From Web Tables to a Knowledge Graph:
Prospects of an End-to-End Solution*

Alexey Shigarov, Nikita Dorodnykh, Alexander Yurin,
Andrey Mikhailov and Viacheslav Paramonov

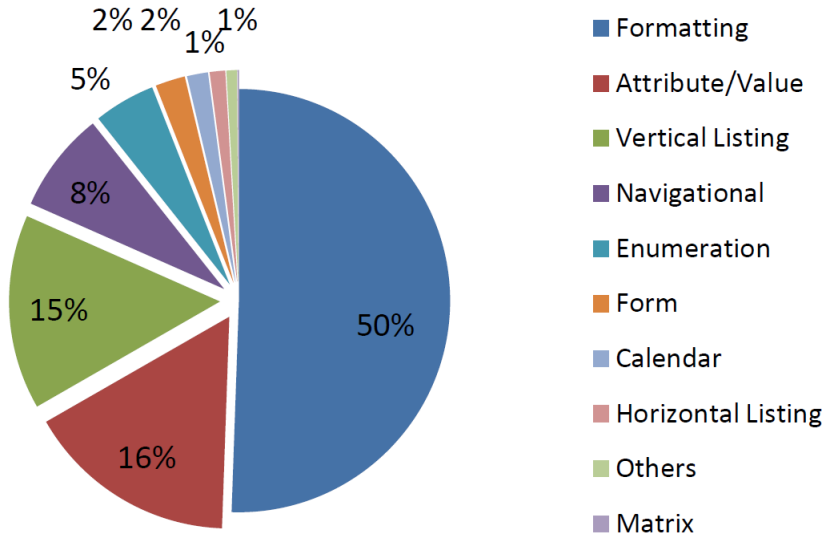
Matrosov Institute for System Dynamics and Control Theory,
Siberian Branch of the Russian Academy of Sciences

ITAMS (Irkutsk, Russia), September 17th, 2021

* This work was supported by the Russian Science Foundation (Grant No. 18-71-10001)

Web Tables

- Cafarella et al. (2008) discovered that the Web contains a large volume of relational data
- **Taxonomy of web table types** (Crestan, 2011)
  - Inferred from a big corpora ~ 1.3B
  - **58% used to layout only** — Formatting, Navigational
  - **40% store relational data** — Attribute-Value, Listing, Enumeration, Matrix, Calendar



Distribution of web table types (~ 1,3B) – copied from (Crestan, 2011)

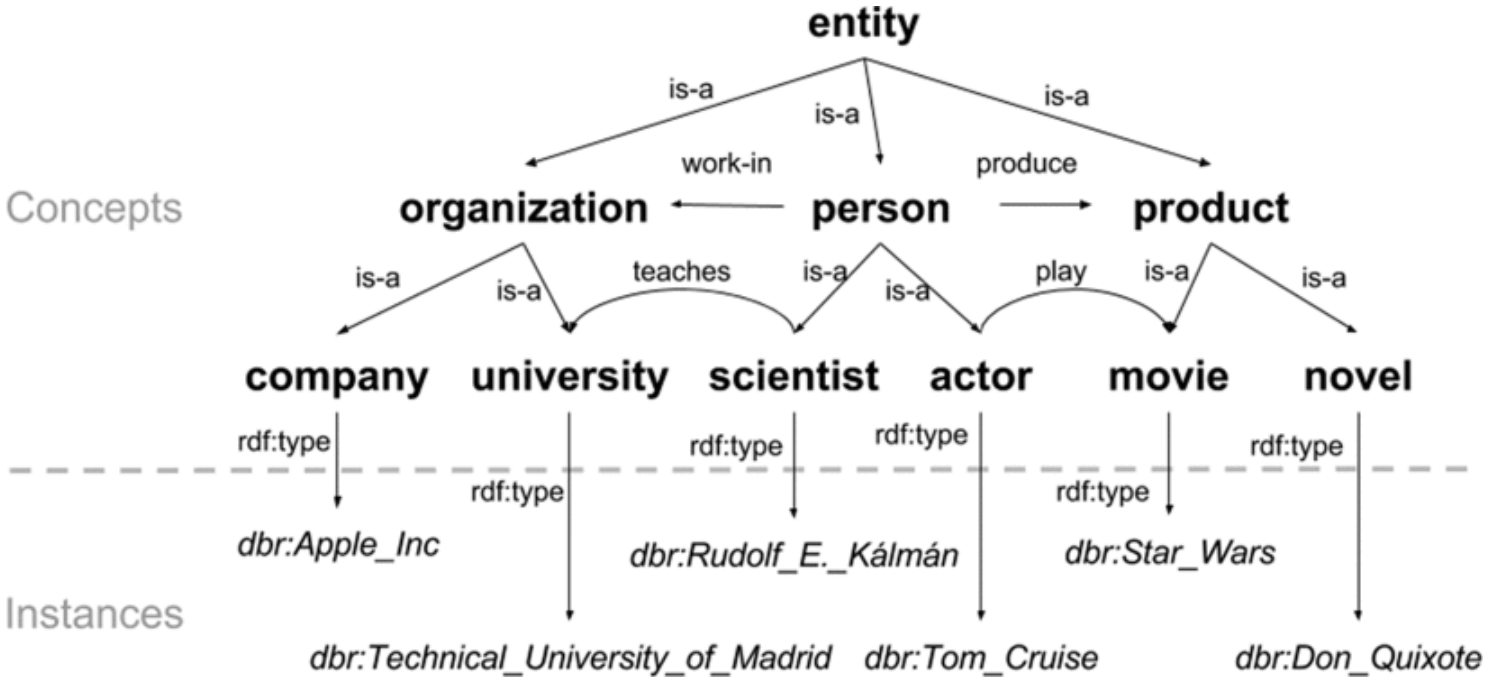
"*The Web is the largest repository of data available, with over 150 million high-quality tables*" (Lautert, 2013)

