APPENDICES

Appendix 1 - Summary of Well-known Data Quality Issues and the Data Quality Dimension Violations Associated

Table 1 - Data Quality Issues and Data Quality Dimension Violations (Adapted from Visengeriyeva and Abedjan, 2020[31])

#	Data Quality Issue	Description	Data Quality Dimensions
1	Missing data	Comprises missing tuples and missing values. Tuple completeness requires that all tuples are present in the table. Missing value issue consists of either null values or disguised values. Value completeness requires that all values are present in the table, while null values indicate that the value is unknown or non-existent.	Accuracy, Completeness
2	Incorrect data	Data that differ from the values of the real-world entity (e.g., wrong date of birth).	Accuracy
3	Misspellings	Syntactic deviation of the data value from its ground truth (e.g., "Smiht" instead of "Smith").	Accuracy
4	Ambiguous data	Data values which might be interpreted in several ways (e.g., abbreviations or cryptic values).	Accuracy, Consistency
5	Extraneous data	Presence of additional data in the attribute value (e.g., the address column contains a person's name in addition to the address).	Consistency, Uniqueness
6	Outdated temporal data	Values that are obsolete or outdated.	Timeliness
7	Misfielded values	Values that are placed inside the wrong attribute.	Accuracy, Consistency, Completeness
8	Incorrect references	Entities that contain wrong information concerning the reference relation (e.g., the employee is associated with a wrong department).	Accuracy
9	Duplicates	Tuples/values that represent the same real-world entity.	Uniqueness
10	Structural conflicts	Conflicting duplicates in different sources.	Consistency, Uniqueness
11	Different word orderings	Values that violate the expected word order (e.g., first name precedes last name).	Consistency, Uniqueness
12	Different aggregation levels	Entities produced by applying different aggregation methods (e.g., entries per quartal vs. entries per year).	Accuracy, Consistency
13	Temporal mismatch	Refers to erroneous data that arise due to non-enforcement of integrity constraints for temporal data.	Accuracy, Timeliness
14	Different units/representations	Occurrence of multiple representations for the same concept (e.g., Price in different currencies).	Consistency
15	Domain violation	Values that violate semantic rules defined on the specific attribute.	Accuracy
16	FD violation	Values that violate previously specified functional dependencies.	Accuracy, Consistency
17	Wrong data type	Values that violate the data type specification of the corresponding attribute, i.e., data type constraint violation.	Consistency
18	Referential integrity violation	Tuples that violate the referential integrity constraints defined on multiple relations (e.g., missing foreign key).	Accuracy, Consistency, Completeness
19	Uniqueness violation	Duplication of values under the uniqueness constraint.	Uniqueness
20	Use of synonyms	Occurrence of synonymous representations for the same concept inside the same column (e.g., "lecturer" and "professor").	Uniqueness
21	Use of special characters (space, no space, dash, parentheses)	Refers to different representations of compound data, such as Social Security Number or phone number.	Consistency
22	Different encoding formats	Inconsistent usage of encodings for values within a dataset (e.g., ASCII or EBCDIC).	Consistency

Appendix 2 - Explanation for selection of 50 database tables from the Prague Relational Learning Repository [21]

These are the steps taken to obtain 50 database tables from the Prague Relational Learning Repository.

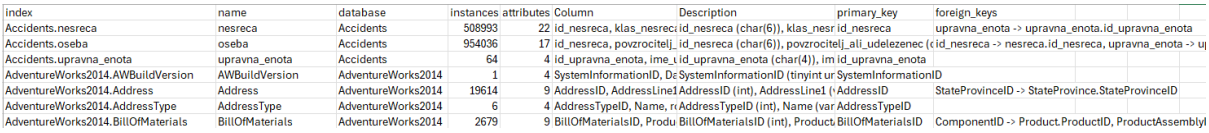
1. The paper that described the Repository (**Motl & Schulte (2024) – [21])** had this Table 1:

