Appendix

Preprocessing and Semantic Type Assignment: Detailed Workflow

Our semantic type detection pipeline employs a comprehensive preprocessing and mapping approach that transforms raw column headers into interpretable semantic annotations via the FinalFormat type system. This process involves a series of systematic steps, each designed to maximize robustness, flexibility, and explainability.

Our semantic type detection pipeline executes the following steps, closely mirroring the production implementation:

1. Initial Parsing and Metadata Extraction:

- Extract numeric prefixes (e.g., "1.ID") for traceability.
- Parse column names by splitting at colons, slashes, or parentheses to isolate core terms.
- Extract free-text descriptions from split points, or from parentheses if none exists, ensuring use of all available metadata.

2.Header Normalization:

- Convert all header tokens to lowercase.
- Remove or replace punctuation and special characters (underscores, hyphens, asterisks).
- Split compound words using spaces, dots, underscores, hyphens, or camelCase patterns (e.g., "AvgPitStopTime" → "avg pit stop time").
- Explicitly tokenize symbols (e.g., "%") for percentage detection.
- Match plural and singular forms (e.g., "chlorides" → "chloride").

3. Abbreviation and Variant Expansion

- Replace abbreviations using a dictionary of over 1000 entries (e.g., DOB → date of birth; bp → blood pressure.
- Apply regular expressions to tokenize and substitute complex patterns, including units (mg/dL, kg/m^2), and non-alphabetic tokens like ºC, cm^2, etc.

4. SourceKeywords Extraction

- Identify key tokens (SourceKeywords) that capture semantic meaning (e.g., "age", "date", "amount").
- Scan descriptions for known tokens, including "has", "is", or numeric/temporal keywords.
- Use primary/foreign key metadata to override assignments where appropriate.

5. Semantic Type Assignment (Rule-Based and Fallback Logic)

- Match tokens to a curated dictionary (over 2700 mappings) to assign one of 39 FinalFormat types.
- Apply priority logic:
    - "binary" overrides all if "has", "is", or "or not" is present.

    - “percentage” takes precedence if “%” or related terms detected.
    - Hierarchical rules resolve conflicts between categorical, numerical, and domain types.
    - Special logic for “name” columns in certain contexts (city, state, etc.).
- If no match is found:
    - Retry abbreviation expansion and check individual tokens (with plural handling).
    - If still unresolved, use ML-based prediction.

6. Ambiguity Resolution

- Use ML fallback (Random Forest, Logistic Regression, etc.) with weighted TF-IDF features.
- Assign confidence scores to all assignments for human-in-the-loop review.

7. Knowledge Graph Annotation

- Perform syntactic and semantic matching (FastText) between normalized headers and KG property/class labels.
- Use substring and similarity logic for best match selection (DBpedia, Schema.org).

8. Final Type Selection and Logging

- Consolidate assignments if column and description differ, prioritizing by semantic importance (e.g., prefer “datetime”, “IDcolumn”, etc.).
- Attribute assignments to specific source keywords for explainability.
- Assign “Nil” and log cases for manual review if no match is found.

Our semantic type detection pipeline applies a comprehensive sequence of normalization, expansion, and mapping steps that transform raw column headers into interpretable semantic annotations via the FinalFormat type system. The workflow (detailed above) enables robust, transparent, and flexible type assignment across real-world tabular data.

Practical Examples of Semantic Type Assignment

Table 8.1 below demonstrates, on a large and diverse set of real columns, how our pipeline’s dictionary logic, abbreviation expansion, and contextual rules deliver precise, explainable semantic types.

Key Interpretation Patterns and Highlights

Binary Columns:

Binary-type columns are often signaled by common words such as “has”, “is”, or phrases like “or not”. The pipeline detects these cues and classifies such columns under the binary FinalFormat, ensuring boolean attributes are annotated correctly even when explicit True/False values are absent.

Disambiguating Categorical Types:

Columns containing the word “range” (e.g., “price range”, “date range”) are accurately identified as categorical, despite potentially misleading co-occurring terms like “date” or “price” that might otherwise point toward temporal or financial types. This illustrates the nuanced, context-aware mapping the system achieves via dictionary-based disambiguation.

Geographic and Demographic Mapping:

Headers with words like “nationality”, “province”, or “capital” are consistently mapped to country, state, or city FinalFormats.

Synonym handling ensures that “ProvinceName” is linked to state, and “capital name” to city, even if these terms don’t match dictionary entries exactly.

Temporal and Timestamp Recognition:
Fields such as “updated”, “day of week”, and “timestamp” are mapped to their appropriate FinalFormats:
“updated” → date
“request timestamp” → datetime
“day of week” → weekday
This allows precise differentiation between temporal and generic date columns.

