MantisTable

an automatic approach for the Semantic Table Interpretation

Marco Cremaschi, Roberto Avogadro, and David Chieregato
Department of Computer Science, Systems and Communication (DISCo)
University of Milano - Bicocca

Semantic Table Interpretation: an example

TABLE

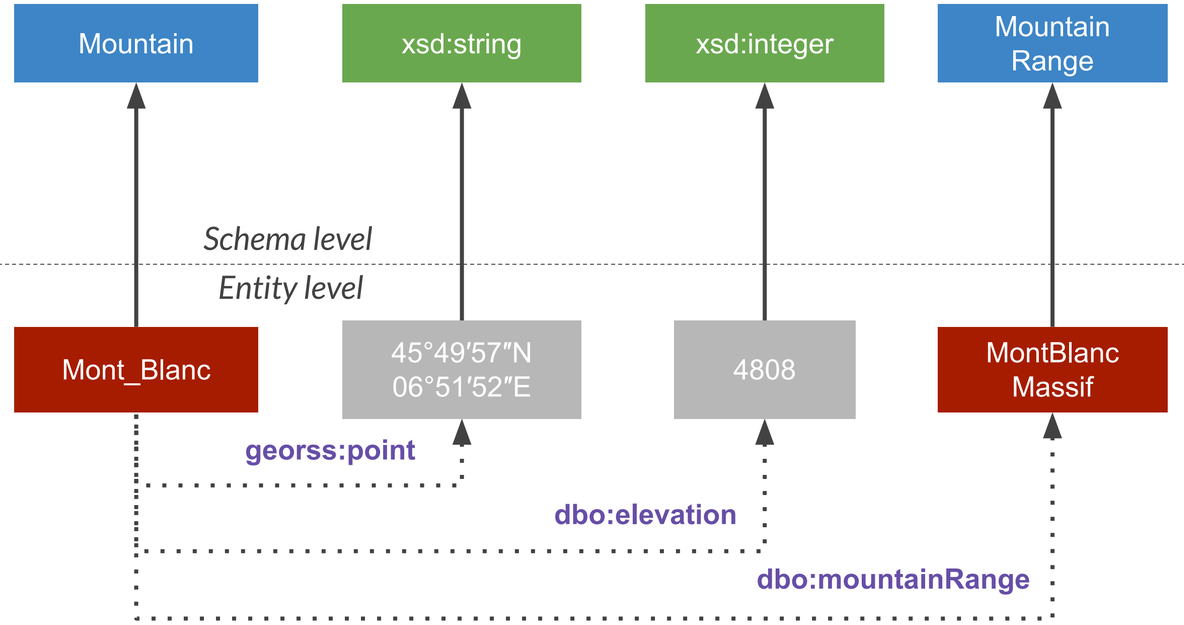
Name	Coordinates	Height	Range
Mont Blanc	**45°49′57″N 06°51′52″E**	**4808**	**Mont Blanc massif**
Lyskamm	45°55′20″N 07°50′08″E	4527	Pennine Alps
Monte Cervino	45°58′35″N 07°39′31″E	4478	Pennine Alps

Subject column (S-column)

Named-Entity column (NE-column)

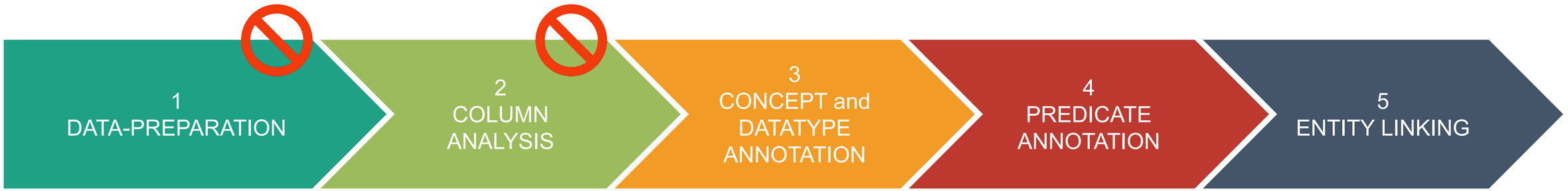
Literal column (L-column)