Database	#Relations	#Instances	Size	Type	Domain	Task
Accidents [18]	4	495760	210.0	Real	Goverment	Class
AdventureWorks	71	30669	234.6	Synth	Retail	Regr
AustralianFootball	4	3036	38.0	Real	Sport	Class
Biodegradability [3]	5	328	3.2	Real	Medicine	Regr
Carcinogenesis [24]	6	329	26.3	Real	Medicine	Class
CCS	6	1000	658.4	Real	Finance	Regr
ClassicModels	8	273	0.5	Synth	Retail	Regr
Countries	4	247	8.6	Real	Geography	Regr
Credit	9	10084	443.6	Synth	Retail	Class
CS	8	100	0.3	Synth	Finance	Class
Dunur [11]	20	276	0.8	Real	Kinship	Class
Elti [11]	14	1081	0.7	Real	Kinship	Class
Employee	7	2838426	344.6	Synth	Retail	Regr
Financial [2]	8	682	94.1	Real	Finance	Class
FTP	2	29555	7.5	Synth	Retail	Class
Genes [6]	3	862	1.9	Real	Medicine	Class
Hepatitis	7	500	2.2	Real	Medicine	Class
Hockey [23]	23	7759	15.5	Real	Sport	Class
IMDb [23]	7	794625	614.6	Real	Entertainment	Class
MovieLens [22]	7	6039	151.9	Real	Entertainment	Class
Lahman	25	23111	84.0	Real	Sport	Regr
LegalActs	5	564268	238.2	Real	Gouverment	Class
Mesh [9]	32	223	1.1	Real	Industry	Regr
Mondial [14]	33	454	3.3	Real	Geography	Class
MooneyFamily [20]	72	92	3.3	Synth	Kinship	Class
Mutagenesis [8]	3	188	0.9	Real	Medicine	Class
Nations	3	14	2.1	Real	Geography	Class
NBA [23]	4	30	0.3	Real	Sport	Class
NCAA	10	268	40.6	Real	Sport	Class
Northwind	29	830	1.1	Synth	Retail	Regr
Pima	14	768	0.8	Real	Medicine	Class
PremiereLeague [19]	4	363	11.3	Real	Sport	Class
PTE [25]	41	299	7.3	Real	Medicine	Class
Pubs	11	18	0.4	Synth	Retail	Regr
Sakila	16	15991	6.6	Synth	Retail	Regr
SalesDB	4	6148886	539.3	Synth	Retail	Regr
SameGen	7	1081	0.3	Real	Kinship	Class
Stats	8	38357	621.4	Real	Education	Regr
StudentLoan [17]	13	1000	0.9	Real	Education	Class
PTC [10]	4	343	7.8	Real	Medicine	Class
Thrombosis [7]	3	806	1.9	Real	Medicine	Class
TPCC [12]	9	28433	174.1	Synth	Retail	Class
TPCDS [16]	24	99550	4587.5	Synth	Retail	Class
TPCH [4]	8	148255	1925.1	Synth	Retail	Regr
Trains [15]	2	20	0.1	Synth	Logistic	Class
University	5	38	0.3	Synth	Education	Class
UW-CSE [21]	4	278	0.2	Real	Education	Class
VOC	8	8215	2.7	Real	Logistic	Class
World	3	239	0.8	Real	Geography	Class

Table 1: List of databases in the repository

In it we can see, in the first line, important information that allow us to choose the Databases.
There are 49 Databases described here, some with numbers that point to research papers associated.
There are the #Relations, that show the number of tables in each database.
There are the #Instances, which show the count of rows in the target table in each database.
There is the Size in MB including indexes.
There is the Type, Real or Synth (synthetically generated).
There are many different Domains.
And there are two possible kinds of Tasks: Regr(ession) and Class(ification).

2. The next step was to take a look at the tables themselves at the Prague Relational Learning Repository. We used MySQL Workbench to connect with the following parameters described at [21]:
   Hostname: db.relational-data.org
   Port: 3306
   Username: guest
   Password: relational
3. After we understood the data we decided to extract the metadata, as we had done previously from UCI, including new information about Primary and Foreign Keys which are available as the tables are relational.
4. The next step was to download the information available then. There were over 2000 tables at that time. We just wanted 50.
5. We brought the following information, similar to what was brought earlier from all the 622 datasets of the UCI Machine Learning Repository:



index	name	database	instances	attributes	Column	Description	primary_key	foreign_keys
Accidents.nesreca	nesreca	Accidents	508993	22	id_nesreca, klas_nesreci	id_nesreca (char(6)), klas_nesr	id_nesreca	upravna_enota -> upravna_enota.id_upravna_enota
Accidents.oseba	oseba	Accidents	954036	17	id_nesreca, povzrocitelj_	id_nesreca (char(6)), povzrocitelj_ali_udelezenec (		id_nesreca -> nesreca.id_nesreca, upravna_enota -> u
Accidents.upravna_enota	upravna_enota	Accidents	64	4	id_upravna_enota, ime_	id_upravna_enota (char(4)), im	id_upravna_enota	
AdventureWorks2014.AWBuildVersion	AWBuildVersion	AdventureWorks2014	1	4	SystemInformationID, Da	SystemInformationID (tinyint ur	SystemInformationID	
AdventureWorks2014.Address	Address	AdventureWorks2014	19614	9	AddressID, AddressLine1	AddressID (int), AddressLine1 (\	AddressID	StateProvinceID -> StateProvince.StateProvinceID
AdventureWorks2014.AddressType	AddressType	AdventureWorks2014	6	4	AddressTypeID, Name, r	AddressTypeID (int), Name (var	AddressTypeID	
AdventureWorks2014.BillOfMaterials	BillOfMaterials	AdventureWorks2014	2679	9	BillOfMaterialsID, Produ	BillOfMaterialsID (int), Product	BillOfMaterialsID	ComponentID -> Product.ProductID, ProductAssemblyI