Financial and Monetary Attributes:
Lexical cues like “cash”, “currency”, “usd”, and “price” reliably trigger the assignment to the money FinalFormat, enabling wide coverage of financial information.

Communication and Identifier Columns:
Terms like “fax”, “cellular phone”, and “author channel id” are normalized to phone and IDcolumn, respectively, with common abbreviations (e.g., “nr” for number) mapped to numerical.

Scientific and Measurement Values:
Abbreviations and specialized terms, such as “wgt avg” (weight average) or “trestbps” (blood pressure), are interpreted using expansion rules, with assignments to numerical>=0 and bloodpressure, reflecting the system’s capacity to handle both generic and domain-specific semantics.

Web and Address Columns:
Patterns like “web address”, “link”, or “href” result in URLformat annotations, unifying various forms of URL-related fields.
Postal and address information, including “zipcode”, “provider zip code”, and “purchase address”, is reliably mapped to postalcode or street.

String and Free-Text Columns:
Terms such as “abstract”, “desc 1”, and “spec abstract” are interpreted as free text and assigned the string FinalFormat.

Other Notable Patterns:

- Medical/Health: Medical headers like “bp” or “trestbps” are mapped to bloodpressure.
- Percentage: Any header containing the “%” symbol or “percent” is assigned percentage.
- Temporal granularity: “Month”, “hour”, “week number”, “year” are all mapped to their corresponding temporal FinalFormats.

Through these patterns, Table 8.1 provides a transparent view of our pipeline’s ability to handle abbreviation, synonymy, compound phrases, and real-world header messiness, delivering precise, actionable semantic types for data quality assessment.

Table 8.1

Original Header	**Normalized**	**SourceKeywords**	**FinalFormat**
Members with age 5 - 17 years old	members with age 5 17 years old	age	age
Ankle_Ground_Angle	ankle ground angle	angle	angle
has_birth_date	has birth date	**has**	**binary**
IsBasedOnRealStory	is based on real story	**is**	**binary**
BookedHotelOrNot	booked hotel or not	**or not**	**binary**
bp	**bp**	**blood pressure**	bloodpressure
trestbps	**trestbps**	**blood pressure**	bloodpressure
Reported Influenza Activity	reported influenza activity	activity	categorical
Aircraft Manufacturer	aircraft manufacturer	aircraft	categorical
Pclass_1	pclass 1	class	categorical
ATTEND_DEPT	attend dept	department	categorical
Date Range*	**date range***	**range**	**categorical**
price_range	**price range**	**range**	**categorical**
Father's Birth Place	father's birth place	**birth place**	**city**
CapitalName	capital name	**capital**	**city**
Author, Country	author, country	country	country
team_one_player_one_nationality	team one player one nationality	**nationality**	**country**
BIRTHDATE	birthdate	birthdate	date
dtRef	**dt** ref	date	date
Data Last Updated	data last updated	**updated**	**date**
Order Date and Time	order date and time	date and time	datetime
Request timestamp	request timestamp	**timestamp**	**datetime**
school_day	school day	day	day
FLAG_EMAIL	flag email	email	E-mailformat
order_hour_of_day	order hour of day	hour	hour
authorChannelId	author channel id	id	IDcolumn
Nameserver IP Address	nameserver ip address	ip	IPformat
src_ip_country_code	src ip country code	ip	IPformat
vehicle_gps_latitude	vehicle gps latitude	latitude	latitude
Delivery_location_longitude	delivery location longitude	longitude	longitude
Cash amount	cash amount	**cash**	**money**
discount_price__currency	discount price currency	**currency**	**money**
cons.price.idx	cons price idx	**price**	**money**
Values in Billions USD	values in billions usd	**usd**	**money**
Date_Of_Death_Month	date of death month	month	month
Head Coach	head coach	**coach**	**name**
Artistic Director	artistic director	**director**	**name**
fullname words	fullname words	**fullname**	**name**
LASTNAME	lastname	lastname	name
Filename	filename	name	name
Person Baptised	person baptised	person	name
Allowed Amount	allowed amount	**amount**	**numerical**
Trip_Distance_km	trip distance km	**distance**	**numerical**
Ground Elevation	ground elevation	**elevation**	**numerical**
Flug-Nr.	flug nr	**number**	**numerical**
Deforestation_Area_Ha	deforestation area ha	**area**	**numerical>=0**
free_throw_attempts	free throw attempts	**attempts**	**numerical>=0**
ViewDepth	view depth	**depth**	**numerical>=0**
DirectoryEntryImportSize	directory entry import size	**size**	**numerical>=0**
Wgt. Avg.	wgt avg	**weight**	**numerical>=0**
YearsInCurrentRole	years in current role	**years**	**numerical>=0**
% of Total Supply Owned	% of total supply owned	**%**	**percentage**
Percent of State employment	percent of state employment	**percent**	**percentage**
Number of Cellular phone	number of cellular phone	cellular phone	phone
Fax Numbers	fax numbers	**fax**	**phone**
Customer_Postal_Code	customer postal code	postal code	postalcode
Provider Zip Code	provider zip code	**zip code**	**postalcode**
ProvinceName	province name	**province**	**state**
Purchase Address	purchase address	address	street
Spec. abstract	spec abstract	abstract	string
desc_1	desc 1	description	string
AvgPitStopTime	avg pit stop time	time	time
goods-title-link--jump href	goods title link jump href	**href**	**URLformat**
Free Download Link	free download link	**link**	**URLformat**
Web Address	web address	**web address**	**URLformat**
arrival_date_week_number	arrival date week number	week number	week