KNOWLEDGE GRAPH



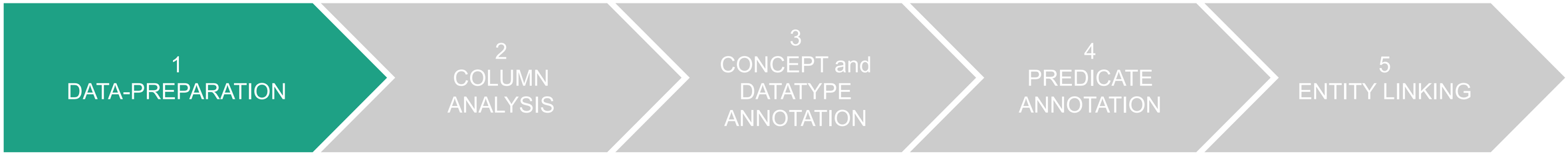
A **RDF* triple** is a subject, predicate, and object construct which makes data easily interlinked

Semantic Table Interpretation: enhanced approach (unsupervised, complete and automatic)



1. *Data Preparation*, which aims to prepare the data inside the table
2. *Column Analysis*, whose tasks are the semantic classification that assigns types to columns (NE-column or L-column), and the detection of the subject column (S-column)
3. *Concept and Datatype Annotation*, which deals with mappings between columns (or headers, if they are available) and semantic elements (concepts or datatypes) in a KG
4. *Predicate Annotation*, whose task is to find relations, in the form of predicates, between the main column and the other columns to set the overall meaning of the table
5. *Entity Linking*, which deals with mappings between cells and entities in a KG

Semantic Table Interpretation: enhanced approach (unsupervised, complete and automatic)

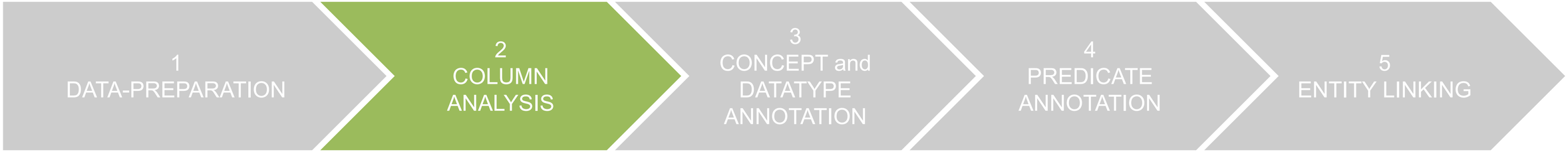


Data Preparation, which aims to prepare the data inside the table

Name	Coordinates	Height	Range
mont blanc	45°49′57″N 06° 51′52″E	4808	mont blanc massif
lyskamm	45°55′20″N 07°50′08″E	4527	pennine alps
monte cervino	45°58′35″N 07°39′31″E	4478	pennine alps

- removal of HTML tags and stop words
- transformation of the text into lowercase
- resolution of acronyms and abbreviation
- normalization of units of measurement by applying regular expressions

Semantic Table Interpretation: enhanced approach (unsupervised, complete and automatic)



Column Analysis, whose tasks are the semantic classification that assigns types to columns (NE-column or L-column), and the detection of the subject column (S-column)

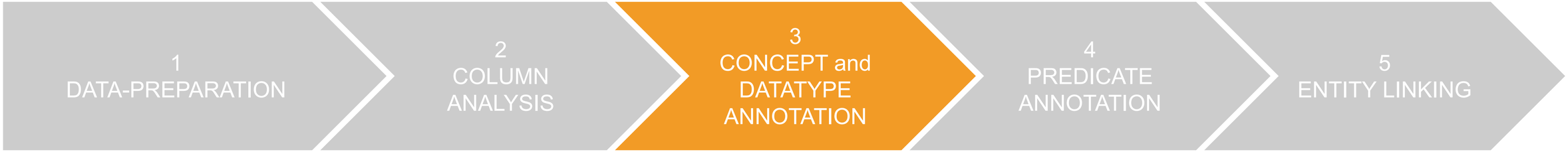
S-column *NE-column* *L-column*

Name	Coordinates	Height	Range
mont blanc	45°49′57″N 06° 51′52″E	4808	mont blanc massif
lyskamm	45°55′20″N 07°50′08″E	4527	pennine alps
monte cervino	45°58′35″N 07°39′31″E	4478	pennine alps

- Detection of L-columns by 16 regular expressions to identify regextype (e.g., geo coordinate, address, hex color code, URL)
- Detection of S-column considers different statistic features

$$subcol(c_j) = \frac{2uc_{norm}(c_j) + aw_{norm}(c_j) - emc_{norm}(c_j)}{\sqrt{df(c_j)+1}}$$