   a. The index, with the content of the database followed by a dot and the table
   b. The name of the table
   c. The name of the database
   d. The number of Instances/records in this table
   e. The number of attributes/columns in this table
   f. The columns in this table
   g. The description, where the columns contain the format associated
   h. The primary_key(s) in this table
   i. The foreign_key(s) in this table

6. From Table 1 of paper [21] we checked if the databases existed in the current repository. Unfortunately, some databases did not exist anymore or had different names at the time.
7. These were the 3 databases not found then: PTC, Thrombosis and VOC.
8. Analyzing the content of all different Domains it was discovered that the Kinship Domain with 4 databases, was not suitable for our research, because all their tables contain only few columns with information about names of relatives, such as brother, father, etc., so nothing usable for our analysis:

Dunur.brother

name1	name2
Alfonso	Sophia
Arthur	Victoria
Colin	Charlotte
Emilio	Lucia
James	Jennifer
Marco	Angela

Dunur.husband

name1	name2
Andrew	Christine
Arthur	Margaret
Charles	Jennifer
Christopher	Penelope
Emilio	Gina

Elti.target

name1	name2	is_elti
ali1	ali2	0
ali1	alp	0
ali1	anil	0
ali1	ayse	0
ali1	ayten	0

9. So, the initial total of 49 databases became 49 – 3 (not found) – 4 (Kinship) = 42 databases.
10. See below the list of the 49 databases, with the Current Name and the Name in Table 1:

Current Name	Name in Table 1	#Relations	#Instances	Size(MB)	Type	Domain	Task
Accidents	Accidents	4	495760	210	Real	Government	Class
AdventureWorks2014	AdventureWorks	71	30669	234.6	Synth	Retail	Regr
AustralianFootball	AustralianFootball	4	3036	38	Real	Sport	Class
Biodegradability	Biodegradability	5	328	3.2	Real	Medicine	Regr
Carcinogenesis	Carcinogenesis	6	329	26.3	Real	Medicine	Class
ccs	CCS	6	1000	658.4	Real	Finance	Regr
classicmodels	ClassicModels	8	273	0.5	Synth	Retail	Regr
Countries	Countries	4	247	8.6	Real	Geography	Regr
Credit	Credit	9	10084	443.6	Synth	Retail	Class
cs	CS	8	100	0.3	Synth	Finance	Class
Dunur	Dunur	20	276	0.8	Real	Kinship	Class
Elti	Elti	14	1081	0.7	Real	Kinship	Class
employee	Employee	7	2838426	344.6	Synth	Retail	Regr
financial	Financial	8	682	94.1	Real	Finance	Class
ftp	FTP	2	29555	7.5	Synth	Retail	Class
genes	Genes	3	862	1.9	Real	Medicine	Class
Hepatitis_std	Hepatitis	7	500	2.2	Real	Medicine	Class
Hockey	Hockey	23	7759	15.5	Real	Sport	Class
imdb_full	IMDb	7	794625	614.6	Real	Entertainment	Class
imdb_MovieLens	MovieLens	7	6039	151.9	Real	Entertainment	Class
lahman_2014	Lahman	25	23111	84	Real	Sport	Regr
legalActs	LegalActs	5	564268	238.2	Real	Government	Class
Mesh	Mesh	32	223	1.1	Real	Industry	Regr
Mondial	Mondial	33	454	3.3	Real	Geography	Class
MooneyFamily	MooneyFamily	72	92	3.3	Synth	Kinship	Class
mutagenesis	Mutagenesis	3	188	0.9	Real	Medicine	Class
nations	Nations	3	14	2.1	Real	Geography	Class
NBA	NBA	4	30	0.3	Real	Sport	Class
NCAA	NCAA	10	268	40.6	Real	Sport	Class
northwind	Northwind	29	830	1.1	Synth	Retail	Regr
Pima	Pima	14	768	0.8	Real	Medicine	Class
PremierLeague	PremiereLeague	4	363	11.3	Real	Sport	Class
Database not found	PTC	4	343	7.8	Real	Medicine	Class
PTE	PTE	41	299	7.3	Real	Medicine	Class
pubs	Pubs	11	18	0.4	Synth	Retail	Regr
sakila	Sakila	16	15991	6.6	Synth	Retail	Regr
SalesDB	SalesDB	4	6148886	539.3	Synth	Retail	Regr