(Cafarella, 2008) Cafarella, M., et al. (2008). WebTables: exploring the power of tables on the web. *Proc. VLDB Endowment, 1*, 538-549
(Creastan, 2011) Crestan, E. & Pantel, P. (2011). Web-scale table census and classification. Proc. 4th ACM Int. Conf. on Web Search and Data Mining, 545-554
(Lautert, 2013) Lautert, L. R., et al. (2013) Web table taxonomy and formalization. ACM SIGMOD Record. 42, 28-33

What is a Knowledge Graph

Knowledge Graph (Bizer, 2009; Fensel, 2020; Hogan, 2020)

- *A collection of interrelated entities*
- **KG-entity**, a thing of the real-world or abstract concept (e.g. "*War and Peace*", "*Leo Tolstoy*")
- **KG-property** is a name of relation of two entities (e.g. "*is-written-by*")
- **KG-class** is a set of entities with a common set of properties (e.g. "*Book*")
- **KG-instance** is a KG-entity belonged to a KG-class
- **KG-category** is an aspect of semantics of a set of entities (e.g. "*Russian novels*")
- **KG-datatype** is a type of literals (e.g. *NUMERIC*, *DATE*, *STRING*, *URI*)



"A tiny example of knowledge graph" — copied from (Zhu, 2017)

(Bizer, 2009) Bizer, C., et al. (2009). DBpedia - A crystallization point for the Web of Data. J. Web Semantics, 7(3), 154-165
(Fensel, 2020) Fensel, D., et al. (2020). Knowledge graphs: Methodology, Tools and Selected Use Cases
(Hogan, 2020) Hogan, A., et al. (2020). Knowledge graphs. *ArXiv, abs/2003.0232*
(Zhu, 2017) Zhu, G. & Iglesias, C. A. (2017). Computing semantic similarity of concepts in knowledge graphs. IEEE Transactions on Knowledge and Data Engineering, 29(1), 72-85

