Data Quality *

Jens Lechtenbörger

Winter Term 2023/2024

Contents

1 Introduction **1**

2 Data Quality Frameworks **3**

3 Data Profiling **4**

4 Sample Cleaning Process **5**

5 Conclusions **7**

1 Introduction

1.1 Learning Objectives

- Explain data quality (DQ) as multidimensional concept
- Identify DQ issues with reference to quality dimensions
  - Apply data profiling (as part of project work)
- Apply cleaning process (as part of project work)

Learning objectives specify what you should be able to do after having worked through a presentation. Thus, they offer guidance for your learning.

Each learning objective consists of two major components, namely an *action* verb and a topic. Action verbs specify what actions you should be able to perform concerning the topic, and they indicate the target level of skill (in Bloom's Taxonomy or its revised version as sketched under the hyperlink above).

You may want to think of learning objectives as sample exam tasks.

1.2 Motivating Examples

- Who wrote this?
  - Lechtenberger, Jens Lechtenboerger, Jens Lechtenborger, Jens Lechtenb?rger, J. Lechtenbörger, `lechten`

*This PDF document is an inferior version of an OER HTML page; free/libre Org mode source repository.

- Where?
  - Grevener Straße 91, Steinfurter Str. 107, Steinfurter Straße 109, Leonardo Campus 3, Leonardo-Campus 3
- When?
  - 2020-11-01, Nov 1st, 1. November, 01/11, 11/01

These sample values hint at different types of data quality issues. Suppose that such values are recorded in different places at an organization, where a single "correct" value should be used.

Please take a moment to think about the following question, which will be revisited later on: How would you classify the data quality issues concerning the author's name?

The last two address variants indicate a spelling issue, which concerns **syntax** and may therefore seem easy to resolve based on a dictionary or registry of streets (where we would find "Leonardo-Campus" but not "Leonardo Campus"). With string similarity measures, which are a topic later on, we might identify the correct spelling. (Indeed, when comparing the candidate `Leonardo Campus` to known street names, `Leonardo-Campus` will be the most similar match.)

Semantic issues are much harder. Indeed, *all* shown addresses were part of the author's office location at different points in time (namely, the Department of Information Systems at the University of Münster in Germany). Notably, the office just moved once, while we see four different addresses (with a renumbering of buildings in one street, where "Straße" is used inconsistently in abbreviated form once, followed by a re-assignment of the building to a different street name later on). String similarity does not help here.

Besides simple typos, the examples here point to the lack of *standards* or *conventions*. For example, higher data quality would arise, if we answered the following questions ahead of time:

- Which abbreviations do we use?
- What string encoding (e.g., UTF-8)?
- What date format?

Similar considerations apply to data values in general. *Data types* and *integrity constraints* of database management systems can enforce some conventions and thereby prevent several types of errors.

1.2.1 An Impressive Example

- See spelling variants for `britney spears` corrected by Google
  - Bottom line: Working with manual user input is challenging

1.3 Why do we care?

- Quotes from [Mar05] (2005)
  - "88 per cent of all data integration projects either fail completely or significantly over-run their budgets"
  - "75 per cent of organisations have identified costs stemming from dirty data"
  - "33 per cent of organisations have delayed or cancelled new IT systems because of poor data"
  - "$611bn per year is lost in the US in poorly targeted mailings and staff overheads alone"
  - "According to Gartner, bad data is the number one cause of CRM system failure"
  - "Customer data typically degenerates at 2 per cent per month or 25 per cent annually"

1.4 Costs of Poor Data Quality

- Attributing costs or assessing impact of poor quality is hard
  - See [HZV11] for four types of costs
    - Two-by-two matrix
      - Costs may be direct or indirect
      - Quality affects operational tasks and strategic decisions
    - E.g., payment errors are direct and operational, poor production planning is indirect and strategic
  - Negative effects cited in [CR19]
    - According to Gartner in 2018, organisations attribute losses of 15 million USD per year on average
    - 2016 IBM research estimates total annual losses in US to be 3 trillion USD
    - According to KPMG 2017 Global CEO Outlook, 56% of CEOs worry about negative impact on their decisions
    - Compliance risks