Day of Week	day of week	**day of week**	**weekday**
athlete_year_birth	athlete year birth	year	year

These examples illustrate how the system transforms raw, often noisy or abbreviated, column headers into precise semantic annotations. By combining normalization, robust abbreviation expansion, and hierarchical rule logic, our pipeline ensures reliable type assignment even in the face of ambiguity, domain-specific terminology, or non-standard header formats. This approach not only supports accurate data profiling but also underpins the effectiveness of our automated data quality assessment framework.

Full SemTab 2024 Metadata2KG [12] Track Analysis

Columns Identified as Correct After Human Review (43 items)

Column values	GT Annotation	Our Annotation	Items	Justification
Year	releaseDate	year	20	More accurate general temporal reference
ISO(2)/ISO(3)	iso31661Code	isoCode	2	General ISO code is semantically suitable
Length	duration	length	1	Physical dimension more appropriate
Government	governmentType	government	4	Precisely matches semantic context
State	location	state	4	Accurate geopolitical specificity
GDP	grossDomesticProduct	**gdp**PerCapita	1	Contextually appropriate economic indicator
Alphabetic**code**	currencyCode	code	4	Correct generalization without context
Granting**Institution**	almaMater	institution	1	Broader semantic applicability
Height(ft), **HEIGHT**INMETERS	elevation	height	5	More precise semantic choice for non-geographical context
Board	owner	board	1	Semantically matches governance/advisory body
		TOTAL	**43**	**30.5%**

Contextually Appropriate Columns (6 items)

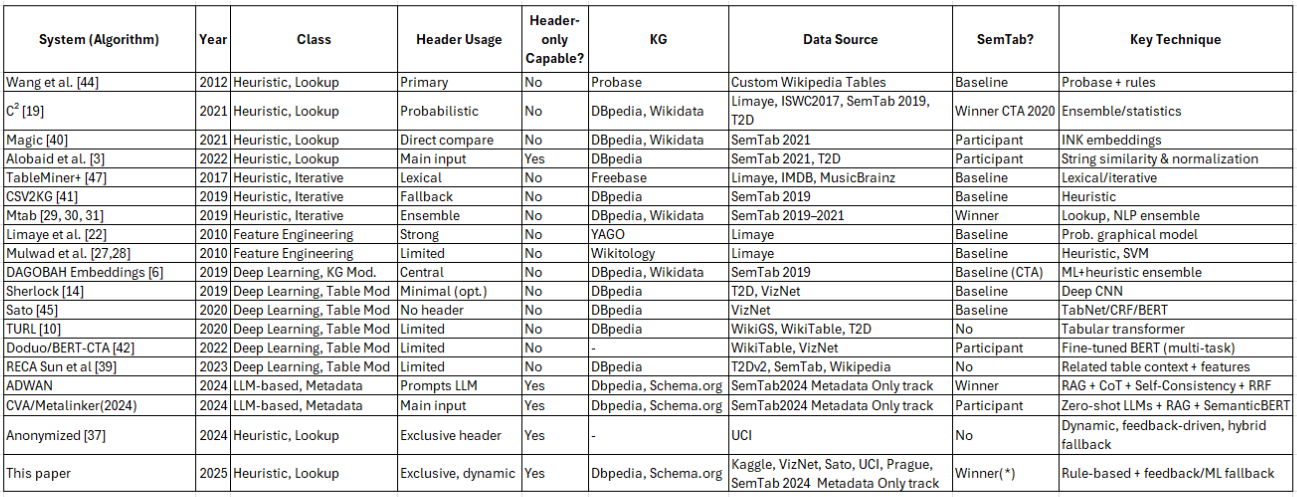
Column values	GT Annotation	Our Annotation	Items	Justification
Size	fileSize	collection**Size**	1	Semantically appropriate given context
Label	developer	distributing**Label**	1	Contextually plausible alternative
System	computingPlatform	**system**Requirements	3	Ambiguous header, plausible semantic choice
watergauge	elevation	water	1	Contextual understanding of water measurement
		TOTAL	**6**	**4%**