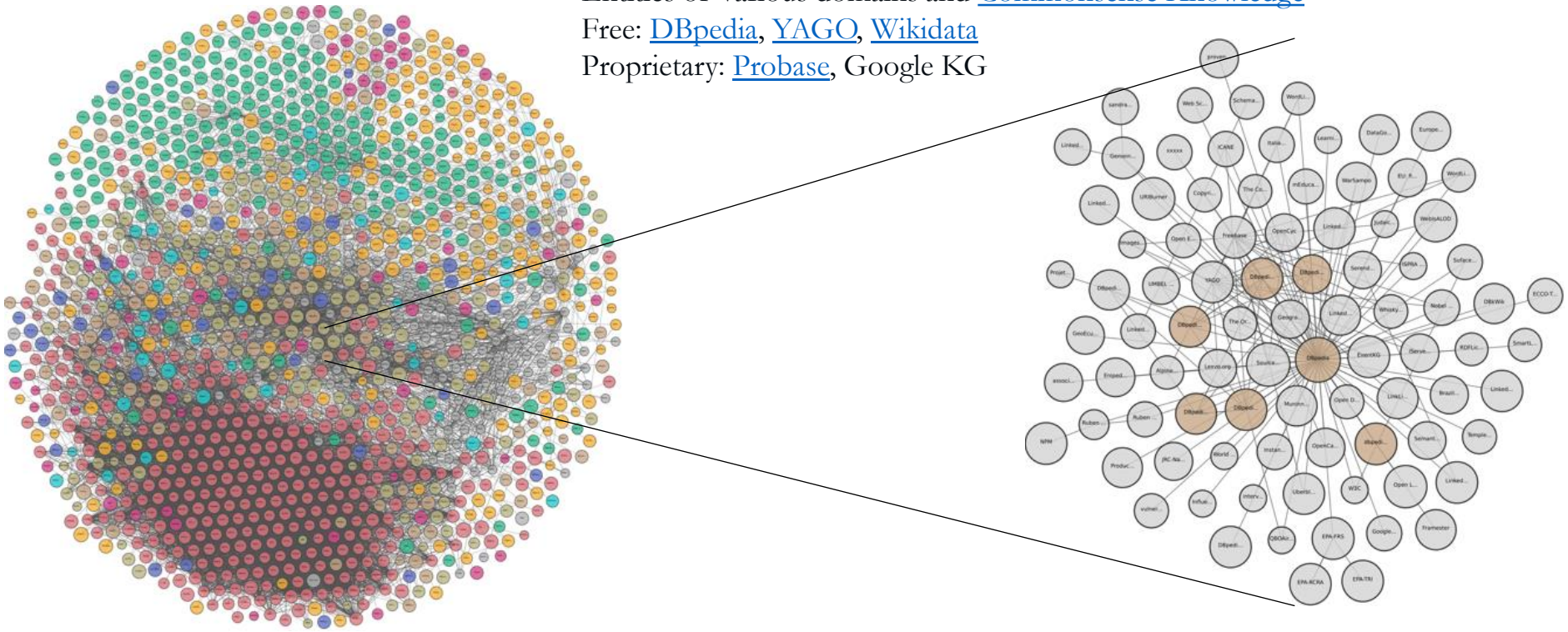
What is a Knowledge Graph

Cross-Domain Knowledge Graph

Entities of various domains and Commonsense Knowledge

Free: DBpedia, YAGO, Wikidata

Proprietary: Probase, Google KG



Linked Open Data (LOD) Cloud – copied from https://lod-cloud.net

Subset of the Cross-Domain Linked Open Data Cloud from lod-cloud.net

Knowledge Engineering & Web Tables

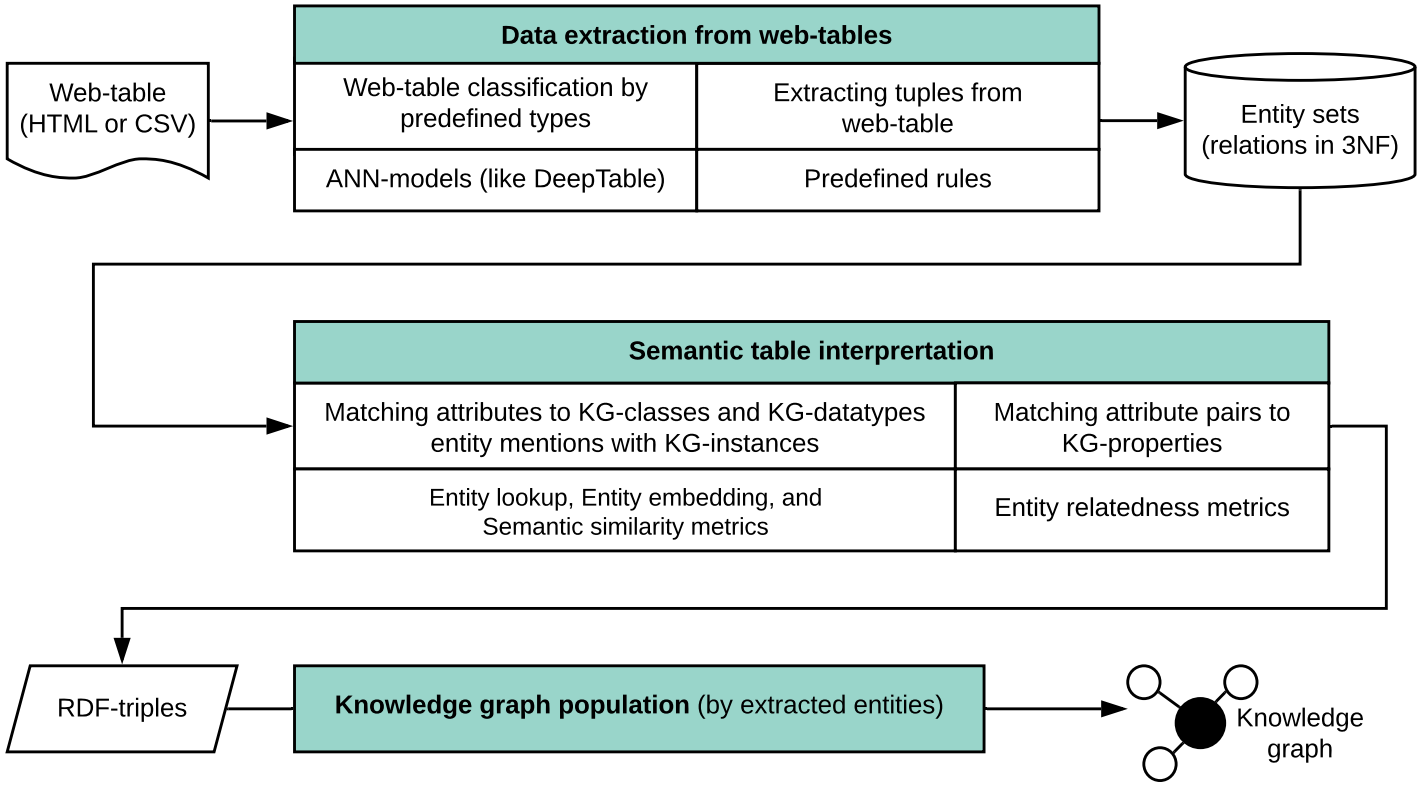
Knowledge Base Construction (KBC)	Knowledge Base Population (KBP)	Knowledge Base Augmentation (KBA)
"*The process of populating a knowledge base […] with information extracted from documents and structured sources*" (Ré, 2014)	"*The task of discovering new facts about entities from a large text corpus, and augmenting a knowledge base with these facts*" (Balog, 2018)	"*Generating new instances of relations using tabular data and updating knowledge bases with the extracted information*" (Zhang&Balog, 2020)

Web tables is a valuable source for the KBC/KBP/KBA
(Ré, 2014; Martinez-Rodriguez, 2020; Zhang&Balog, 2020; Hogan, 2021)

