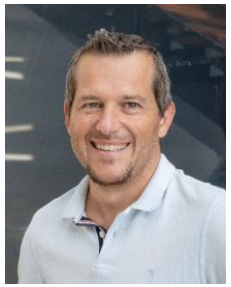
From Tabular Data to Knowledge Graphs

A Survey of Semantic Table Interpretation Tasks and Methods

J. Liu
@yansera

P. Huynh
@vp_huynh

Y. Chabot
@yoan_chabot



R. Troncy
@rtroncy

The article was originally published in the Journal of Web Semantics.
Liu, J., Chabot, Y., Troncy, R., Huynh, V. P., Labbé, T., & Monnin, P. (2022). From tabular data to knowledge graphs: A survey of semantic table interpretation tasks and methods. Journal of Web Semantics, Vol. 76, 2023.
https://doi.org/10.1016/j.websem.2022.100761

Motivation

Movie	Location
Our Happy Lives	Belfort
The French Kissers	Rennes

Q171545
Belfort

Q484170
Commune of France

P31 instance of

P31 instance of

P17 country

P840 narrative location

P495 country of origin

Q142
France

P17 country

Q647
Rennes

P495 country of origin

P840 narrative location

Q3344332
Our Happy Lives

P31 instance of

Q11424
film

P31 instance of

Q745690
The French Kissers

« Our Happy Lives » is a French film, filmed in Belfort.
« Belfort » is a French city.
…

Take aways

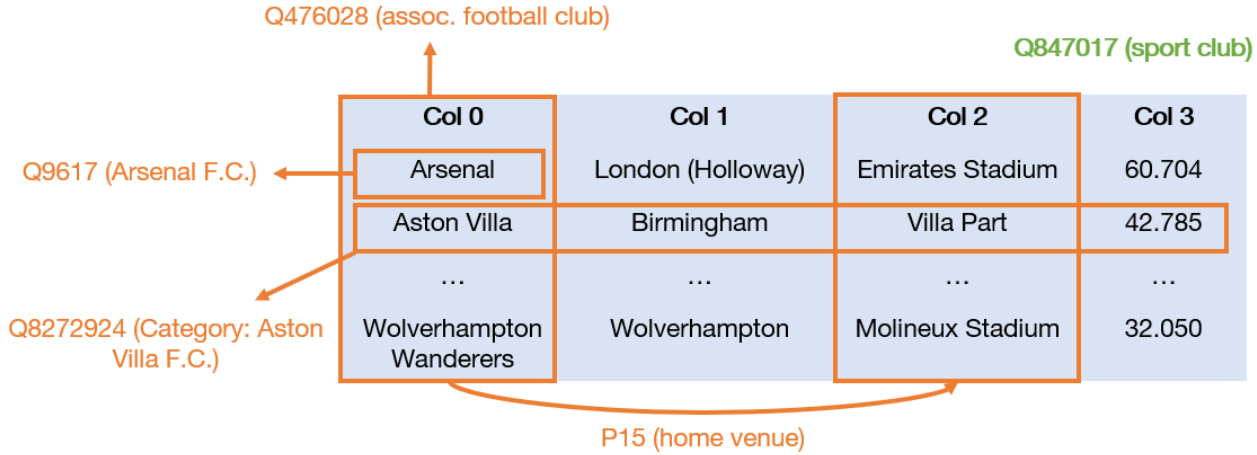
1. We introduce a fine-grained taxonomy of table types

Year	Category	Nominees(s)	Nominated for	Result
1996	Best World Music Album	Cesaria Evora[4]	*Cesaria Evora*	Nominated
1998			*Cabo Verde*	Nominated
1999			*Miss Perfumado*	Nominated
2000			*Café Atlantico*	Nominated
2002			*São Vicente*	Nominated
2004	Best Contemporary World Music Album		*Voz D'Amor*	Won

		Predicted condition	
	Total 8 + 4 = 12	Cancer 7	Non-cancer 5
Actual condition	Cancer 8	6	2
	Non-cancer 4	1	3

