



POZNAN UNIVERSITY OF TECHNOLOGY

DW Loading and Refreshing Techniques: ETL Data Quality

Robert Wrembel
Poznan University of Technology
Institute of Computing Science
Poznań, Poland
Robert.Wrembel@cs.put.poznan.pl
www.cs.put.poznan.pl/rwrembel



Outline

- **Erroneous data**
 - value errors
 - missing values
 - inconsistencies
 - outliers
 - duplicates
- **Dimensions of data quality**



Errors

- Text (edited only during initial writing) written with keyboard includes approx. 0.2% typing errors
- Approx. 70% of typing errors are noted and corrected while writing

J.J. Pollock, A. Zamora: Collection and Characterization of Spelling Errors in Scientific and Scholarly Text. Journal of the American Society for Information Science, Vol. 34, Issue 1, 1983



Errors

- Four basic spelling error operations (all involve one character)
 - insertion, omission, transposition, substitution
- Multiple errors - two or more (not necessarily different) error operations
- A spelling error is most likely to involve third letter

position number of a letter	data source evaluated									
	1	2	3	4	5	6	7	8	9	10
1	3.30	5.23	3.30	11.14	2.66	4.89	3.04	3.75		
2	13.70	9.11	8.60	11.93	9.74	8.96	8.56	6.25		
3	23.00	20.84	17.50	18.36	13.36	17.58	31.77	13.75		
4	15.30	16.38	18.00	14.25	14.01	16.76	15.19	13.75		
5	14.00	11.58	12.60	11.45	11.35	10.88	10.22	15.00		
6	10.60	11.42	11.80	10.17	14.37	12.17	10.50	10.00		
7	8.40	8.91	10.60	7.84	11.43	10.13	8.56	11.25		
8	4.90	5.99	6.60	5.70	8.60	9.95	5.53	11.25		
9	3.80	4.69	4.80	3.67	5.58	4.66	3.87	10.00		
10	3.00	5.85	6.20	5.49	8.90	4.02	2.76	5.00		



Errors

➔ The first position at which the word differs from the misspelling

		ERROR POSITION																			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
W	1	86	14	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	2	21	43	35	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
O	3	14	22	47	17	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	4	16	21	35	22	6	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
R	5	13	18	29	18	17	4	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	6	10	15	25	19	14	13	4	-	-	-	-	-	-	-	-	-	-	-	-	-
D	7	9	11	22	19	14	13	9	3	-	-	-	-	-	-	-	-	-	-	-	-
	8	7	11	17	17	16	15	10	6	2	-	-	-	-	-	-	-	-	-	-	-
	9	5	9	14	15	15	16	13	8	5	2	-	-	-	-	-	-	-	-	-	-
L	10	5	8	12	13	11	13	15	12	8	2	1	-	-	-	-	-	-	-	-	-
	11	4	7	11	12	11	13	12	13	9	6	2	1	-	-	-	-	-	-	-	-
E	12	3	7	10	10	10	12	13	9	10	8	5	2	1	-	-	-	-	-	-	-
	13	4	5	8	10	10	11	9	9	9	10	8	4	2	1	-	-	-	-	-	-
N	14	3	5	8	9	10	10	10	8	8	10	8	4	4	2	1	-	-	-	-	-
	15	3	5	6	7	8	10	8	9	8	8	6	7	4	2	3	0	-	-	-	-
G	16	4	5	5	8	7	9	7	7	9	10	4	4	7	3	1	1	1	-	-	-
	17	4	3	5	5	6	10	10	11	9	7	2	12	4	7	8	2	1	1	-	-
T	18	2	3	2	3	6	9	11	5	7	5	7	10	12	6	7	3	4	3	1	-
	19	0	0	3	4	10	3	12	10	6	7	11	8	10	3	4	4	1	4	0	0
H	20	0	0	5	0	5	3	0	14	5	14	11	11	8	0	5	3	3	5	3	5



Errors

➔ From 80% to 96% misspellings have only one error

- O - character omission
- I - additional character inserted
- S - character substituted
- T - character transposed
- M - multiple errors

Database	Misspellings	Error Type				
		O(%)	I(%)	S(%)	T(%)	M(%)
CA	10,243	38	31	14.5	13	3.5
ACS	5,542	36	29	16	13	6
ACS-2	2,512	38.5	24.5	15.5	15.0	6.5
DOLE	27,844	30	27.5	21	11.5	9.5
BA	4,662	40	18	20.5	15	7
CIN	1,718	39.5	23	19.5	12.5	6
ISA	362	33	34.5	16	10	6.5
PI	80	15	19	16	40	9
ALL	52,963	34	27	19	12.5	7.5

➔ Omission is the most common error operation

- omission of two contiguous letters, often corresponding to a syllable, e.g., additional