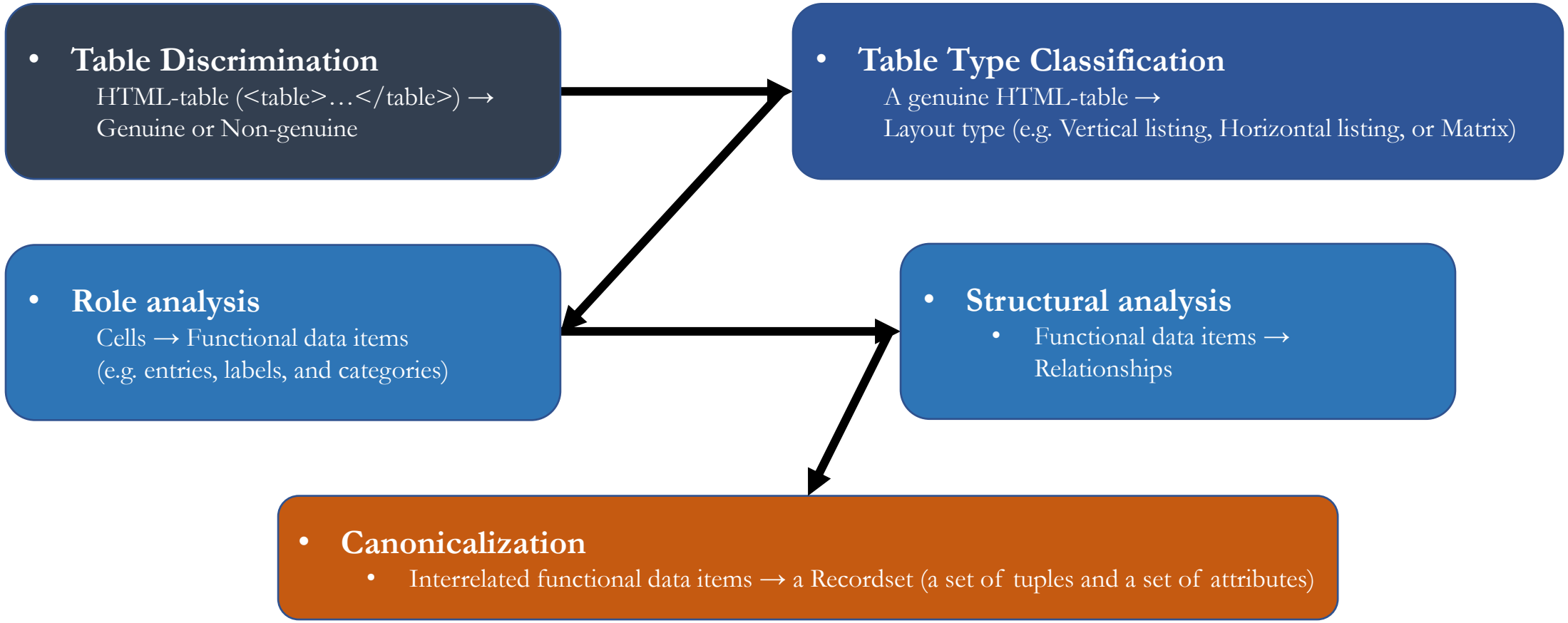
(Balog, 2018) Balog, K. (2018). Populating knowledge bases. Entity-Oriented Search. INRE 39, 189-222
(Hogan, 2021) Hogan, A., et al. (2021). Knowledge graphs. ACM Comput. Surv. 54(4)
(Martinez-Rodriguez, 2020) Martinez-Rodriguez, J. (2020). Information extraction meets the Semantic Web: a survey. Semantic Web, 11(2), 255-335
(Ré, 2014) Ré, C., et al. (2014). Feature engineering for knowledge base construction. IEEE Data Eng. Bull. 37, 26-40
(Zhang&Balog, 2020) Zhang, S. & Balog, K. (2020). Web table extraction, retrieval, and augmentation: a survey. ACM Trans Intell Syst Technol, 11

KBP from Web Tables

- Two sub-tasks: I — **Web Table Extraction (WTE)** & II — **Semantic Table Interpretation (STI)**



Web Table Extraction



Web Table Extraction

- **Challenge**: To support atomic data items not cells

State	STATE ABBREVIATION	2-DIGIT ANSI CODE	2000				2002			
			Total \1 (1,000)	Democratic (1,000)	Republican (1,000)	Percent for leading party	Total \1 (1,000)	Democratic (1,000)	Republican (1,000)	Percent for leading party
United States	US	00	98 800	46 412	46 750	R-47.3	74 707	33 642	37 091	R-49.6
Alabama	AL	01	1 439	486	849	**R-59.0**	1 269	507	695	R-54.7
Alaska	AK	02	274	45	191	R-69.6	228	39	170	R-74.5
Arizona	AZ	04	1 466	558	855	R-58.3	1 194	472	682	R-57.1
Arkansas \2	AR	05	633	355	277	D-56.2	688	392	284	D-57.0
California	CA	06	10 437	5 407	4 446	D-51.8	7 258	3 731	3 226	D-51.4
Colorado	CO	08	1 624	496	969	R-59.7	1 397	589	753	R-53.9
Connecticut \2	CT	09	1 313	699	591	D-53.2	989	509	466	D-51.5
Delaware	DE	10	313	96	212	R-67.6	228	61	165	R-72.1
Florida \2	FL	12	5 011	1 976	2 852	R-56.9	3 767	1 537	2 161	R-57.4

A fragment of a table from SAUS

Q: What is **R-59.0**?