Semantic Table Interpretation: enhanced approach (unsupervised, complete and automatic)



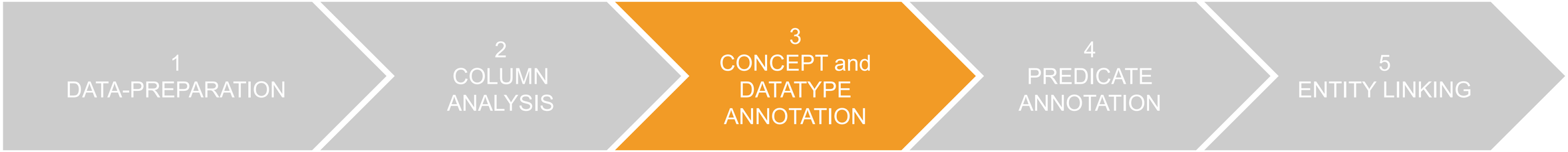
Concept and Datatype Annotation, which deals with mappings between columns (or headers, if they are available) and semantic elements (concepts or datatypes) in a KG

Name	Coordinates	Height	Range
mont blanc	45°49′57″N 06° 51′52″E	4808	mont blanc massif
lyskamm	45°55′20″N 07°50′08″E	4527	pennine alps
monte cervino	45°58′35″N 07°39′31″E	4478	pennine alps

MOUNTAIN PLACE HEIGHT MASSIF

- Retrieval of a set of candidate entities performing the entity-linking by searching the Knowledge Graph with the content of a cell
- Retrieval of abstract and concepts for each item in the set of retrieved entities
- Application of heuristics for the identification of the most frequent concept of the column

Semantic Table Interpretation: enhanced approach (unsupervised, complete and automatic)



Concept and Datatype Annotation, which deals with mappings between columns (or headers, if they are available) and semantic elements (concepts or datatypes) in a KG

$$econtext(e_{i,j}) = |bow(abstract(e_{i,j})) \cap bow(rcontent(i,j))| + |bow(abstract(e_{i,j})) \cap bow(hcontent(i,j))|$$

Abstract of the entity inside the KG

Row of the table

Header of the column

$$ename(e_{i,j}) = editDistance(tx(i,j), e_{i,j})$$

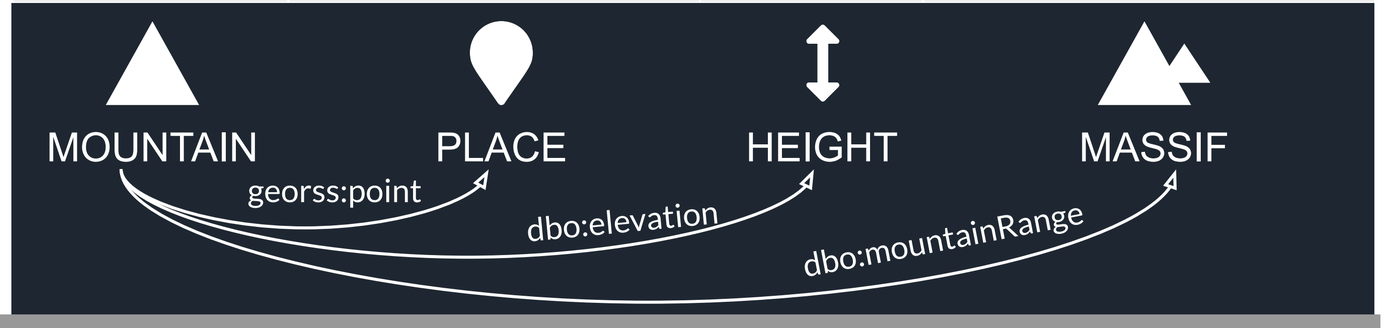
Text in the cell

Semantic Table Interpretation: enhanced approach (unsupervised, complete and automatic)

1 DATA-PREPARATION → 2 COLUMN ANALYSIS → 3 CONCEPT and DATATYPE ANNOTATION → **4 PREDICATE ANNOTATION** → 5 ENTITY LINKING