Errors

⇒ Other examples

- ŁĄŻNIKI - ŁĄŻNIKI
- WIERZBOWICE - WIERZBOCICE
- KRETA - KRĘTA
- DOLNA - DOLINA
- PRZEMYSŁAWA - PRZEMYSŁOWA
- STUDZIENICE - STUDZIENIEC
- KAROLEWO - KORALEWO
- WOŁOWA - WAŁOWA
- GARBACKA - GARBATKA
- WŁOCINA - SŁOCINA
- SOKOŁOWICE - SOKOŁOWIEC



Typical text value errors

- ⇒ ' %' - leading space(s)
- ⇒ '% ' - trailing space(s)
- ⇒ values composed of two text strings (f. names, l. names, street names)
 - '% - %'
 - '%- %'
 - '% - %'
- ⇒ **0** instead of **O**
- ⇒ names (first, last, street, city, country, ...)
 - leading or trailing characters other than letters
 - other characters than letters in any place of a value
 - '%.%', '%..%'



Other value checks

- ➔ **postal code**
 - **length**
 - **values other than digits**
 - **with or without '-'**
- ➔ **email format**
 - **text@text.text**
- ➔ **addresses**

KROSINKO / UL. ZIELONA
SIKORSKIEGO 20/33
SIEDLEC12/3 60-123
WŁOCŁAWSKA 13/24 62-7



Inconsistencies

BOH II WOJNY ŚWIAT
BOH. 2
BOH 2 OJNY ŚW
BOH 2 WOJ ŚWIAT
BOH 2 WOJ ŚWIATOWEJ
BOH DRUGIEJ WOJNY ŚWIATOW
BOH II WOJNY ŚW
BOH II ŚWIATOWEJ
BOH. II W.
BOH II W ŚWIATOWEJ
BOH II W ŚW
BOH II W ŚWIAT
BOH II W ŚWIATOWEJ
BOH II WIJ ŚWIAT
BOH. II WOJ. SW.
BOH II WOJ SW
BOH II WOJ ŚWIAT
BOH II WOJ ŚWIATOWEJ
BOH. II WOJ. ŚW.
BOH II WOJ. ŚW

BOH II WOJ. ŚW.
BOH II WOJ ŚW
BOH II WOJ. ŚWIAT.
BOH II WOJ ŚWIAT
BOH II WOJ ŚWIATOWEJ
BOH II WOJN ŚW
BOH II WOJNA ŚWIAT
BOH. II WOJNY ŚWIATOWEJ
BOH II WOJNY ŚWIATOWEJ
BOH. II WOJNY ŚWIATOWEJ
BOH II WOJNY ŚWIATOWEJ
BOH II WOJ-SWIAT
BOH II WOJ.SWIATOWEJ
BOH. II WOJ.ŚWI.
BOH II WOJ-ŚWIAT
BOH II WOJ.ŚWIAT
BOH II WOJ.ŚWIAT.
BOH. II W.Ś.
BOH IIWOJ ŚWIAT
BOH IIWOJN ŚWIAT



Inconsistencies

➤ First names

- more than one name
 - MIROSLAWPAWEŁ
 - PRZEMYSŁAWANDRZEJ
 - Robert, Andrzej
 - ...
- other values not related to first names



Discovering data quality

- Errors/outliers found by **profiling**
- Frequency of name distribution
 - in a given data set
 - globally



Correcting errors

⇒ Errors to be cleaned automatically

- ⇒ ' %' - leading space(s)
- ⇒ '% ' - trailing space(s)
- ⇒ values composed of two text strings (first, last, street names)
 - '% - %'
 - '%- %'
 - '% - %'
- ⇒ 0 instead of O
- ⇒ names (first, last, street, city, country, ...)
 - leading or trailing characters other than letters
 - other characters than letters in any place of a value
 - '%.%%', '%..%'

⇒ Errors to be cleaned semi-automatically

- consistent naming
- with the support of translation dictionaries
 - street names
 - city names
 - ...

BOH II WOJNY ŚWIAT
BOH 2 WOJNY ŚW
BOH 2 WOJ ŚWIAT
BOH 2 WOJ ŚWIATOWEJ
BOH DRUGIEJ WOJNY ŚWIATOW
BOH II WOJNY ŚW
BOH II ŚWIATOWEJ



Correcting errors

⇒ Errors to be cleaned manually

- reference dictionaries need to be consulted
 - city name + postal code may allow to find out which name is correct, e.g., **STUZIENICE - STUZIENIEC**
- other examples
 - **MIROSLAWPAWEŁ**
 - **PRZEMYSŁAWANDRZEJ**



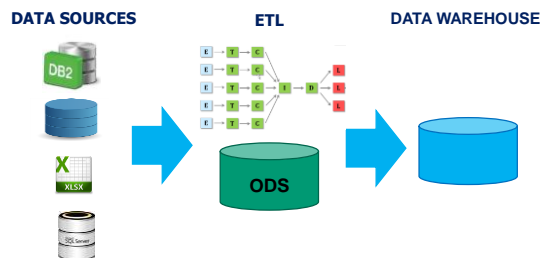
Data quality

⇒ Fitness for use

- Wang, R. Y., & Strong, D. M. (1996) Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems 12(4)