2 Data Quality Frameworks

2.1 DQ Framework Overview

- Data quality frameworks offer processes for strategic DQ improvement
- See [CR19] for an overview
  - Comparison of twelve DQ frameworks that cover
    - DQ definition
    - DQ assessment
    - DQ improvement
  - Decision guide for organizations
    - Criteria to narrow down choice of framework

2.1.1 Data Quality (DQ)

- "Fitness for use" (with background in quality literature, see [WS96])
  - Quality is judged by consumer according to context and purpose
- Lots of data quality dimensions, see [Sid+12] for survey
  - According to [CR19], most commonly (going back to [WS96]):
    - Completeness: Sufficient breadth, depth, scope for task at hand
    - Accuracy: Correct, reliable, certified
    - Timeliness: Age is appropriate for task at hand
    - Consistency: Same formats, compatible with previous data
    - Accessibility: Available, or easily and quickly retrievable

As stated here, data quality is a multidimensional concept, for which ultimately data consumers will judge whether the quality of a given data set is fit, or good enough, for their intended use.

In the context of data examples shown earlier, I asked how you would classify some DQ issues. What were your thoughts back then?

Please revisit that slide with the dimensions mentioned here. Which ones are problematic for what purposes? Anything else that comes to mind?

2.1.2 DQ Assessment

- Need metric per quality dimension
  - Subjective, e.g., measurements with surveys among data consumers
  - Objective, e.g., count NULL values, constraint violations, duplicates, or measure number of erroneous decisions

2.1.3 DQ Improvement

- Improve information products
  - E.g., de-duplicate data, fill in missing values, standardize, fix errors
    - See cleaning process later on
- Improve information processes
  - May start from root cause analysis
    - Why did low-quality data arise?
    - Change processes to avoid root causes.

3 Data Profiling

3.1 Getting to Know Your Data

- See Chapter 3 in [HKP11]
  - Inspect types of attributes
    - and their statistical properties such as mean, media, mode, quartiles, variance
  - Use visualizations
  - Along the way, identify data quality issues

3.2 Data Profiling Aspects

- Methodical inspection of data instead of manual "eye-balling"
- See [AGN15] for a survey
  - "Data profiling is the set of activities and processes to determine the metadata about a given dataset."
    - Single column, e.g., cardinalities, value distributions, data types
    - Multiple columns, e.g., correlations, topic overlap, duplicates

        * Dependencies, e.g., (foreign) keys, functional dependencies and their violations

- Aside
  - SIGMOD 2017 tutorial slides based on [AGN15]
    * With lists of industrial and research tools

3.3 Talend Open Studio for Data Quality

- Sample tool with data profiling capabilities
  - Java, free/libre and open source (Apache License, Version 2.0)
  - Project at SourceForge
- See user's guide for sample analysis results

4 Sample Cleaning Process

Kimball, Dealing with Dirty Data, DBMS, 1996

- Process description
  - Six steps, may be part of ETL process
    1. Elementizing
       * Split non-atomic values, e.g., names, addresses, dates (recall initial examples)
         · Regular expressions may help
    2. Standardizing (next slides)
    3. Verifying
       * Check whether elements are mutually consistent, e.g., ZIP code 48149 cannot be in Bavaria (08..., 09...)
       * Integrity constraints
    4. Matching
       * Check whether "equivalent" element does already exist; if yes, augment with new information (subsequent slides)
    5. Householding
       * Try to group elements, e.g., married couples
    6. Documenting

4.1 Standardizing (1/2)

- Consider "gender"
  - Varying source representations
    - Possibly "real" values mixed with salutations, academic titles
      - Possibly all of male/female/diverse/etc., m/w, m/f, Mann/Frau, Herr/Frau/Firma, Dipl.-Ing., Dr., Prof., ?, unknown, NULL
      - Use profiling results (e.g., count distinct, histogram)

Lookup	Data Source	Source	TargetCode
	S1	unknown	0
	S1	female	1
	S1	male	2
	S1	diverse	3
	S2	f	1
	S2	m	2
	S2	?	0
	S2	NULL	0
	S3	1	2
	S3	2	1
	...	...	...