Country	France
Region	Île-de-France
Department	Paris
Intercommunality	Métropole du Grand Paris
Subdivisions	20 arrondissements
Government	
• Mayor (2020–2026)	Anne Hidalgo[1] (PS)
Area[1]	105.4 km² (40.7 sq mi)
• Urban (2020)	2,853.5 km² (1,101.7 sq mi)
• Metro (2020)	18,940.7 km² (7,313.0 sq mi)
Population (2023)[2]	2,102,650
• Density	20,000/km² (52,000/sq mi)
• Urban (2019[3])	10,858,852

2. We propose five main tasks for Semantic Table Interpretation (STI)
   - Cell-entity Annotation
   - Column-Type Annotation
   - Column-property Annotation
   - Topic Annotation
   - Row-to-Instance Annotation



3. We review and group STI methods and systems into 3 families

4. We provide insights on current performance and describe the remaining challenges

Context

Massive amount of heterogeneous tables … but few of them are annotated

- Web Tables extracted from the Common Crawl: 4.2B tables http://webdatacommons.org/structureddata/schemaorgtables/
- Encyclopedic tables: Wikipedia

Largest metropolitan areas in France
2019 census

https://en.wikipedia.org/wiki/France#Major%20cities

Rank	Name	Region	Pop.	Rank	Name	Region	Pop.
1	Paris	Île-de-France	13,114,718	11	Grenoble	Auvergne-Rhône-Alpes	717,469
2	Lyon	Auvergne-Rhône-Alpes	2,280,845	12	Rouen	Normandy	705,627
3	Marseille	Provence-Alpes-Côte d'Azur	1,873,270	13	Nice	Provence-Alpes-Côte d'Azur	615,126
4	Lille	Hauts-de-France	1,510,079	14	Toulon	Provence-Alpes-Côte d'Azur	573,230
5	Toulouse	Occitania (administrative region)	1,454,158	15	Tours	Centre-Val de Loire	519,778
6	Bordeaux	Nouvelle-Aquitaine	1,363,711	16	Nancy	Grand Est	510,306
7	Nantes	Pays de la Loire	1,011,020	17	Clermont-Ferrand	Auvergne-Rhône-Alpes	507,479
8	Strasbourg	Grand Est	853,110	18	Saint-Étienne	Auvergne-Rhône-Alpes	498,849
9	Montpellier	Occitania (administrative region)	801,595	19	Caen	Normandy	472,161
10	Rennes	Brittany	755,668	20	Orléans	Centre-Val de Loire	451,373

- Non-encyclopedic tables, Enterprise tables

https://pencilandpaper.io/articles/ux-pattern-analysis-enterprise-data-tables/

ID	Name	Status	Submission
BASD-55498	Danesh Bashir	Approved	12/12/2020
FINA-97846	Alexander Finn	Pending	-
DESI-48765	Isabelle Desmarais	Denied	Resubmit
SHEW-11687	Wendy Shea	Abandoned	Resubmit
WILJ-10348	Johnny Wilson	Approved	04/01/2021

Context

Tabular data is highly heterogenous

- Complex table layouts (multi-level headers, orientation, merged cells, etc.)
- Semantic heterogeneity: ambiguous, hidden information
- Multiple datatype (numeric, datetime, unit, symbol, etc.)
- Incomplete, and potentially dynamic

Club	Season	League			National Cup[a]		State League[b]	
		Division	Apps	Goals	Apps	Goals	Apps	Goals
Santos	2009	Série A	33	10	3	1	12	3
	2010	Série A	31	17	8	11		
	2011	Série A	21	13	—			
	2012	Série A	17	14	—			
	2013	Série A	1	0	4	1		
	Total		102	54	15	13		

