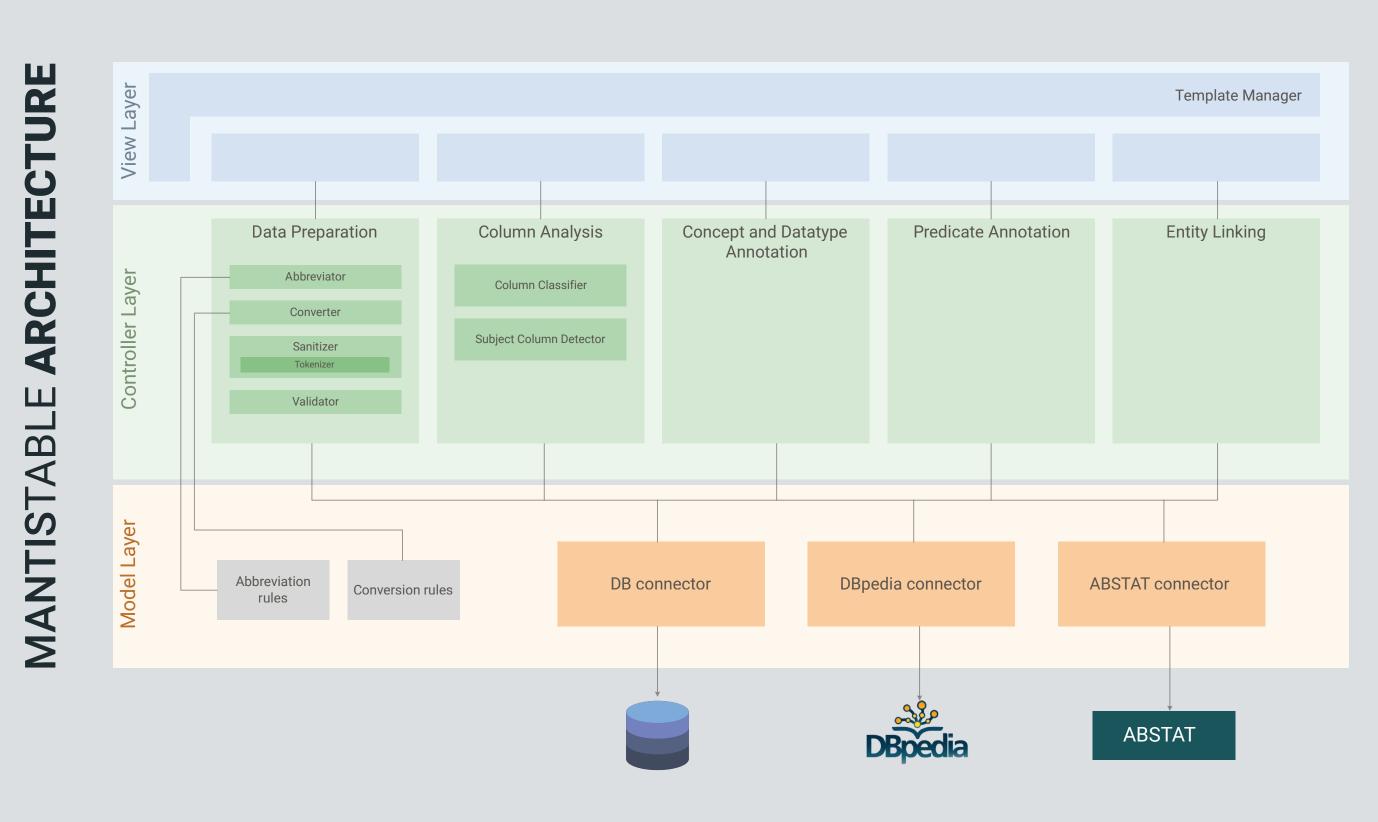






MantisTable is an open source Semantic Table Interpretation tool that automatically annotates, manages and makes accessible to humans and machines the **semantic of tables**.

45°49'58"N 6°51'54"E **Mont Blanc Massif** 4808 **Mont Blanc** 45°55'33"N 7°49'51"E Pennine Alps 4527 Lyskamm 4478 45°58'35.0"N 7°39'31"E Pennine Alps **Monte Cervino** xsd:string xsd:integer dbo:MountainRange dbo:Mountain dbo:mountainRange dbo:elevation



Mantis Table architecture is designed to be modular:

View Layer provides a graphic user interface to serve different types of tasks such as:

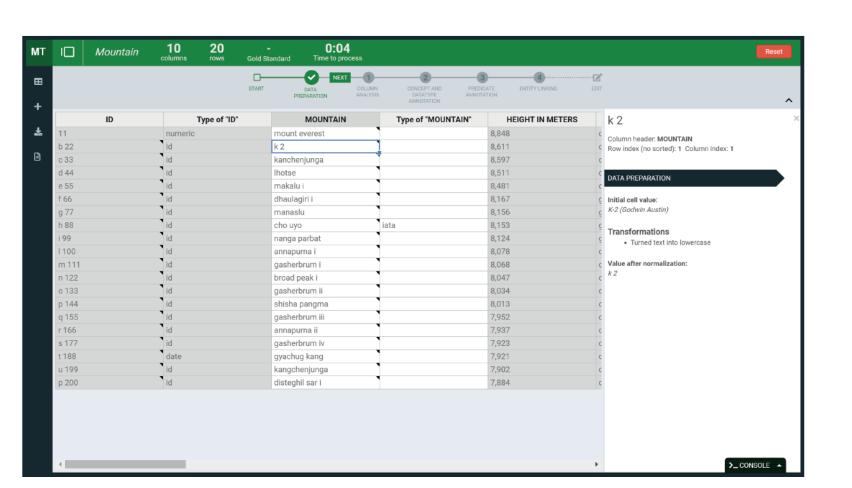
- storing and loading tables - execution of the STI steps
- exploration of the annotated tables
- annotations editing

Controller Layer creates all the abstraction between the View layer and the Model layer and implements all the STI steps.

Model Layer considers mainly data access for communicating with an application's data sources such as DB connector or DBpedia connector.

MANTISTABLE PROCESS





COLUMN ANALYSIS

	START	DATA COLUMN PREPARATION ANALYSIS	CONCEPT AND DATATIVE ANNOTATION	ENTITY LINKING EDIT			
ID NA	MOUNTAIN S		RANGE NE	CONQUERED ON L	MOUNTAIN		
1.1	mount everest	8,848	himalayas	may 29 1953			
b 22	k2	8,611	karakoram	july 31 1954	COLUMN ANALYSIS		
33	kanchenjunga	8,597	himalayas	wednesday 25 may 1955	Baseline Control of the Control of t		
d 44	Ihotse	8,511	himalayas	429926400	Column type: Named Entity		
55	makalu i	8,481	himalayas	may 15 1955			
66	dhaulagiri i	8,167	himalayas may 13 1960		Subject Column Detection		
77	manaslu	8,156	himalayas	may 9 1956	Fraction of empty cell: 0.0		
188	cho uyo	8,153	himalayas	oct 19 1954	Fraction of cell with unique content: 0.1		
99 nanga parbat 100 annapurna i		8,124	himalayas	july 3 1953	Distance from the first NE-column: 2.0		
		8,078 8,068	himalayas	june 3 1950	Average number of words: 1.0		
	n 111 gasherbrum i		karakoram	july 5 1958	Normalized values		
1 122 broad peak I		8,047	karakoram	June 9 1957	Fraction of empty cell: 0.0		
133	gasherbrum ii	8,034	karakoram	july 7 1956	Fraction of cell with unique content: 1.0		
144	shisha pangma	8,013	himalayas	may 2 1964	Distance from the first NE-column: 1.0		
155	gasherbrum iii	7,952	karakoram	aug 11 1975	Average number of words: 1.0		
166	annapurna ii	7,937	himalayas	may 17 1960			
					3.0		
			The state of the s		Final score		
200	disteghil sar i	7,884	karakoram june 9 1960				
s 177 t 188 u 199 p 200	gasherbrum iv gyachug kang kangchenjunga disteghil sar i	7,923 7,921 7,902 7,884	karakoram himalayas himalayas karakoram	aug 6 1958 apr 10 1964 may 26 1974 June 9 1960	3.0 Final score		

Columns are classified as named-entity column (NE-column) or literal column (L-column) and a subject column (S-column) is detected

- 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

CONCEPT AND DATATYPE ANNOTATION

	Mountain	10 20 rows	- Gold Standa	0:11 ard Time to process	s	li i			Reset	Ļ	
			START	DATA CO	OLUMN NALYSIS	CONCEPT AND DATATYPE ANNOTATION	PREDICATE ANNOTATION	ENTITY LINKING EDIT		٨	
1	ID NA	MOUNTAIN	s 🚹 🛗	HEIGHT IN METERS	LB	RANGE	NE 👩	CONQUERED ON L	mount everest	×	
11		mount everest	8,8	48		himalayas	Į.	may 29 1953	Column header: MOUNTAIN		
b 22		k2	8,6	11		karakoram		july 31 1954	Row index (no sorted): 0 Column index: 1		
c 33		kanchenjunga Ihotse makalu i		97		himalayas		wednesday 25 may 1955		A0000	
d 44				11		himalayas	30	429926400	CONCEPT AND DATATYPE ANNOTATION		
e 55				81		himalayas		may 15 1955	OUNCE TAILS BAIATH EARING TAILON	Allon	
f 66		dhaulagiri i	8,1	67		himalayas		may 13 1960	CANDIDATE ENTITIES CANDIDATE CONCEPTS		
g 77		manaslu	8,1	56		himalayas	2	may 9 1956	CANDIDATE ENTITIES CANDIDATE CONCEPTS		
h 88		cho uyo nanga parbat annapurna I		53		himalayas	2	oct 19 1954	http://dbpedia.org/resource/Mount_Everest_web		
199				24		himalayas	ì	july 3 1953	Edit Distance Score: 0.35	20000000000000000000000000000000000000	
1100				78		himalayas		June 3 1950	Row Context Score: 8 (Normalized: 0.222222222222222222222222222222222222		
m 111		gasherbrum i	8,0	68		karakoram		july 5 1958	Bonus Score: 0 Final score: 0.52222222222224		
n 122		broad peak i	8,0	47		karakoram		june 9 1957			
o 133		gasherbrum ii		34		karakoram		july 7 1956	http://dbpedia.org/resource/Mount_Everest		
p 144		shisha pangma	8,0	13		himalayas		may 2 1964	Edit Distance Score: 0 Row Context Score: 36 (Normalized: 1)		
q 155		gasherbrum iii		52		karakoram		aug 11 1975	Cell Context Score; 4 (Normalizad: 1) Bonus Score; 0		
r 166		annapurna ii	7,9	37		himalayas	- 8	may 17 1960	Final score: 2		
s 177		gasherbrum iv	7,9	23		karakoram		aug 6 1958	http://dhuadle.com/sono/stayet Forest Fore		
t 188		gyachug kang	7,9	21		himalayas		apr 10 1964	http://dbpedia.org/resource/Mount_Everest_Fou Edit Distance Score: 0.45833333333333333		
u 199		kangchenjunga	7,9	02		himalayas		may 26 1974	x3 Row Context Score: 6 (Normalized: 0.1666666666666666666666666666666666666)	
p 200	p 200	disteghil sar i		84	karakoram			june 9 1960	Bonus Score: 0		
Posterior			1.00						Final score: 0.250000000000001 http://dbpedia.org/resource/1996_Mount_Everes Edit Distance Score: 0.5185185185185185 x 4 Cell Context Score: 10 (Normalized: 1) Bonus Score: 0 Final score: 0.2407407407407407	B)	
(1)									http://dhnedia.org/resource/Fy		