Same_Gen	SameGen	7	1081	0.3	Real	Kinship	Class
stats	Stats	8	38357	621.4	Real	Education	Regr
Student_loan	StudentLoan	13	1000	0.9	Real	Education	Class
Database not found	Thrombosis	3	806	1.9	Real	Medicine	Class
tpcc	TPCC	9	28433	174.1	Synth	Retail	Class
tpcds	TPCDS	24	99550	4587.5	Synth	Retail	Class
tpch	TPCH	8	148255	1925.1	Synth	Retail	Regr
trains	Trains	2	20	0.1	Synth	Logistic	Class
university	University	5	38	0.3	Synth	Education	Class
UW_std	UW-CSE	4	278	0.2	Real	Education	Class
Database not found	VOC	8	8215	2.7	Real	Logistic	Class
world	World	3	239	0.8	Real	Geography	Class

11. An analysis was made, and this was the distribution of the databases:

Category	Subcategory	Count
Domain	Retail	12
	Medicine	7
	Sport	6
	Geography	4
	Education	4
	Kinship	0
	Finance	3
	Government	2
	Entertainment	2
	Industry	1
	Logistic	1
	Total	**42**
Size	Small (< 1,000 instances)	16
	Medium (1,000 - 100,000 instances)	18
	Large (> 100,000 instances)	8
	Total	**42**
Type	Real	28
	Synthetic	14
	Total	**42**
Task	Classification	25
	Regression	17
	Total	**42**

12. The next step was filtering to 25 databases. It was found out that the Accidents' Database tables had Czech words in titles, not English words, which means that our code could not check the words in them. So, one more database was cancelled. It was from the Government Domain. A proportion was made regarding the Count of Domain databases:

Category	Subcategory	Count	Filtered Count
Domain	Retail	12	7
	Medicine	7	4
	Sport	6	3
	Geography	4	2
	Education	4	2
	Kinship	0	0
	Finance	3	2
	Government	2	1
	Entertainment	2	2
	Industry	1	1
	Logistic	1	1
	Total	**42**	**25**

13. This was the list of the 25 databases chosen, by alphabetical order of Domain:

Current Name	Name in Table 1	#Relations	#Instances	Size (MB)	Type	Domain	Task
stats	Stats	8	38357	621.4	Real	Education	Regr
UW_std	UW-CSE	4	278	0.2	Real	Education	Class
imdb_full	IMDb	7	794625	614.6	Real	Entertainment	Class
imdb_MovieLens	MovieLens	7	6039	151.9	Real	Entertainment	Class
ccs	CCS	6	1000	658.4	Real	Finance	Regr
financial	Financial	8	682	94.1	Real	Finance	Class
Countries	Countries	4	247	8.6	Real	Geography	Regr
Mondial	Mondial	33	454	3.3	Real	Geography	Class
legalActs	LegalActs	5	564268	238.2	Real	Government	Class
Mesh	Mesh	32	223	1.1	Real	Industry	Regr
trains	Trains	2	20	0.1	Synth	Logistic	Class
Biodegradability	Biodegradability	5	328	3.2	Real	Medicine	Regr
Carcinogenesis	Carcinogenesis	6	329	26.3	Real	Medicine	Class
Hepatitis_std	Hepatitis	7	500	2.2	Real	Medicine	Class
PTE	PTE	41	299	7.3	Real	Medicine	Class
classicmodels	ClassicModels	8	273	0.5	Synth	Retail	Regr
Credit	Credit	9	10084	443.6	Synth	Retail	Class
ftp	FTP	2	29555	7.5	Synth	Retail	Class
northwind	Northwind	29	830	1.1	Synth	Retail	Regr
sakila	Sakila	16	15991	6.6	Synth	Retail	Regr
SalesDB	SalesDB	4	6148886	539.3	Synth	Retail	Regr
tpcds	TPCDS	24	99550	4587.5	Synth	Retail	Class
AustralianFootball	AustralianFootball	4	3036	38	Real	Sport	Class
Hockey	Hockey	23	7759	15.5	Real	Sport	Class
lahman_2014	Lahman	25	23111	84	Real	Sport	Regr