Context:
D — Democratic Party
R — Republican Party

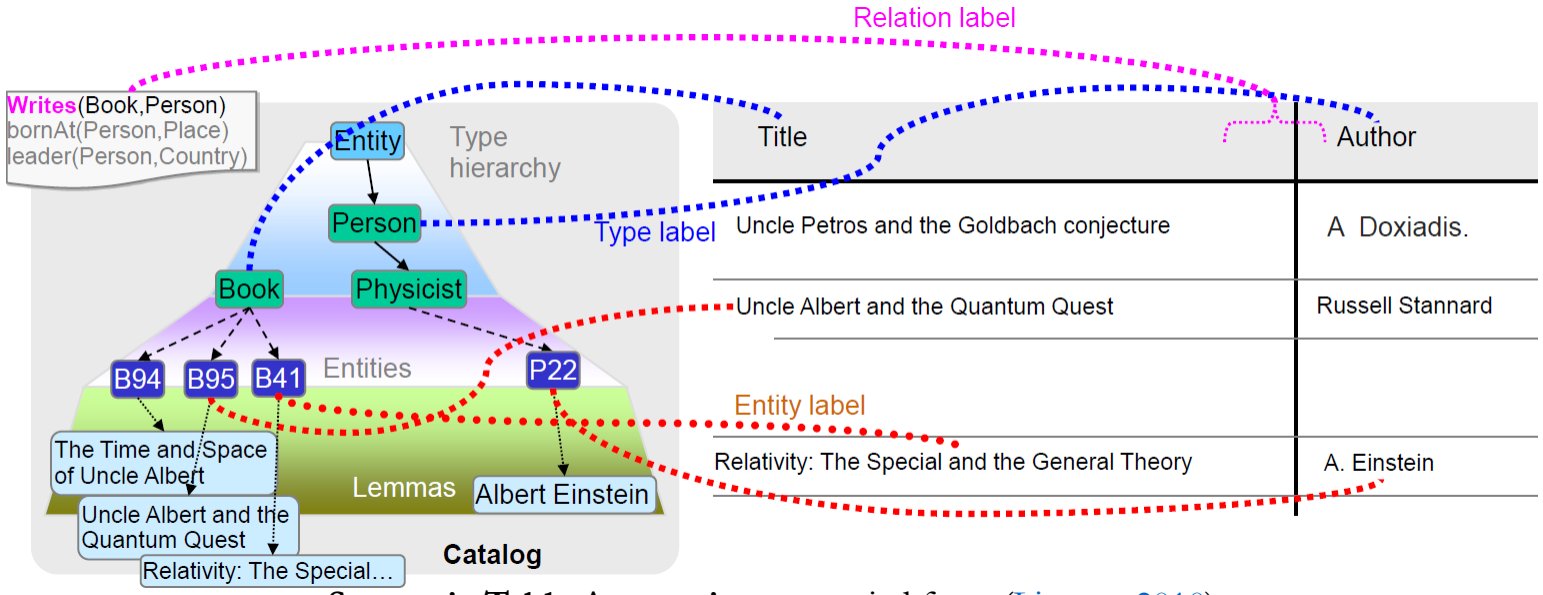
A: **R-59.0** means 2 facts:

- In 2000 year, in Alabama, US, the percent for leading party was — **59.0**
- In 2000 year, in Alabama, US, the percent for leading party was Republican — **R**

Semantic Table Interpretation

- Two works (Limaye, 2010; Syed, 2010) kicked off **Semantic Table Annotation**
- '**Semantic Table Interpretation'** term was introduced in (Zhang, 2014; Zhang, 2017)

- **3 sub-tasks** (Zhang&Balog, 2020)
  - Entity Linking
  - Column Type Identification
  - Relation Extraction



Semantic Table Annotation — copied from (Limaye, 2010)

(Limaye, 2010) Limaye, G., et al. (2010). Annotating and searching web tables using entities, types and relationships. Proc. VLDB Endowment, 3, 1338-1347
(Syed, 2010) Syed, Z., et al. (2010). Exploiting a web of semantic data for interpreting tables. Proc. 2nd Web Science Conference, 26-27
(Zhang, 2014) Zhang, Z. (2014). Towards efficient and effective semantic table interpretation. The Semantic Web -- ISWC 2014, 8796 LNCS, 487-502
(Zhang, 2017) Zhang, Z. (2017). Effective and Efficient Semantic Table Interpretation Using TableMiner+. Semantic Web, vol. 8, no. 6, pp. 921-957
(Zhang&Balog, 2020) Zhang, S. & Balog, K. (2020). Web table extraction, retrieval, and augmentation: a survey. ACM Trans Intell Syst Technol, 11

