic enrichment functionalities. Features exclusive to SemTUI, which are not supported by the other tools, are indicated by *no*. The features labeled *partially* denote those offered by the respective tool but enhanced to some extent in SemTUI. Features marked *yes* indicate functionalities that are comparable to those provided by SemTUI.

3. Main Features and Demonstration

For more details, we refer to [9], where we also report some preliminary user studies; we summarize the main features and some insights into pilot applications here below.

Annotation and reconciliation services. End-to-end STI annotations (computed by clicking on the Automatic Annotation button shown on the top-right corner in Figure 1) are currently served by a selected service (Alligator). Other reconciliation services interoperate following the

[6] https://openrefine.org/

[7] Another limitation of the current version of SemTUI w.r.t. ASIA is the lack of support for manual editing of schema annotations; this feature is planned for future integration

specification of W3C Reconciliation Service API v0.2[8] and include, among others: geocoding services based on GeoNames and Here APIs, linking services based on GeoNames APIs and Wikidata (via Alligator [11] and OpenRefine APIs[9]).

Extension services. Extension services are accessed by clicking on a reconciled column that forms the input for the extension service and by specifying in a widget other input conditions (more optional columns) and the features to add in new columns. Extension services currently supported include, among others: Wikidata properties, geographical data (GeoNames attributes, and route information from Here APIs), and meteorological data (from OpenMeteo[10]).

Visual exploration and edit functionalities. Entities identified by shared systems of identifiers can be inspected by the users. The users can look at the best candidates for each cell, modify the link, or enter a new identifier. Visual codes (colors and shapes) are introduced to communicate the status of reconciliation for each cell (confidence, manual revision, etc.).

Pilot applications. We summarize various pilot applications of SemTUI for data enrichment tested so far. 1) Enrichment from annotations computed by STI approaches, where additional data are collected from an annotated table using found links (see the example in Figure 1 or a short online demo. 2) Enrichment of data from digital marketing campaigns, where historical performance data is enriched using API services to obtain coordinates and weather data; this approach streamlines an enrichment use case discussed in prior work [5]. 3) Enrichment of procurement data with information from Wikidata, where organization names are linked to Wikidata for additional information to support classification algorithms (see also a short online demo). 4) Enrichment of urban data with further insights from human interaction [10, 16]. 5) Enrichment of company data with corporate knowledge graphs, integrating Atoka's entity reconciliation and data extension services to streamline client data enrichment without needing specialized data scientists.

Demonstration. In the demonstration, we plan to use data from the first and the second scenarios. In the first scenario, we use tables from the SemTab challenges and show an example of end-to-end annotation process with the Alligator STI framework and show examples of entity links found with other reconciliation, e.g., the Wikidata lookup service; we will show how to explore the results and change links that are not correct; finally, we will show examples of enrichment using Wikidata-based extension services. In the second scenario, we use digital marketing data reconciled and enriched with third-party services to demonstrate the service-based generalization of the link & extend paradigm using APIs. In both scenarios, we will walk the audience through different enrichment steps, discussing the features of SemTUI.

4. Acknowledgements

This work has been partially funded by the European innovation action *enRichMyData* (HE 101070284) and the Italian PRIN project *Discount Quality for Responsible Data Science: Human-in-the-Loop for Quality Data* (202248FWFS) funded by the European Community - Next Generation EU.

[8] https://www.w3.org/community/reports/reconciliation/CG-FINAL-specs-0.2-20230410/
[9] https://wikidata.reconci.link/
[10] https://open-meteo.com/

References

[1] M. Hameed, F. Naumann, Data preparation: A survey of commercial tools, SIGMOD Rec. 49 (2020) 18–29.

[2] M. Ciavotta, V. Cutrona, F. De Paoli, N. Nikolov, M. Palmonari, D. Roman, Supporting semantic data enrichment at scale, in: Technologies and Applications for Big Data Value, Springer, 2022, pp. 19–39.

[3] T.-C. Bucher, X. Jiang, O. Meyer, S. Waitz, S. Hertling, H. Paulheim, scikit-learn pipelines meet knowledge graphs: The python kgextension package, in: ESWC 2021 Satellite Events: Revised Selected Papers, Springer, 2021, pp. 9–14.

[4] A. Harari, G. Katz, Automatic features generation and selection from external sources: a DBpedia use case, Information Sciences 582 (2022) 398–414.

[5] V. Cutrona, F. De Paoli, A. Košmerlj, N. Nikolov, M. Palmonari, F. Perales, D. Roman, Semantically-enabled optimization of digital marketing campaigns, in: ISWC, Springer, 2019, pp. 345–362.

[6] C. Sarthou-Camy, G. Jourdain, Y. Chabot, P. Monnin, F. Deuzé, V.-P. Huynh, J. Liu, T. Labbé, R. Troncy, DAGOBAH-UI: a new hope for semantic table interpretation, in: ESWC Demo Papers, Springer, 2022, pp. 107–111.

[7] B. Steenwinckel, F. De Turck, F. Ongenae, MAGIC: Mining an augmented graph using INK, starting from a CSV., in: SemTab@ ISWC, 2021, pp. 68–78.

[8] J. Liu, Y. Chabot, R. Troncy, V.-P. Huynh, T. Labbé, P. Monnin, From tabular data to knowledge graphs: A survey of semantic table interpretation tasks and methods, Journal of Web Semantics 76 (2023).

[9] M. Ripamonti, F. De Paoli, M. Palmonari, SemTUI: a framework for the interactive semantic enrichment of tabular data, arXiv preprint arXiv:2203.09521 (2022).

[10] F. De Paoli, M. Ciavotta, R. Avogadro, E. Hristov, M. Borukova, D. Petrova-Antonova, I. Krasteva, An interactive approach to semantic enrichment with geospatial data, Data Knowledge Engineering (2024).

[11] R. Avogadro, M. Ciavotta, F. De Paoli, M. Palmonari, D. Roman, Estimating link confidence for human-in-the-loop table annotation, in: IEEE/WIC International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE, 2023, pp. 142–149.

[12] M. Cremaschi, R. Avogadro, D. Chieregato, s-elBat: a semantic interpretation approach for messy table-s, in: SemTab @ ISWC, CEUR-WS. org, 2022.

[13] V. Cutrona, M. Ciavotta, F. De Paoli, M. Palmonari, et al., ASIA: a tool for assisted semantic interpretation and annotation of tabular data, in: ISWC Demo Papers, volume 2456, CEUR-WS.org, 2019, pp. 209–212.

[14] M. Cremaschi, A. Rula, A. Siano, F. De Paoli, MantisTable: a tool for creating semantic annotations on tabular data, in: ESWC Demo Papers, Springer, 2019, pp. 18–23.

[15] O. Hassanzadeh, N. Abdelmageed, V. Efthymiou, J. Chen, V. Cutrona, M. Hulsebos, E. Jiménez-Ruiz, A. Khatiwada, K. Korini, B. Kruit, et al., Results of SemTab 2023, in: SemTab @ ISWC, volume 3557, CEUR-WS.org, 2023, pp. 1–14.

[16] I. Krasteva, D. Petrova-Antonova, F. De Paoli, E. Hristov, M. Borukova, M. Ciavotta, R. Avogadro, Geospatial enrichment of urban data for advanced city planning: a pilot study, in: BigData, IEEE, 2023, pp. 3139–3143.