  - Define convention, e.g.:

TargetCode*	TargetName
0	unknown
1	female
2	male
3	diverse

  - May use lookup table
    - Data type of column Source?
    - Join source data with lookup table
      - Beware of NULLs!
    - View or ETL process to produce target data

4.2 Standardizing (2/2)

- NULL values are related to the completeness dimension of DQ
- Avoid NULL values, explicitly represent degree of knowledge
  - **Three** types of NULLs
    - Not existing (inapplicable); no incompleteness issue
    - Value exists for sure, but we don't know it; incomplete
    - We don't know whether a value exists; unknown whether incomplete

ID	Name	Surname	Birthdate	E-Mail
1	John	Smith	03/17/1974	smith@abc.it
2	Edward	Monroe	02/03/1967	NULL (does not exist)
3	Anthony	White	01/01/1936	NULL (existing but unknown)
4	Marianne	Collins	11/20/1955	NULL (not known if existing)

(Source: [SMB05])

4.3 Matching

- **Matching** = object identification = duplicate detection
  - Matching is easy for exact duplicates or with "real" keys (e.g., social security number)
    - For computers at least; see this challenge for human beings
  - Otherwise, need **quasi-identifiers**
    - Groups of attributes, possibly with similarity matching for probabilistic matching
    - E.g., name, date of birth, and address in presence of spelling mistakes

4.3.1 Matching Outlook

- Finding of **similar items** to be revisited in several sessions
  - More efficient approaches than naive comparison of all pairs with quadratic complexity?
  - Measures/metrics for similarity?
- Afterwards, **data fusion** is necessary
  - Given duplicates, create single object representation while resolving conflicting values

5 Conclusions

5.1 Summary

- Data quality
  - is a pressing topic in practice,
  - is a multidimensional concept,
  - can be improved with cleaning steps (e.g., as part of ETL processes),
  - is the focus of data quality frameworks for strategic approaches.

Bibliography

[AGN15] Ziawasch Abedjan, Lukasz Golab, and Felix Naumann. "Profiling relational data: a survey". In: *The VLDB Journal* 24.4 (2015), pp. 557–581. URL: https://doi.org/10.1007/s00778-015-0389-y.

[CR19] Corinna Cichy and Stefan Rass. "An overview of data quality frameworks". In: *IEEE Access* 7 (2019), pp. 24634–24648. DOI: 10.1109/ACCESS.2019.2899751. URL: https://doi.org/10.1109/ACCESS.2019.2899751.

[HKP11] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. 3rd ed. Morgan Kaufmann Publishers, 2011.

[HZV11] Anders Haug, Frederik Zachariassen, and Dennis Van Liempd. "The costs of poor data quality". In: *Journal of Industrial Engineering and Management (JIEM)* 4.2 (2011), pp. 168–193. DOI: 10.3926/jiem.2011.v4n2.p168-193. URL: https://doi.org/10.3926/jiem.2011.v4n2.p168-193.

[Mar05] Richard Marsh. "Drowning in dirty data? It's time to sink or swim: A four-stage methodology for total data quality management". In: *Journal of Database Marketing & Customer Strategy Management* 12.2 (2005), pp. 105–112. URL: https://doi.org/10.1057/palgrave.dbm.3240247.

[Sid+12] F. Sidi et al. "Data quality: A survey of data quality dimensions". In: *2012 International Conference on Information Retrieval & Knowledge Management*. 2012, pp. 300–304. DOI: 10.1109/InfRKM.2012.6204995. URL: https://doi.org/10.1109/InfRKM.2012.6204995.

[SMB05] Monica Scannapieco, Paolo Missier, and Carlo Batini. "Data Quality at a Glance". In: *Datenbank-Spektrum* 14 (2005), pp. 6–14. URL: https://web.archive.org/web/20070823223630/http://www.datenbank-spektrum.de/v2/archiv/beitrag.html?key=dbspektrum/ScannapiecoMB05&nummer=14.

[WS96] Richard Y Wang and Diane M Strong. "Beyond accuracy: What data quality means to data consumers". In: *Journal of Management Information Systems* 12.4 (1996), pp. 5–33. URL: https://www.tandfonline.com/doi/abs/10.1080/07421222.1996.11518099.

License Information