14. This list of 25 databases contained 331 tables. The code used to download the information, 'BringingDatabaseInformation.ipynb' and the resulting file created 'all_selected_databases_info.xlsx' are available on our GitHub [29].
15. The final definition was based in analysis of the content, to choose interesting tables from the maximum amount of the 25 databases.
16. Some databases did not contain interesting information for attribute analysis, such as Carcinogenesis and Mesh, and were discarded.
17. Another decision was made regarding the number of Foreign Keys. In a future step of our research, we will analyse the relationship between different tables, and the fact that there are tables with many FKs will enable that, and this was an important decision taken.
18. This is the final definition of the fifty tables:

	index	name	database	area	instances	attributes	Column	Description	primary_key	foreign_keys
1	AustralianFootball.players	players	AustralianFootball	Sport	1911	6	pid, dob, first_name, last_name, h	pid (int), dob (datetime), first_name (varchar(255)), last_name (varc	pid	
2	AustralianFootball.teams	teams	AustralianFootball	Sport	18	2	tid, tname	tid (int), tname (varchar(255))	tid	
3	Biodegradability.molecule	molecule	Biodegradability	Medicine	328	4	molecule_id, activity, logp, mweig	molecule_id (varchar(255)), activity (float), logp (float), mweight (flo	molecule_id	
4	ccs.transactions_1k	transactions_1k	ccs	Finance	1000	9	TransactionID, Date, Time, Custom	TransactionID (int), Date (date), Time (time), CustomerID (int), CardI	TransactionID	
5	classicmodels.customers	customers	classicmodels	Retail	122	13	customerNumber, customerName	customerNumber (int), customerName (varchar(50)), contactLastNan	customerNumber	salesRepEmployeeNumber -> employees.employeeN
6	Countries.Metadata - Countries	Metadata - Countries	Countries	Geography	247	5	Country Code, Country Name, Reg	Country Code (varchar(255)), Country Name (varchar(255)), Region (v	Country Code	
7	Credit.charge	charge	Credit	Retail	1600000	8	charge_no, member_no, provider_	charge_no (int), member_no (int), provider_no (int), category_no (in	charge_no	category_no -> category.category_no, member_no -> r
8	financial.trans	trans	financial	Finance	1056320	10	trans_id, account_id, date, type, o	trans_id (int), account_id (int), date (date), type (varchar(6)), operati	trans_id	account_id -> account.account_id
9	ftp.session	session	ftp	Retail	30000	4	session_id, start_time, end_time,	session_id (char(6)), start_time (datetime), end_time (datetime), ge	session_id	
10	Hepatitis_std.dispat	dispat	Hepatitis_std	Medicine	500	4	m_id, sex, age, Type	m_id (int unsigned), sex (varchar(45)), age (varchar(45)), Type (varch	m_id	
11	Hockey.Master	Master	Hockey	Sport	7761	31	playerID, coachID, hofID, firstNam	playerID (varchar(255)), coachID (varchar(255)), hofID (varchar(255)),	firstName (varchar(255))	coachID -> Coaches.coachID
12	imdb_full.actors	actors	imdb_full	Entertainment	98690	3	actorid, name, sex	actorid (mediumint unsigned), name (varchar(250)), sex (enum('M','F	actorid	
13	imdb_full.movies	movies	imdb_full	Entertainment	3832	3	movieid, title, year	movieid (mediumint unsigned), title (varchar(400)), year (varchar(10	movieid	
14	imdb_full.movies2actors	movies2actors	imdb_full	Entertainment	138349	4	movieid, actorid, as_character, le	movieid (mediumint unsigned), actorid (mediumint unsigned), as_ch	movieid,actorid	movieid -> movies.movieid, actorid -> actors.actorid
15	imdb_MovieLens.movies	movies	imdb_MovieLens	Entertainment	3832	5	movieid, year, isEnglish, country,	movieid (mediumint unsigned), year (int), isEnglish (enum('T','F')), c	movieid	
16	imdb_MovieLens.u2base	u2base	imdb_MovieLens	Entertainment	996159	3	userid, movieid, rating	userid (int), movieid (mediumint unsigned), rating (varchar(45))	userid,movieid	movieid -> movies.movieid, userid -> users.userid
17	imdb_MovieLens.users	users	imdb_MovieLens	Entertainment	6039	4	userid, age, u_gender, occupation	userid (int), age (varchar(5)), u_gender (varchar(5)), occupation (varc	userid	
18	lahman_2014.hallOfFame	hallOfFame	lahman_2014	Sport	3883	8	hofID, yearID, votedBy, ballots, ne	hofID (varchar(10)), yearID (int), votedBy (varchar(64)), ballots (int), r	hofID,yearID	
19	lahman_2014.players	players	lahman_2014	Sport	16407	33	lahmanID, playerID, managerID, h	lahmanID (int), playerID (varchar(9)), managerID (varchar(10)), hofID	lahmanID	hofID -> halloffame.hofID
20	lahman_2014.salaries	salaries	lahman_2014	Sport	23111	5	yearID, teamID, lgID, playerID, sal	yearID (int), teamID (varchar(3)), lgID (varchar(2)), playerID (varchar(	'yearID,teamID,lgID,play	teamID -> teams.teamID
21	lahman_2014.teams	teams	lahman_2014	Sport	2715	48	yearID, lgID, teamID, franchID, div	yearID (int), lgID (varchar(2)), teamID (varchar(3)), franchID (varchar(	'yearID,lgID,teamID	
22	legalActs.legalacts	legalacts	legalActs	Government	580049	22	id, hash, update, Court, CaseKind,	id (int), hash (char(32)), update (timestamp), Court (varchar(100)), Ca	id	
23	Mondial.city	city	Mondial	Geography	3111	6	Name, Country, Province, Populat	Name (varchar(35)), Country (varchar(4)), Province (varchar(35)), Pop	Name,Province	Country -> country.Code, Province -> province.Name,