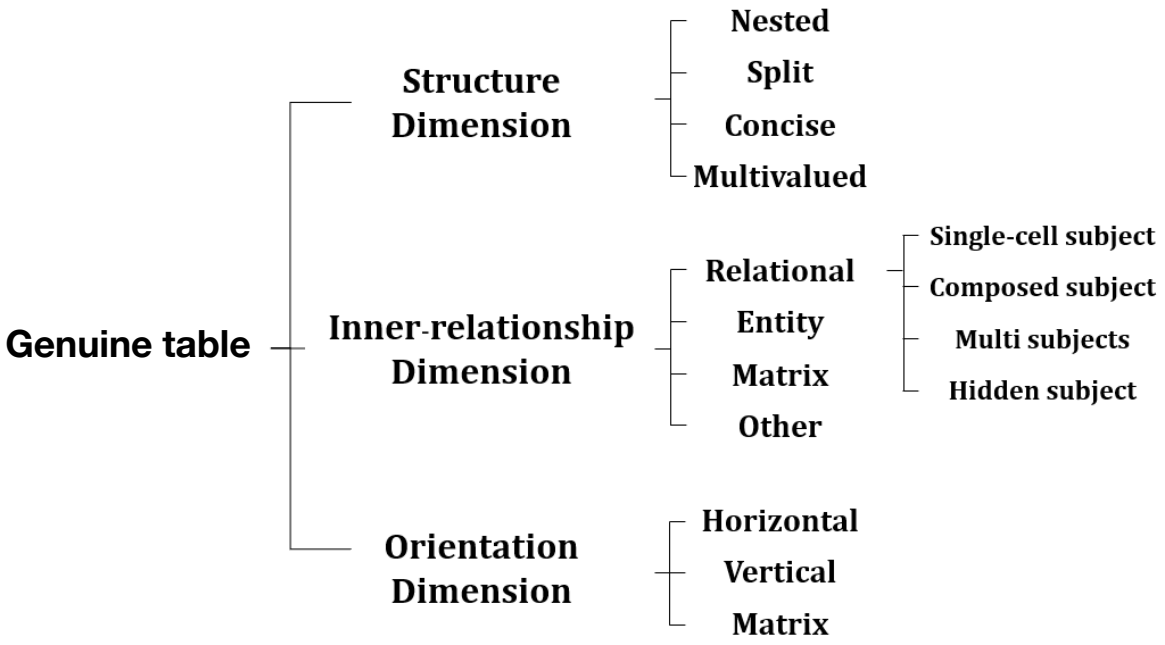
1	10 juin 1998	Brésil	2 - 1	Écosse
2	10 juin 1998	Maroc	2 - 2	Norvège
17	16 juin 1998	Écosse	1 - 1	Norvège
18	16 juin 1998	Brésil	3 - 0	Maroc
35	23 juin 1998	Écosse	0 - 3	Maroc
36	23 juin 1998	Brésil	1 - 2	Norvège

		Predicted condition	
	Total 8 + 4 = 12	Cancer 7	Non-cancer 5
Actual condition	Cancer 8	6	2
	Non-cancer 4	1	3

Year	Category	Nominees(s)	Nominated for	Result
1996	Best World Music Album	Cesaria Evora[4]	*Cesaria Evora*	Nominated
1998			*Cabo Verde*	Nominated
1999			*Miss Perfumado*	Nominated
2000			*Café Atlantico*	Nominated
2002			*São Vicente*	Nominated
2004	Best Contemporary World Music Album		*Voz D'Amor*	Won

Country	France
Region	Île-de-France
Department	Paris
Intercommunality	Métropole du Grand Paris
Subdivisions	20 arrondissements
Government	
• Mayor (2020–2026)	Anne Hidalgo[1] (PS)
Area[1]	105.4 km² (40.7 sq mi)
• Urban (2020)	2,853.5 km² (1,101.7 sq mi)
• Metro (2020)	18,940.7 km² (7,313.0 sq mi)
Population (2023)[2]	2,102,650
• Density	20,000/km² (52,000/sq mi)
• Urban (2019[3])	10,858,852
• Urban density	3,800/km² (9,900/sq mi)
• Metro (Jan. 2017[4])	13,024,518