Semantic Table Interpretation

- **Challenge:** To support both entity-focused and multidimensional tables

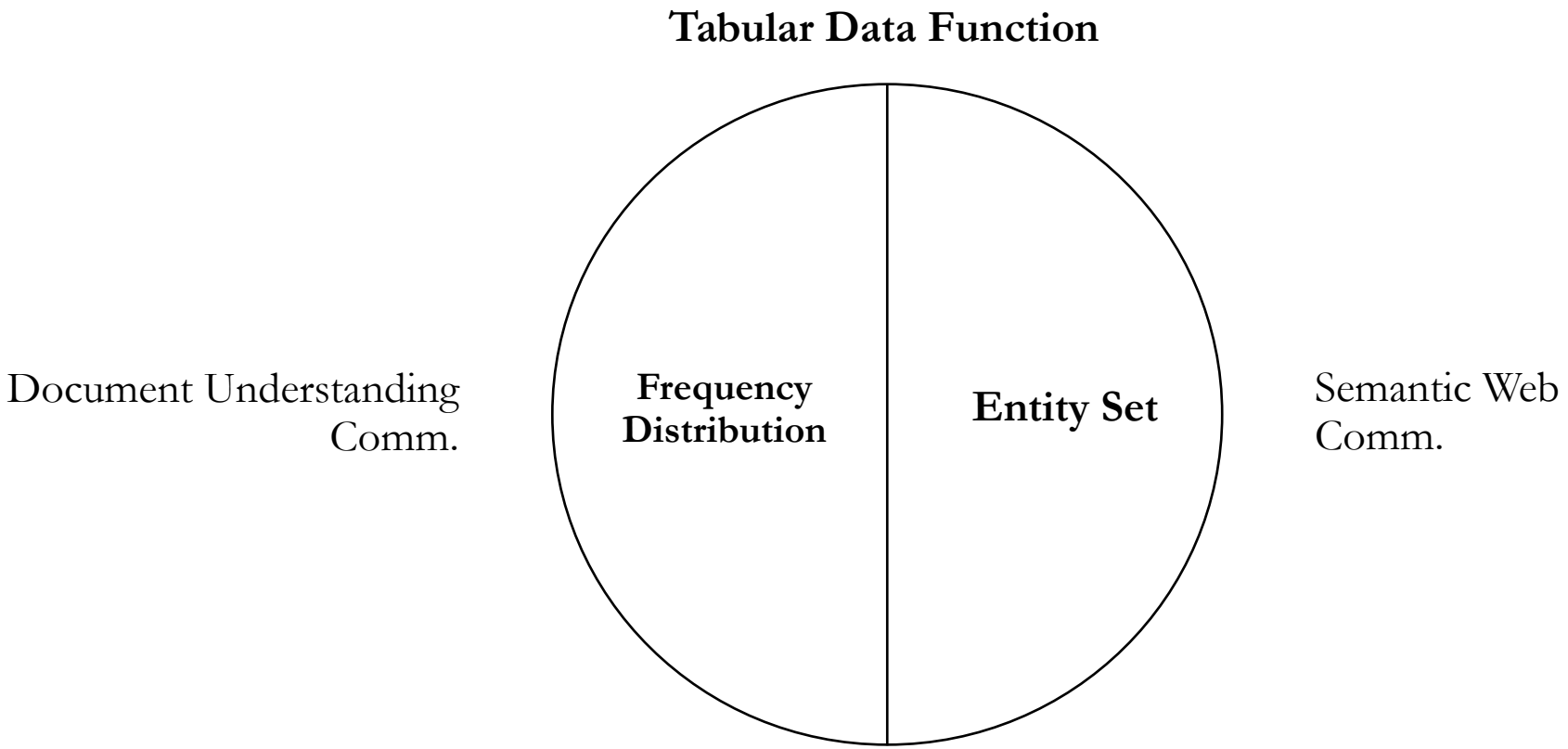


Table as an Entity Set

- **Entity Set**
  - **Relational table** — "*view of the relation*" (Embley, 2018; Johnston, 2014)
  - ER-model (Chen, 1976) defines that a tuple presents an entity, "*a thing of the real-world*"

President	#	Party	Election	End of Presidency	Vice President
Donald Trump	45	Republican	2016	Incumbent	Mike Pence
Barack Obama	44	Democratic	2008	January 20, 2017	Joe Biden
George Bush	43	Republican	2000	January 20, 2009	Dick Cheney

An entity set (U.S. presidents) as a relational table with the entity column 'President'

  - **Entity Column** is an approximate key containing names of entities (Lehmberg, 2016)
  - Entity Column is similarly defined in terms of ER-model in (Ouassarah, 2016)
  - **Entity-focused table** (Zhang&Balog, 2017)

(Chen, 1976) Chen, P. (1976). The entity-relationship model — toward a unified view of data. ACM Transactions on Database Systems. 1(1) 9-36
(Johnston, 2014) Johnston, T. (2014). Chapter 3 - the relational paradigm: mathematics. In: Johnston T. (eds) Bitemporal Data, 35-41
(Lehmberg, 2016) Lehmberg, O. & Bizer, C. (2016). Web table column categorization and profiling. Proc. 19th Int. W. Web and Databases, Article 4
(Ouassarah, 2016) Ouassarah, A. A. (2016) ADI: A NoSQL system for bi-temporal databases. Université de Lyon. PhD thesis
(Zhang&Balog, 2017) Zhang, S. & Balog, K. (2017). EntiTables: Smart assistance for entity-focused tables. Proc. 40th Int. ACM SIGIR Conf. Res. and Dev. in Inf. Retr. 255-264

Entity Set to Knowledge Graph

- **Existing methods of the Semantic Table Interpretation require**
  - Entity Column (Zhang&Balog, 2020)
  - To be in 3NF (Braunschweig, 2015)

Single-Concept Table (Braunschweig, 2015)
A tuple presents an instance of one concept

Multi-Concept Table (Braunschweig, 2015)
A tuple matches instances of few concepts

City	Mayor	Elevation(m)	Country	Area (km²)	Population (Mio.)
London	B. Johnson	35	England	130,395	53
Manchester	N. u. Hassan	38	England	130,395	53
Dublin	O. Quinn	-	Ireland	70,273	4.59
Berlin	K. Wowereit	34	Germany	357,021	80.5
Paris	B. Delanoë	35	France	551,695	63.4

City

Name	Mayor	Elevation(m)	Country
London	B. Johnson	35	England
Manchester	N. u. Hassan	38	England
Dublin	O. Quinn	-	Ireland
Berlin	K. Wowereit	34	Germany
Paris	B. Delanoë	35	France

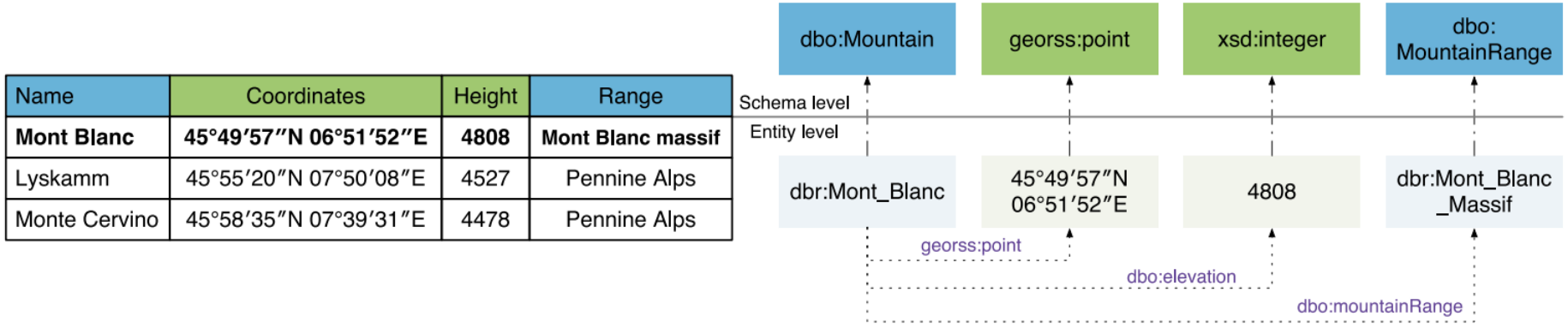
Country

Name	Area (km²)	Population (Mio.)
England	130,395	53
Ireland	70,273	4.59
Germany	357,021	80.5
France	551,695	63.4