24	Mondial.country	country	Mondial	Geography	238	6	Name, Code, Capital, Province, Ar	Name (varchar(35)), Code (varchar(4)), Capital (varchar(35)), Province	Code	
25	Mondial.politics	politics	Mondial	Geography	238	4	Country, Independence, Depende	Country (varchar(4)), Independence (date), Dependent (varchar(4)),	Country	Country -> country.Code, Dependent -> country.Code
26	Mondial.population	population	Mondial	Geography	238	3	Country, Population_Growth, Infa	Country (varchar(4)), Population_Growth (float), Infant_Mortality (flo	Country	Country -> country.Code
27	Mondial.province	province	Mondial	Geography	1450	6	Name, Country, Population, Area,	Name (varchar(35)), Country (varchar(4)), Population (int), Area (floa	Name,Country	Country -> country.Code
28	northwind.Invoices	Invoices	northwind	Retail	2155	26	ShipName, ShipAddress, ShipCity,	ShipName (varchar(40)), ShipAddress (varchar(60)), ShipCity (varchar(15)), ShipRegion (varchar(15)), ShipPostalCode (varchar(10)), ShipCountry (varc		
29	northwind.Order Details Extende	Order Details Extended	northwind	Retail	2155	7	OrderID, ProductID, ProductName	OrderID (int), ProductID (int), ProductName (varchar(40)), UnitPrice (decimal(10,4)), Quantity (smallint), Discount (double(8,0)), ExtendedPrice (dou		
30	northwind.Orders	Orders	northwind	Retail	830	14	OrderID, CustomerID, EmployeeID	OrderID (int), CustomerID (varchar(5)), EmployeeID (int), OrderDate	OrderID	CustomerID -> Customers.CustomerID, EmployeeID ->
31	northwind.Products	Products	northwind	Retail	77	10	ProductID, ProductName, Supplier	ProductID (int), ProductName (varchar(40)), SupplierID (int), Categor	ProductID	CategoryID -> Categories.CategoryID, SupplierID -> Sup
32	northwind.Suppliers	Suppliers	northwind	Retail	29	12	SupplierID, CompanyName, Conta	SupplierID (int), CompanyName (varchar(40)), ContactName (varchar	SupplierID	
33	PTE.pte_atm	pte_atm	PTE	Medicine	9189	5	drug_id, atom_id, atom_type, Arg	drug_id (varchar(80)), atom_id (varchar(80)), atom_type (varchar(80)	atom_id	drug_id -> pte_drug.drug_id
34	PTE.pte_drug	pte_drug	PTE	Medicine	340	1	drug_id	drug_id (varchar(80))	drug_id	
35	sakila.address	address	sakila	Retail	603	8	address_id, address, address2, dis	address_id (smallint unsigned), address (varchar(50)), address2 (varc	address_id	city_id -> city.city_id
36	sakila.city	city	sakila	Retail	600	4	city_id, city, country_id, last_upda	city_id (smallint unsigned), city (varchar(50)), country_id (smallint ur	city_id	country_id -> country.country_id
37	sakila.country	country	sakila	Retail	109	3	country_id, country, last_update	country_id (smallint unsigned), country (varchar(50)), last_update (ti	country_id	
38	sakila.customer	customer	sakila	Retail	599	9	customer_id, store_id, first_name	customer_id (smallint unsigned), store_id (tinyint unsigned), first_n	customer_id	address_id -> address.address_id, store_id -> store.sto
39	sakila.film	film	sakila	Retail	1000	13	film_id, title, description, release	film_id (smallint unsigned), title (varchar(255)), description (text), re	film_id	language_id -> language.language_id, original_languag
40	sakila.inventory	inventory	sakila	Retail	4581	4	inventory_id, film_id, store_id, la	inventory_id (mediumint unsigned), film_id (smallint unsigned), sto	inventory_id	film_id -> film.film_id, store_id -> store.store_id
41	sakila.language	language	sakila	Retail	6	3	language_id, name, last_update	language_id (tinyint unsigned), name (char(20)), last_update (timest	language_id	
42	sakila.rental	rental	sakila	Retail	16044	7	rental_id, rental_date, inventory_	rental_id (int), rental_date (datetime), inventory_id (mediumint uns	rental_id	customer_id -> customer.customer_id, inventory_id ->
43	sakila.staff	staff	sakila	Retail	2	11	staff_id, first_name, last_name, a	staff_id (tinyint unsigned), first_name (varchar(45)), last_name (varc	staff_id	address_id -> address.address_id, store_id -> store.sto
44	sakila.store	store	sakila	Retail	2	4	store_id, manager_staff_id, addre	store_id (tinyint unsigned), manager_staff_id (tinyint unsigned), adc	store_id	address_id -> address.address_id, manager_staff_id ->
45	SalesDB.Sales	Sales	SalesDB	Retail	6715221	5	SalesID, SalesPersonID, Customer	SalesID (int), SalesPersonID (int), CustomerID (int), ProductID (int), C	SalesID	SalesPersonID -> Employees.EmployeeID, CustomerID
46	stats.users	users	stats	Education	40325	14	Id, Reputation, CreationDate, Disp	Id (int), Reputation (int), CreationDate (datetime), DisplayName (var	Id	
47	tpcds.item	item	tpcds	Retail	18000	22	i_item_sk, i_item_id, i_rec_start_d	i_item_sk (int), i_item_id (char(16)), i_rec_start_date (date), i_rec_en	i_item_sk	
48	tpcds.store_sales	store_sales	tpcds	Retail	2880404	23	ss_sold_date_sk, ss_sold_time_sk	ss_sold_date_sk (int), ss_sold_time_sk (int), ss_item_sk (int), ss_cust	ss_item_sk,ss_ticket_nu	ss_addr_sk -> customer_address.ca_address_sk, ss_cus
49	trains.cars	cars	trains	Logistic	63	10	id, train_id, position, shape, len_s	id (int), train_id (int), position (int), shape (varchar(255)), len (varcha	id	train_id -> trains.id
50	UW_std.person	person	UW_std	Education	278	6	p_id, professor, student, hasPositi	p_id (int), professor (varchar(11)), student (varchar(11)), hasPosition	p_id	