Incorrect Columns (24 items)

Column values	GT Annotation	Our Annotation	Items	Issue

Basincountries, Staat(en)	country	Nil	2	Multilingual/compound words issues
8.848, Feets, Meters, H?he	elevation	Nil	5	Numeric/unit headers lacking context
GNI, GDPnominal(US$M)	giniCoefficient, grossDomesticProduct	Nil	2	Abbreviations and economic indicators
Ashton-under-Lyne, Editeur, Endroit, Stadt	location, owner	Nil	5	Non-English terms, ambiguous without context
Europe, Japan, Rel.	releaseDate	Nil	5	Contextual mismatch with GT annotation
NorthAmerica	releaseDate	northWestPlace	1	Contextual mismatch with GT annotation
Rationale	knownFor	ratio	2	Semantic misunderstanding due to lexical similarity
1T	iataAirlineCode	Nil	1	headers lacking context
Max.-Tiefe-(m)	depth	max	1	headers lacking context
		TOTAL	**24**	**17%**

Table 5.1



System (Algorithm)	Year	Class	Header Usage	Header-only Capable?	KG	Data Source	SemTab?	Key Technique
Wang et al. [44]	2012	Heuristic, Lookup	Primary	No	Probase	Custom Wikipedia Tables	Baseline	Probase + rules
C^2 [19]	2021	Heuristic, Lookup	Probabilistic	No	DBpedia, Wikidata	Limaye, ISWC2017, SemTab 2019, T2D	Winner CTA 2020	Ensemble/statistics
Magic [40]	2021	Heuristic, Lookup	Direct compare	No	DBpedia, Wikidata	SemTab 2021	Participant	INK embeddings
Alobaid et al. [3]	2022	Heuristic, Lookup	Main input	Yes	DBpedia	SemTab 2021, T2D	Participant	String similarity & normalization
TableMiner+ [47]	2017	Heuristic, Iterative	Lexical	No	Freebase	Limaye, IMDB, MusicBrainz	Baseline	Lexical/iterative
CSV2KG [41]	2019	Heuristic, Iterative	Fallback	No	DBpedia	SemTab 2019	Baseline	Heuristic
Mtab [29, 30, 31]	2019	Heuristic, Iterative	Ensemble	No	DBpedia, Wikidata	SemTab 2019–2021	Winner	Lookup, NLP ensemble
Limaye et al. [22]	2010	Feature Engineering	Strong	No	YAGO	Limaye	Baseline	Prob. graphical model
Mulwad et al. [27,28]	2010	Feature Engineering	Limited	No	Wikitology	Limaye	Baseline	Heuristic, SVM
DAGOBAH Embeddings [6]	2019	Deep Learning, KG Mod.	Central	No	DBpedia, Wikidata	SemTab 2019	Baseline (CTA)	ML+heuristic ensemble
Sherlock [14]	2019	Deep Learning, Table Mod	Minimal (opt.)	No	DBpedia	T2D, VizNet	Baseline	Deep CNN
Sato [45]	2020	Deep Learning, Table Mod	No header	No	DBpedia	VizNet	Baseline	TabNet/CRF/BERT
TURL [10]	2020	Deep Learning, Table Mod	Limited	No	DBpedia	WikiGS, WikiTable, T2D	No	Tabular transformer
Doduo/BERT-CTA [42]	2022	Deep Learning, Table Mod	Limited	No	-	WikiTable, VizNet	Participant	Fine-tuned BERT (multi-task)
RECA Sun et al [39]	2023	Deep Learning, Table Mod	Limited	No	DBpedia	T2Dv2, SemTab, Wikipedia	No	Related table context + features
ADWAN	2024	LLM-based, Metadata	Prompts LLM	Yes	Dbpedia, Schema.org	SemTab2024 Metadata Only track	Winner	RAG + CoT + Self-Consistency + RRF
CVA/Metalinker(2024)	2024	LLM-based, Metadata	Main input	Yes	Dbpedia, Schema.org	SemTab2024 Metadata Only track	Participant	Zero-shot LLMs + RAG + SemanticBERT
Anonymized [37]	2024	Heuristic, Lookup	Exclusive header	Yes	-	UCI	No	Dynamic, feedback-driven, hybrid fallback
This paper	2025	Heuristic, Lookup	Exclusive, dynamic	Yes	Dbpedia, Schema.org	Kaggle, VizNet, Sato, UCI, Prague, SemTab 2024 Metadata Only track	Winner(*)	Rule-based + feedback/ML fallback