⇒ Assured by ETL processes

⇒ Discovered by profiling



© R.Wrembel - Poznan University of Technology, Institute of Computing Science

15



Data quality

⇒ DQ dimensions

- L. Cai, Y. Zhu: The Challenges of Data Quality and Data Quality Assessment in the Big Data Era.
<https://datascience.codata.org/articles/10.5334/dsj-2015-002/>
- **D1: Availability**
- **D2: Usability**
- **D3: Reliability**
- **D4: Relevance**
- **D5: Presentation quality**

© R.Wrembel - Poznan University of Technology, Institute of Computing Science

16



Data quality

⇒ D1: Availability

- **Accessibility**
 - easiness of accessing data
 - openness of data sources to provide data and metadata
- **Latency/timeliness**
 - time between data generation and ingestion
- **Authorization**
 - access rights to use data



Data quality

⇒ D2: Usability

- **Documentation**
 - the quality, details, and completeness of documentation
- **Credibility/trustworthiness**
 - reliability of data sources
 - reliability of data collection methods
- **Metadata**
 - the quality, details, and completeness



Data quality

⇒ D3: Reliability

impacted by duplicates

- **Accuracy**
 - whether stored data reflect real-world data
- **Consistency**
 - correctness w.r.t. to some criteria (see consistency in databases)
- **Integrity**
 - data have defined and complete structures, e.g., are normalized (db notion)
 - accurate and consistent
- **Completeness**
 - no missing pieces of data
- **Auditability**
 - well documented data life-cycle (provenance, lineage)



Data quality

⇒ D4: Relevance

- **Fitness**
 - degree to which available data fit user needs
 - how much of available data are really used

⇒ D5: Presentation quality

- **Readability**
 - how easy one can understand a data source structure and data
- **Structure**
 - structured vs. unstructured
 - effort to transform unstructured to structured



DQ in DW architecture

Dimension		DW arch.	
D1: Availability	Accessibility	high	<ul style="list-style-type: none"> ▪ outdated ▪ some parts missing
	Latency	24h	
	Authorization	yes	
D2: Usability	Documentation	medium - good	customers' data: <ul style="list-style-type: none"> ▪ outdated ▪ duplicated ▪ multiple versions sensor data: <ul style="list-style-type: none"> ▪ missing ▪ outliers ▪ wrong
	Credibility	high	
	Metadata	rich	
D3: Reliability	Accuracy	medium - high	customers' data: <ul style="list-style-type: none"> ▪ no integrity constr. checking sensor data: <ul style="list-style-type: none"> ▪ no correctness checking
	Consistency	medium - high	
	Integrity	high	
	Completeness	medium - high	
D4: Relevance	Auditability	high	customers' data: <ul style="list-style-type: none"> ▪ missing (zip, phone no, ...) sensor data: <ul style="list-style-type: none"> ▪ missing measurements
	Fitness	high	
D5: Present. quality	Readability	high	
	Structure	high	



Big Data

⇒ The most popular characteristics

- Volume
- Velocity
- Variety



Big data quality

⇒ Traditional DW

- data profiling + ETL

⇒ Big Data eco-system

- **volume** → full data profiling may not be feasible
 - sampling & approximations
 - data quality with probability ranges
- **variety** → profiling is more difficult
 - different techniques for different data types
 - different quality measures for different data types
- **velocity** → real-time DQ assessment is difficult
 - profiling often impossible for fast arriving data (overhead, hardware resources needed)
 - sampling
 - data quality may frequently change in time



Big data quality

⇒ Different domains expose different DQ dimensions

- **human generated (e.g., social media)**
 - low credibility
 - low: accuracy, integrity, consistency, completeness, auditability
 - rather high accessibility
 - low-medium fitness
 - unstructured
- **machine generated (e.g., sensors)**
 - high(er) credibility
 - high(er): accuracy, integrity, consistency, completeness
 - high fitness
 - high readability, well structured



Big data quality assurance

- ⇒ M.H. Farid, A. Roatis, I.F. Ilyas, H. Franziska Hoffmann, X. Chu: CLAMS: Bringing Quality to Data Lakes. SIGMOD, 2016
- ⇒ **RDF to store heterogeneous data**
- ⇒ **Relational view V to transform on-the-fly RDF data into a tabular format**
- ⇒ **Automatic constraint discovery + manual constraint building**
- ⇒ **Denial constrains defined for V to control DQ**
 - **DQ: instead of 'saying' what is allowed it 'says' what is not allowed**



DQ summary

Dimension		DW arch.	DL arch.
D1: Availability	Accessibility	high	low - medium - high
	Latency	24h	RT, minutes, hours
	Authorization	yes	yes - no
D2: Usability	Documentation	good	some - poor - none
	Credibility	high	low
	Metadata	rich	some - poor - none
D3: Reliability	Accuracy	medium - high	low
	Consistency	medium - high	low
	Integrity	high	low - none
	Completeness	medium - high	low - none
	Auditability	high	low - none
D4: Relevance	Fitness	high	low - medium
D5: Present. quality	Readability	high	low - medium
	Structure	high	none - low - medium