This final definition is in the file Fifty_Database_Tables_from_Prague_Repository.xlsx [29]

This is the analysis of the fifty final tables:

Database Distribution Analysis

Current Name	Number
sakila	10
Mondial	5
northwind	5
lahman_2014	4
imdb_full	3
imdb_MovieLens	3
AustralianFootball	2
PTE	2
tpcds	2
Biodegradability	1
ccs	1
classicmodels	1
Countries	1
Credit	1
financial	1
ftp	1
Hepatitis_std	1
Hockey	1
legalActs	1
SalesDB	1
stats	1
trains	1
UW_std	1
Carcinogenesis	0
Mesh	0
Total	**50**

Total number of databases: 23

Area Distribution Analysis

- Retail: 21 tables
- Sport: 7 tables
- Entertainment: 6 tables
- Geography: 6 tables
- Medicine: 4 tables
- Education: 2 tables
- Finance: 2 tables
- Government: 1 table
- Logistic: 1 table

Total number of areas: 9

Table Size Distribution Analysis

- Small (0-1,000 instances): 23 tables
- Medium (1,001-100,000 instances): 20 tables
- Large (100,001+ instances): 7 tables

Notable large tables:

1. Credit.charge: 1,600,000 instances
2. financial.trans: 1,056,320 instances
3. SalesDB.Sales: 6,715,221 instances
4. tpcds.store_sales: 2,880,404 instances

Primary Key Analysis

Total tables: 50 Tables

Tables with defined Primary Keys: 47 Tables (39 use single-column primary keys and 8 use composite keys).