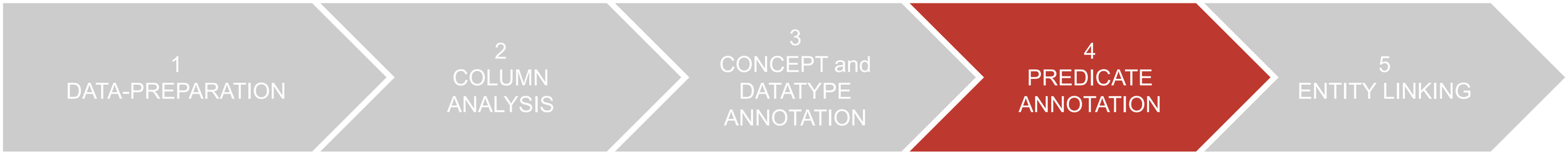
Predicate Annotation, whose task is to find relations, in the form of predicates, between the main column and the other columns to set the overall meaning of the table

Name	Coordinates	Height	Range
mont blanc	45°49′57″N 06° 51′52″E	4808	mont blanc massif
lyskamm	45°55′20″N 07°50′08″E	4527	pennine alps
monte cervino	45°58′35″N 07°39′31″E	4478	pennine alps



- The winning concept of the S-column are considered as the subject of the relationship and annotations of the other columns as objects
- The Knowledge Graph is searched for the subject and the object to collect possible predicates

Semantic Table Interpretation: enhanced approach (unsupervised, complete and automatic)



Predicate Annotation, whose task is to find relations, in the form of predicates, between the main column and the other columns to set the overall meaning of the table [Zhang 2017]

$$pcontext(p_j) = dice(p_j, x_j) = \frac{2 \cdot \sum_{w \in bowset(p_j) \cap bowset(x_j)} (freq(w, bow(p_j)) + freq(w, bow(x_j)))}{|bow(p_j)| + |bow(x_j)|}$$

Predicate Contexts

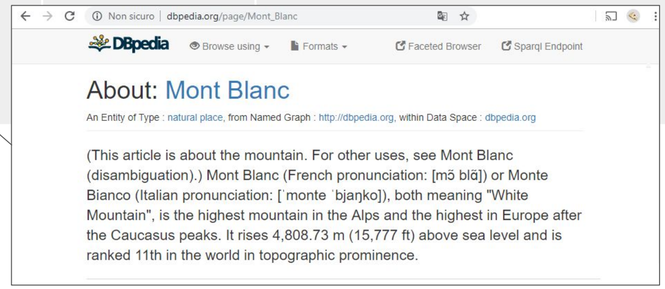
$$pfreq(p_j) = \frac{|p_j|}{\sum_j |p_j|}$$

Semantic Table Interpretation: enhanced approach (unsupervised, complete and automatic)

1 DATA-PREPARATION → 2 COLUMN ANALYSIS → 3 CONCEPT and DATATYPE ANNOTATION → 4 PREDICATE ANNOTATION → **5 ENTITY LINKING**

Entity Linking, which deals with mappings between cells and entities in a KG

Name	Coordinates	Height	Range
mont blanc dbr:Mont_Blanc	45°49′57″N 06°51′52″E	4808	mont blanc massif dbr:Mont_Blanc_massif
lyskamm dbr:Lyskamm	45°55′20″N 07°50′08″E	4527	pennine alps dbr:Pennine_Alps
monte cervino dbr:Monte_Cervino	45°58′35″N 07°39′31″E		



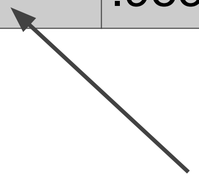
- Already discovered annotations are used to create a query for the disambiguation of the cell content
- If more than one entity is returned for a cell, the one with a smaller edit distance is taken

Semantic Table Interpretation: enhanced approach (unsupervised, complete and automatic)

CTA		
	Primary score	Secondary score
Round 1	.929	.933
Round 2	1.049	.247
Round 3	1.648	.269
Round 4	1.682	.322

CEA		
	Primary score	Secondary score
Round 1	1	1
Round 2	.614	.673
Round 3	.633	.679
Round 4	.973	.983

CPA		
	Primary score	Secondary score
Round 1	.965	.991
Round 2	.460	.544
Round 3	.518	.595
Round 4	.787	.841



Search for the path in the graph that links all the entities in the row

- Load tables in JSON format
- Download annotations (RDF/XML, N3, NTriples, Turtle and JSON-LD)
- Possibility to explore the output of each phase
- Manual annotation editing function
- Integration of the API provided by ABSTAT for auto-completion and suggestions

MANTISTABLE

Thank you

Marco Cremaschi
PhD Student@UNIMIB

marco.cremaschi@unimib.it