"Example of a multi-concept table, with two corresponding single-concept tables" — copied from (Braunschweig, 2015)

(Braunschweig, 2015) Braunschweig, K., et al. (2015). From Web Tables to Concepts: A Semantic Normalization Approach. Conceptual Modeling. ER 2015. 9381 LNCS, 247-260
(Zhang&Balog, 2020) Zhang, S. & Balog, K. (2020). Web table extraction, retrieval, and augmentation: a survey. ACM Trans Intell Syst Technol, 11

Entity Set to Knowledge Graph



Semantic Table Annotation (Entity Set Case) — copied from (Cremaschi, 2020)

How to map an Entity Set to a Knowledge Graph

Cell-Entity-Annotation	Entity mention	→	KG-instance
Column-Type-Annotation	Column of values	→	KG-class or KG-datatype
Column-Property-Annotation	Pair of columns (E,P), where E is the entity column, P is any other	→	KG-property

(Cremaschi, 2020) Cremaschi, M., et al. (2020). A fully automated approach to a complete Semantic Table Interpretation, Future Generation Computer Systems, 112, 478-500

Table as a Frequency Distribution

- **A multi-dimensional table with the hierarchical data model** (Wang, 1996; Hurst, 2006)
  - Category → Subcategory → Sub-Subcategory → …
  - Frequency distribution

The average marks for 1991-1992

	Assignments			**Examinations**		Grade
	A1	A2	A3	Midterm	Final	
1991						
Winter	85	80	75	60	75	75
Spring	80	65	75	60	70	70
Fall	80	85	75	55	80	75
1992						
Winter	85	80	70	70	75	75
Spring	80	80	70	70	75	56
Fall	75	70	65	60	80	70

An example of a table from (Wang, 1996)

- Xinxin Wang declared this table has
- **3 categories of top-level**

  YEAR = {1991, 1992}
  TERM = {Winter, Spring, Fall}
  MARK = {ASSIGNMENTS, EXAMINATIONS, Grade}

- MARK **consists of 2 subcategories**

  ASSIGNMENTS = {A1, A2, A3}
  EXAMINATIONS = {Midterm, Final}

(Hurst, 2006) Hurst, M. (2006). Towards a theory of tables. IJDAR, 8, 123-131
(Wang, 1996) Wang, X. (1996). Tabular abstraction, editing, and formatting. University of Waterloo, PhD Thesis

Frequency Distribution to Knowledge Graph

- **From the hierarchical (Wang's model) to the relational data model: subcategories treated as data**

The average marks for 1991-1992

YEAR	TERM	Assignments			Examinations		Grade
		A1	A12	A13	Midterm	Final	
1991	Winter	85	80	75	60	75	75
	Spring	80	65	75	60	70	70
	Fall	80	85	75	55	80	75
1992	Winter	85	80	70	70	75	75
	Spring	80	80	70	70	75	56
	Fall	75	70	65	60	80	70

An example of a table from (Wang, 1996)

YEAR	TERM	UNNAMED (SUBCATEGORY)	UNNAMED (LABEL)	AVERAGE MARK
1991	Winter	ASSIGNMENTS	A1	85
1991	Winter	ASSIGNMENTS	A2	80
1991	Winter	ASSIGNMENTS	A3	75
1991	Winter	EXAMINATIONS	Midterm	60
1991	Winter	EXAMINATIONS	Final	75
1991	Winter	-	Grade	75
…	…	…	…	…

(Wang, 1996) Wang, X. (1996). Tabular abstraction, editing, and formatting. University of Waterloo, PhD Thesis

Frequency Distribution to Knowledge Graph

- **Multi-dimensional tables for frequency distribution**
  - **No entity columns;** A candidate key is a super key **(*A*)**
  - **Hierarchical case: A table (a relation in 1NF) joins a few relations in 3NF (*B*)**
    <YEAR,TERM,UNNAMED(LABEL)> -> <UNNAMED(SUBCATEGORY)> & <UNNAMED(LABEL)> -> <UNNAMED(SUBCATEGORY)>

YEAR	TERM	UNNAMED (SUBCATEGORY)	UNNAMED (LABEL)	AVERAGE MARK
1991	Winter	ASSIGNMENTS	A1	85
1991	Winter	ASSIGNMENTS	A2	80
1991	Winter	ASSIGNMENTS	A3	75
1991	Winter	EXAMINATIONS	Midterm	60
1991	Winter	EXAMINATIONS	Final	75
1991	Winter	-	Grade	75
…	…	…	…	…

A. *Wang's table before normalization*
(1 relation in 1NF)

FK

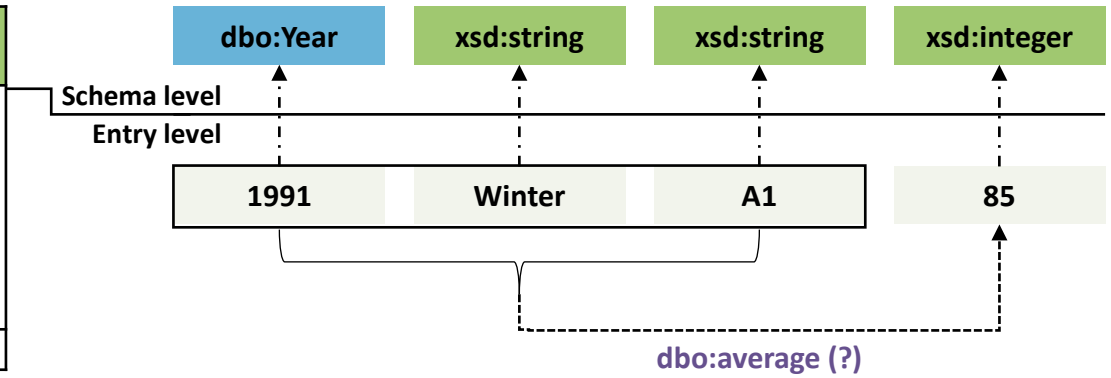
YEAR	TERM	UNNAMED (LABEL)	AVERAGE MARK
1991	Winter	A1	85
1991	Winter	A2	80
1991	Winter	A3	75
1991	Winter	Midterm	60
1991	Winter	Final	75
1991	Winter	Grade	75
…	…	…	…