Towards a Classification of Tables



Three dimensions

- **Structure Dimension [1]**
  if the table is nested into another table, if cells are merged or multivalued
- **Inner-relationship Dimension [2]**
  focus on exploring the topology of semantic connection between cells
- **Orientation Dimension [1,3]**
  studies the horizontal/vertical/matrix arrangement of the attribute values of the subject

(a) Relational table

Year	Category	Nominees(s)	Nominated for	Result
1996	Best World Music Album	Cesaria Evora[4]	*Cesaria Evora*	Nominated
1998			*Cabo Verde*	Nominated
1999			*Miss Perfumado*	Nominated
2000			*Café Atlantico*	Nominated
2002			*São Vicente*	Nominated
2004	Best Contemporary World Music Album		*Voz D'Amor*	Won

(b) Matrix table

		Predicted condition	
	Total 8 + 4 = 12	Cancer 7	Non-cancer 5
Actual condition	Cancer 8	6	2
	Non-cancer 4	1	3

(c) Entity table

Country	France
Region	Île-de-France
Department	Paris
Intercommunality	Métropole du Grand Paris
Subdivisions	20 arrondissements
Government	
• Mayor (2020–2026)	Anne Hidalgo[1] (PS)
Area[1]	105.4 km^2 (40.7 sq mi)
• Urban (2020)	2,853.5 km^2 (1,101.7 sq mi)
• Metro (2020)	18,940.7 km^2 (7,313.0 sq mi)
Population (2023)[2]	2,102,650
• Density	20,000/km^2 (52,000/sq mi)
• Urban (2019[3])	10,858,852

[1] Lautert et al.. Web table taxonomy and formalization. ACM SIGMOD Record, 42(3):28–33, 2013.
[2] Ritze. Web-scale web table to knowledge base matching. PhD thesis, 2017.
[3] Eberius et al. Building the Dresden web table corpus: A classification approach. In 2015 IEEE/ACM 2nd International Symposium on Big Data Computing (BDC), pages 41–50. IEEE, 2015.

Semantic Table Interpretation tasks

Input Table

Col 0	Col 1	Col 2	Col 3
Arsenal	London (Holloway)	Emirates Stadium	60.704
Aston Villa	Birmingham	Villa Part	42.785
…	…	…	…
Wolverhampton Wanderers	Wolverhampton	Molineux Stadium	32.050

Column-Type Annotation (CTA)
Q476028 (assoc. football club)

Topic Annotation
Q847017 (sport club)

Annotated Table

Col 0	Col 1	Col 2	Col 3
Arsenal	London (Holloway)	Emirates Stadium	60.704
Aston Villa	Birmingham	Villa Part	42.785
…	…	…	…
Wolverhampton Wanderers	Wolverhampton	Molineux Stadium	32.050

Cell-Entity Annotation (CEA)
Q9617 (Arsenal F.C.)

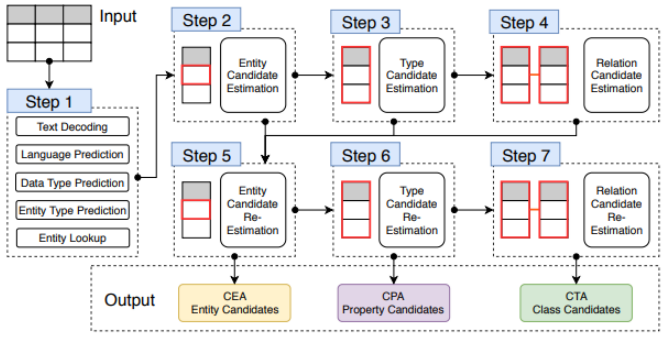
Row-to-Instance
Q8272924 (Category: Aston Villa F.C.)

