Master Thesis Scientific Paper – Final version, November 30th, 2017

Are chains in need of other requirements regarding data quality compared to just one company?

The case of the Dutch Energy Chain

M. Pronk

Complex Systems Engineering and Management (CoSEM), Delft University of Technology

Abstract

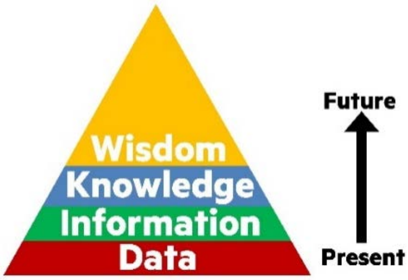
High quality data is essential for organizations to deduce correct information from it. The four main data quality dimensions are accuracy, completeness, currency and consistency, and are frequently used to measure data quality within single organizations. When organizations are operating in a chain they may be dependent on data originating from other organizations in the chain, which may hold other data requirements and quality standards. Therefore data chains with multiple participants may require different requirements regarding data quality than do single entities. In this paper we try to answer the question what are the data requirements for data in the Dutch electricity chain.

In the case study an analysis of the data flow in the energy chain in the Netherlands and its quality problems was made. It was found that low data quality originating from one party cause problems elsewhere in the chain and are contributing to costs for other parties in the chain. Moreover, data transferred from one party to another often are not up to standards; are not consistent or may even differ between similar organizations in the chain. Four data quality dimensions were found that may remedy and help prevent data quality problems in the future; these are: responsibility, usability, interoperability and comparability. Therefore a new framework is recommended combining the four traditional data quality dimensions with the four dimensions that are more suitable for data chains where multiple companies are involved.

Keywords: data quality, data chain, data quality dimensions

1 Introduction

Most decisions in businesses are based on information that is derived from a lot of data. More importantly, if data cannot be trusted fully, the analysis that has been done loses value and affects the quality of the decision. An analysis based on data is always dependent on the reliability of that data (Weidema, 1996). Hence within organisations it is crucial that data is of high quality to perform better. High quality data therefore is an essential prerequisite to get value out of the data (Cai, 2015). When data is of insufficient quality you can risk that the data and analysis will not be used (Lee, 2017). Therefore, using high quality data is important in business and research. Low quality data namely impacts at three levels: operational, tactical and strategical (Redman, 1998). A lot of business decisions are based on information that has been deducted from data. But what happens when the data quality is uncertain? Can decisions still be made? What is the effect on the other parties involved?



Improvements that lead to a higher data quality cost money. High quality data however can lead to savings on current spending. Costs are being avoided by making sure data is correct initially. Showing that putting focus on data quality from an early point could therefore be rewarding. Decisions based on incorrect data can be financially disastrous for an organisation. An estimate of IBM is that low data quality is costing the US three trillion dollars on a yearly basis (IBM, 2014).

When data can be combined, for instance in a production and distribution chain, its value can increase not just for the single company but for all the organisations involved. Since the energy sector is described by the energy chain, it is necessary for data to be communicated and transmitted in the chain. Data in chains is the communication between different parties and the transmission of data from one party to the other. However, data in chains is not always an easy task. Even at an (single) organisational level, the department which creates the data is not always the one that will use it or benefit from it. Therefore, the goals on data quality and context are not always aligned which leads to less improvements on data. The same problem arises and probably will be worse when data are being exchanged between different organisations.

In this paper different data dimensions will be analysed. Through a literature study different popular dimensions will be identified. A case study research executed in the electricity chain will shed light on problems observed when multiple stakeholders come into play. The question that will be answered is: Which data quality dimensions describe the needs in the Dutch electricity chain best?

2 Data quality dimensions

What is the definition of high quality data and how could the data quality be determined?
A dimension is "a set of data quality attributes that represent a single aspect or construct of data quality" (Wang, 1996). Good data quality dimensions can limit the uncertainty and can improve the analysis of the process (Weidema, 1996). So, using the correct data quality dimensions to assess data quality is important. There are lots of dimensions to describe data quality according to professionals (Wang, 1996) and research (HIQA, 2011). Accuracy, timeliness, consistency and completeness are four data quality dimensions that are used a lot in important literature (Scannapieco, 2005). These four dimensions will be further explored.