Column headers are mapped to **semantic elements** (concepts or datatypes) of a **Knowledge Graph**

- Retrieval of a set of candidate entities performing the **entity-linking** by searching the Knowledge Graph with the content of a cell. The entity with the highest confident score is used to annotate the cell
- Extraction of the rdf:type values for each winning entity. The most frequent type is used to annotate the column

PREDICATE ANNOTATION

T I	I Mountain		old Standard TART	O:05 Time to process ODATA COLUM PREPARATION ANALYS	S DAT	PT AND ATYPE A	PREDICATE NINOTATION	NEXT BY ENTITY LINCONG EDIT			Reset
	ID NA	MOUNTAIN S	8 ≅ HE	IGHT IN METERS L	8 R	ANGE	NE 🖯	CONQUERED ON L	MOUNTAIN		
	11	mount everest	8,848		himala	/as		may 29 1953			
	b 22	k2	8,611		karako	am		july 31 1954	5		
	c 33	kanchenjunga	8,597		himala	/as		wednesday 25 may 1955	PREDICATE ANNOTA	TION	
	d 44	Ihotse	8,511		himala	/as		429926400	0/00/07/17/17		
	e 55	makalu i	8,481		himala	/as		may 15 1955	Subject	Predicate	Object
	f 66	dhaulagiri i	8,167		himala	/as		may 13 1960	Lancourrent		HEIGHT IN METERS xsd:integer
	g 77	manaslu	8,156		himala	/as		may 9 1956	MOUNTAIN Mountain	elevation	
	h 88	cho uyo	8,153		himala	/as		oct 19 1954			
	199	nanga parbat	8,124		himala	/as		July 3 1953	MOUNTAIN	mountainRange	RANGE MountainRange
	I 100	annapurna i	8,078		himala	/as		june 3 1950	Mountain		
	m 111	gasherbrum i	8,068		karako	am		july 5 1958	MOUNTAIN	firstAscentYear	CONQUERED O
	n 122	broad peak i	8,047		karako	am		june 9 1957	Mountain		
	o 133	gasherbrum ii	8,034		karako	am		july 7 1956	TANKET WE	in location	COORDINATES xsd:string
	p 144	shisha pangma	8,013		himala	/as		may 2 1964	MOUNTAIN Mountain		
	q 155	gasherbrum iii	7,952		karako	am		aug 11 1975			
	r166	annapurna ii	7,937		himala	/as		may 17 1960	MOUNTAIN Mountain	not found	BOOLEAN xsd:boolean
	s 177	gasherbrum iv	7,923		karako	am		aug 6 1958	Mountain		xsu.boolean
	t 188	gyachug kang	7,921		himala	/as		apr 10 1964	MOUNTAIN	isPrimaryTopicOf	URL xsd:anyURI
	u 199	kangchenjunga	7,902		himala	/as		may 26 1974	Mountain	710000	
	p 200	disteghil sar i	7,884		karako	am		june 9 1960	MOUNTAIN	not found	DESCRIPTION
								2	Mountain		xsd:string

Relations (predicates) between the subject column and the other columns are identified

- 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

ENTITY LINKING

The **content of cells** is mapped to entities in the **Knowledge Graph**

- 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

The data in the table are cleaned and uniformed

- Deletion of HTML tags and some characters
- Transformation of text into lowercase
- Deletion of text in brackets
- Explanation of acronyms and abbreviations using Oxford English Dictionary
- Normalisation of units of measurement applying regular expressions