UNNAMED (LABEL)	UNNAMED (SUBCATEGORY)
A1	ASSIGNMENTS
A2	ASSIGNMENTS
A3	ASSIGNMENTS
Midterm	EXAMINATIONS
Final	EXAMINATIONS
Grade	NONE

B. *Wang's table after normalization*
(2 relations in 3NF)

Frequency Distribution to Knowledge Graph

YEAR	TERM	UNNAMED (LABEL)	AVERAGE MARK
1991	**Winter**	**A1**	**85**
1991	Winter	A2	80
1991	Winter	A3	75
1991	Winter	Midterm	60
1991	Winter	Final	75
1991	Winter	Grade	75
…	…	…	…



Semantic Table Annotation (Frequency Distribution Case) — Hypothetically, It might look like this

How to map a Frequency Distribution to a Knowledge Graph

Cell-Entity-Annotation	Entity mention	→	KG-instance
Column-Type-Annotation	Column of values	→	KG-class or KG-datatype
Column-Dependent-Annotation	Functional dependency $<V_1,\ldots,V_n>$ -> F	→	KG-property (?)

Proposed Methods

- **Web Table Extraction**
  - Table type taxonomy: DWTC (Eberius, 2015)
  - DL-based table type classification: DeepTable (Habibi, 2020)
  - Rule-based data extraction from tables: TabbyXL (Shigarov, 2017)
  - Performance evaluation: TOMATE (Roldán, 2021)
- **Semantic Table Interpretation**
  - Entity lookup: DBpedia (Ritze, 2015)
  - Entity embedding: RDF2Vec (Ristoski, 2016), KGloVe (Cochez, 2017)
  - Hybrid (Efthymiou, 2017)
  - DL-based column type classification: ColNet (Chen, 2019)
  - Performance evaluation: SemTab (Jiménez-Ruiz, 2020)

(Chen, 2019) Chen, J., et al. (2019). ColNet: Embedding the semantics of web tables for column type prediction. Proc. 33rd AAAI Conference on Artificial Intelligence. 29-36
(Cochez, 2017) Cochez, M., et al. (2017). Global RDF vector space embeddings. The Semantic Web – ISWC 2017. 10587 LNCS 190-207
(Eberius, 2015) Eberius, J., et al. (2015). Eberius, J., et al. (2015). Building the Dresden web table corpus: a classification approach. Proc. IEEE/ACM 2nd Int. S. on Big Data Comp. 41-50
(Efthymiou, 2017) Efthymiou, V., et al. (2017). Matching web tables with knowledge base entities: from entity lookups to entity embeddings. ISWC 2017. 10587 LNCS, 260-277
(Habibi, 2020) Habibi, M., et al. (2020) DeepTable: a permutation invariant neural network for table orientation classification. Data Min Knowl Disc 34, 1963-1983
(Jiménez-Ruiz, 2020) Jiménez-Ruiz, E., at al. (2020). SemTab 2019: Resources to benchmark tabular data to knowledge graph matching systems. ESWC. LNCS 12123, 514-530
(Ristoski, 2016) Ristoski, P. & Paulheim, H. (2016). RDF2Vec: RDF graph embeddings for data mining. The Semantic Web – ISWC 2016. 9981 LNCS, 498-514
(Ritze, 2015) Ritze, D., et al. (2015). Matching HTML tables to DBpedia. WIMS '15. Article 10, 1-6
(Roldán, 2021) Roldán, J., et al. (2021). TOMATE: A heuristic-based approach to extract data from HTML tables. Information Sciences. 577, 49-68
(Shigarov, 2015) Shigarov, A. (2015). Table understanding using a rule engine. Expert Systems with Applications. 42(2), 929-937
(Shigarov, 2017) Shigarov, A. & Mikhailov, A. (2017). Rule-based spreadsheet data transformation from arbitrary to relational tables. Information Systems. 71, 123-136

Conclusion

- **SOTA in STI** showed on SemTab 2020
  - **CEA-task** — F_1 **= 0.907** (The best), F_1 **= 0.54** (TOP-10 average) — On natural tables (2T dataset)
  - **CTA-task** — F_1 **= 0.728** (The best), F_1 **= 0.59** (TOP-10 average) — On natural tables (2T dataset)
  - **CPA-task** — F_1 **= 0.93-0.97** (TOP-10 average) — On synthetic tables
- **Limitation** of STI solutions
  - Only entity-focused tables (Zhang&Balog, 2020)
  - Only single-concept tables (Braunschweig, 2015)
  - KG's low coverage (Zhang&Balog, 2020)
- **Goals**
  - To develop an end-to-end solution for KBP from web tables
    - **WTE**: To support atomic data items not cells
    - **STI**: To support both entity-focused tables and multidimensional ones
    - To publish as open source software
  - To participate in SemTab 2023 competitions

(Braunschweig, 2015) Braunschweig, K., et al. (2015). From Web Tables to Concepts: A Semantic Normalization Approach. Conceptual Modeling. ER 2015. 9381 LNCS, 247-260
(Zhang&Balog, 2020) Zhang, S. & Balog, K. (2020). Web table extraction, retrieval, and augmentation: a survey. ACM Trans Intell Syst Technol, 11

Thanks!