Columns-Property Annotation (CPA)
P15 (home venue)

Families of STI approaches

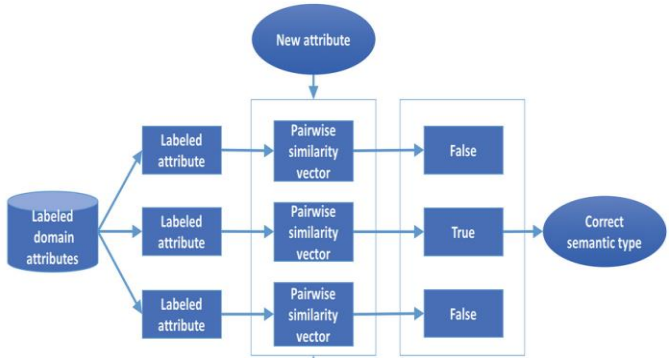
Heuristic

- Heuristically process the annotation on a lookup service
- String similarity measures, majority voting, TFIDF or probabilistic frameworks
- For example, Mtab[1]



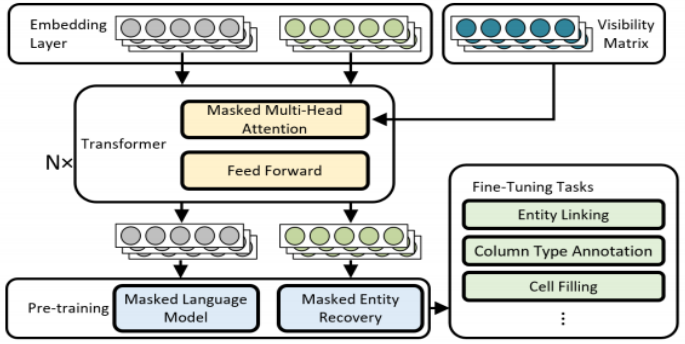
Feature Enginerring

- Statistical and lexical features
- SVM, Random Forest, KNN, etc.
- For example, DSL[2]



Deep Learning

- KG embedding
- Table embedding
- For example, TURL[3]



[1] Nguyen, Phuc, et al. "MTab: Matching tabular data to knowledge graph using probability models." In Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab), 2019.

[2] Minh, P., Suresh, A., Craig, A.K., Pedro, S.: Semantic Labeling: A Domain-Independent Approach. In: 15th International Semantic Web Conference (ISWC). pp. 446—462. Springer (2016)

[3] Deng, X., Sun, H., Lees, A., Wu, Y., Yu, C.: TURL: Table Understanding through Representation Learning. arXiv:2006.14806 (2020)