3 Tables without defined Primary Keys:

1. Hockey.Master
2. northwind.Invoices
3. northwind.Order Details Extended

Composite Key Analysis

Total tables with composite keys: 8

Tables with composite keys:

1. imdb_full.movies2actors (movieid, actorid)
2. imdb_MovieLens.u2base (userid, movieid)
3. lahman_2014.halloffame (hofID, yearID)
4. lahman_2014.salaries (yearID, teamID, lgID, playerID)
5. lahman_2014.teams (yearID, lgID, teamID)
6. Mondial.city (Name, Province)
7. Mondial.province (Name, Country)
8. tpcds.store_sales (ss_item_sk, ss_ticket_number)

- The lahman_2014 database uses composite keys most frequently, with 3 out of its 4 tables having composite keys.
- The Mondial database uses composite keys for geographic entities, likely to handle cases where names might be repeated across different regions.
- The imdb databases use composite keys for relationship tables, which is a common practice in many-to-many relationships.

Foreign Keys analysis

Total tables: 50 Tables

Tables with defined Foreign Keys: 26 Tables

index	#FKs
tpcds.store_sales	9
Credit.charge	4
Mondial.city	3
northwind.Orders	3
sakila.rental	3
SalesDB.Sales	3
imdb_full.movies2actors	2
imdb_MovieLens.u2base	2
Mondial.politics	2
northwind.Products	2
sakila.customer	2
sakila.film	2
sakila.inventory	2
sakila.staff	2
sakila.store	2
classicmodels.customers	1
financial.trans	1
Hockey.Master	1
lahman_2014.players	1
lahman_2014.salaries	1
Mondial.population	1
Mondial.province	1
PTE.pte_atm	1
sakila.address	1
sakila.city	1
trains.cars	1

Observe that the databases with more tables chosen contain FKs, such as sakila, Mondial, and northwind, allowing future FK analysis:

Current Name	Number
sakila	10
Mondial	5
northwind	5
lahman_2014	4
imdb_full	3
imdb_MovieLens	3
AustralianFootball	2
PTE	2
tpcds	2

Appendix 3 – Extraction of Abbreviation–Full Term Pairs from the NameGuess Dataset

To enhance the semantic type detection process, the abbreviations dictionary was expanded by leveraging evaluation data from the *NameGuess* project [37]. The *NameGuess* system, introduced at EMNLP 2023, focuses on reducing data quality risks from cryptic column headers through bidirectional transformation between abbreviated and expanded column names.

1. **Repository and Dataset Structure**

The *NameGuess* GitHub repository (https://github.com/amazon-science/nameguess) provides a collection of JSON files representing evaluation datasets. These files are located in a structured subdirectory (e.g., nameguess/data/) and each JSON object contains:

- table_id: Identifier of the source table,
- technical_name: The abbreviated column name,
- gt_label: The expanded, ground truth label,
- difficulty: Optional metadata field.

2. **Parsing and Extraction Process**

A Python notebook file (‘**Obtaining Items from Name Guess.ipynb’** [29]) was developed to iterate through the relevant JSON files (eval_chicago.json, eval_la.json and eval_sf.json). The extraction followed these steps:

a. **Selection**: JSON files containing the substring “eval” in their names were selected for processing.
b. **Parsing**: The script used the json module to load each file and iterate through entries.
c. **Matching**: For each entry, technical_name was compared to gt_label. If the two differed and a valid abbreviation–full term pattern was found, the pair was extracted.
d. **Aggregation**: Each extracted pair was recorded along with the file name and table ID. The collected data were exported into a CSV file for inspection.

3. **Dictionary Expansion and Integration**

The resulting list of abbreviation–full form pairs was reviewed, deduplicated, and integrated with the existing dictionary. This increased the size of the abbreviations dictionary from 300 to **839 entries**, significantly enhancing coverage of commonly used column header abbreviations.

These new abbreviations also contributed to expanding the *Formats Dictionary*, particularly for semantic types commonly represented through abbreviations (e.g., 'hrs' → 'hours', 'amt' → 'amount', 'dob' → 'date_of_birth'). This update improved the accuracy of rule-based semantic type detection, particularly in real-world data sources with non-standardised column labels.

See some interesting abbreviations:

10th: tenth
abbrev: abbreviation
abdnt: abandonment
abndnd: abandoned
abs: absolute
admin: administrator
advsry: advisory
aero: aerospace
afdble: affordable
affiliatn: affiliation
affirmatn: affirmation
affltd: affiliated