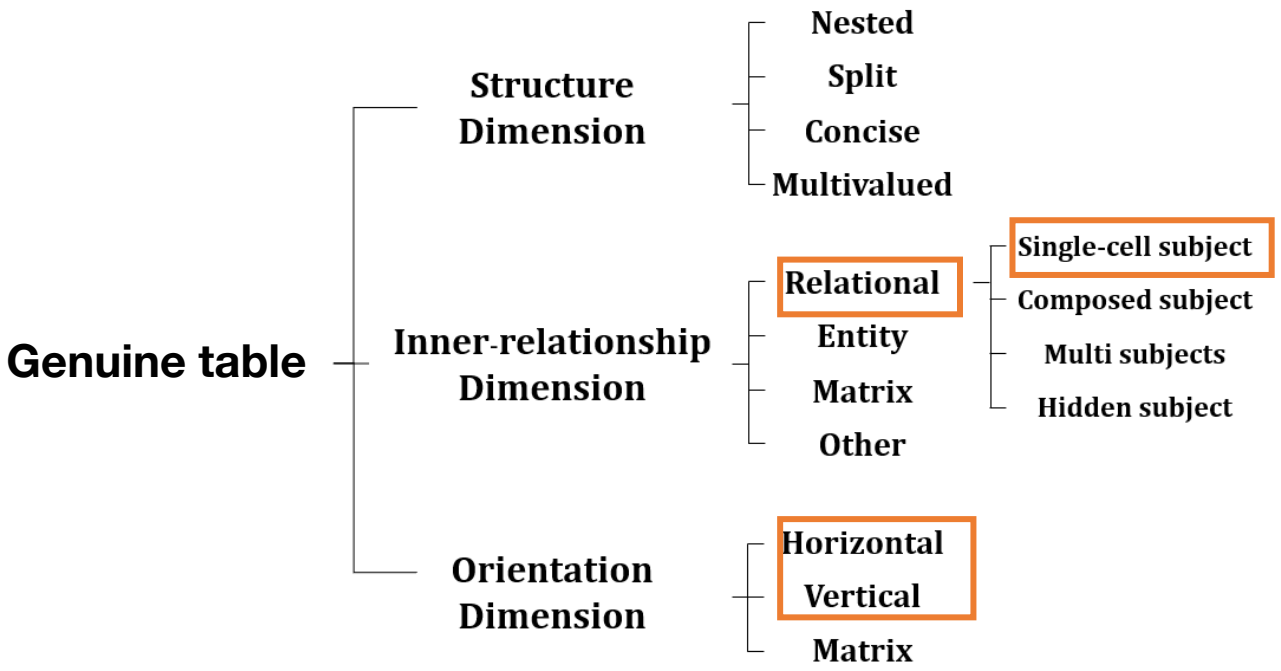
Performance

Best F1 scores on different datasets for Semantic Table Interpretation (STI) tasks – as of December 2022

Dataset	#Tables	CEA	CTA	CPA
SyntheticTable – SemTab	12173	0.99 (MTab)	0.98 (MTab)	0.99 (MTab)
WebTables - Limaye	437	0.89 (TabEAno)	0.88 (T2K++)	0.89 (Mulward et al.)
WebTables – T2D	779	0.91 (TabEAno)	0.98 (ColNet)	0.91 (T2K++)
WebTables - ToughTable	180	0.95 (DAGOBAH-SL)	0.83 (DAGOBAH-SL)	x
Gittables	1101	x	0.69 (KGCODE-Tab)	x
BioDivTab	45	x	0.87 (KGCODE-Tab)	x
SOTAB	7026	x	0.85 (Doduo)	0.80 (Doduo)

Extensive evaluation on numerous datasets for many systems in the paper

But, this is just the beginning…



Low coverage of state-of-the-art approaches

Name	Pinnacle height	Year	Country	Town	Remarks
Tokyo Skytree	634 m (2,080 ft)	2011	Japan	Tokyo	
Kyiv TV Tower	385 m (1,263 ft)	1973	Ukraine	Kyiv	
Dragon Tower	336 m (1,102 ft)	2000	China	Harbin	
Tokyo Tower	333 m (1,093 ft)	1958	Japan	Tokyo	
WITI TV Tower	329.4 m (1,081 ft)	1962	United States	Shorewood, Wisconsin	
St. Petersburg TV Tower	326 m (1,070 ft)	1962	Russia	Saint Petersburg	

The research area is becoming more and more active

Challenge SemTab@ISWC

Corpus: GitTable + BioTable

Workshop Tabular Data Analysis@VLDB

2019 2020 2021 2022 2023

Corpus:
- WebTables
- T2D
- Limaye

- Corpus: ToughTable
- Emergence of Table representation learning with LM

- Corpus: SOTAB
- Workshop Table Representation Learning@Neurips

Conclusion and Open Challenges

- Several STI systems are now available

https://mtab.app/mtab

https://bitbucket.org/disco_unimib/selbat

https://developer.orange.com/apis/table-annotation

- Remaining challenges to be solved to make STI more effective in practice:
  - Ability to handle simultaneously various table types and layouts
  - End-to-End Semantic Table Annotation
  - Efficient KG indexing and exploitation, KG varying in terms of completeness and size

- Potential future works:
  - Table representation learning: learning latent representation from structured table using LLM.
  - Emerging abilities of generative LLM: unified text-2-structure framework can perform multiple STI tasks at once (i.e. multitask learning). [1]

[1] Huynh et al., Towards Generative Semantic Table Intepretation, Tabular Data Analysis@VLDB 2023