2-1 Accuracy

"Accuracy of data refers to how closely the data correctly captures what it was designed to capture" (HIQA, 2011). Accuracy is needed to see if the correct value has been recorded. When a street name is matched to a wrong connection this can be considered inaccurate. When wrong payment data is connected to a customer this can be considered also an inaccuracy mistake. Some mistakes are more important than others since they have a bigger impact (Batini, 2009). Accuracy therefore is variable in the impact it can have on an analysis.

2-2 Completeness

"Completeness of data refers to the extent to which the data collected matches the data set that was developed to describe a specific entity." (HIQA, 2011). Data records that have all fields filled in are of course the dream of every organisation. Unfortunately, that is not always the case. Not all data is complete and therefore parties need to handle the non-completeness.

2-3 Timeliness

"Timeliness is the extent to which age of the data is appropriate for the task at hand" (Scannapieco, 2005). When data is old it still can be considered good for the task. Or based upon the characteristic it can be outdated and therefore not current (Batini, 2009). Within the electricity chain physical infrastructures are there for long periods of time. These infrastructural data elements within a chain are not really losing value after years. Electricity usage data however need to be much more recent in order to have value. The electricity usage value is most valuable when it is very recent as the planning of the power to be generated is based on them. But in 20 years time their added value will be close to zero. It is important that the usage data is correct. However not all data needs the same timeliness in order to keep the data current and thereby contributing to the data quality.

2-4 Consistency

Consistency is the process of checking data whether they are consistent. When Jaffalaan 5 is the street name and Delft is the city then the zip-code should be 2628BX. When this is not the case then the data is not consistent. Data can be inconsistent when different parties are using different procedures. As argued different departments in a company can be involved in data handling and not everybody may be sticking to the same principles. Standardisation of the data entries is preferred since it will then be easier to turn data into reliable information.

Now we have explored the data dimensions important within a single company, the question arises what additional requirements can be found when studying data quality in a chain? Therefore, we will analyse the data quality challenges in the Dutch electricity chain. The case study will explore data handling at KPN, the largest telecom provider in The Netherlands. KPN is also a big user of electricity and therefore there is an active role from KPN when it comes to energy data handling.

3 Observations in the electricity chain

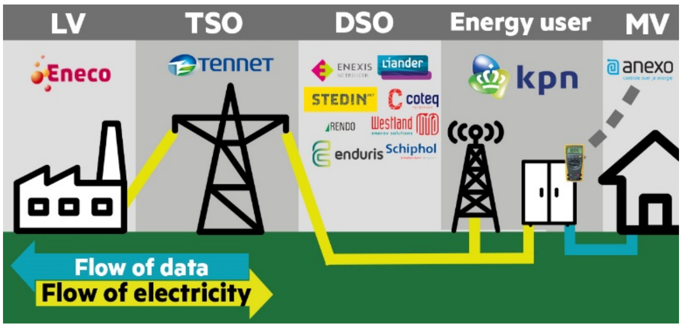


Figure 1: Overview of the parties in the electricity chain

In Figure 1 an overview is given of the Dutch electricity chain. Multiple parties like Suppliers (LV) Transmission System Operator (TSO), Distribution System Operators (DSO), the measuring organisations (MV) and the energy user are important actors. All these parties interact with each other in order to ensure the delivery of electricity in the Netherlands.

Each party is handling the data in a different way and they need the data for different reasons. This makes the quality and transmission of the data problematic. However, within a chain there also lies a joint responsibility. All parties rely on each other's data. Organisations in chains should therefore realise that a mistake from their side can have negative consequences down the chain. Parties create their own administration of the activities they are performing. Therefore, data exchange can be problematic and have a big impact on the different parties in the chain.

The problem is that the department which creates the data is not always the one that will use it or benefit from it. Therefore, the goals on data quality and context are not always aligned which leads to insufficient data quality.

Within a chain some mistakes are more important than others since they have a bigger impact on the operation (Batini, 2009). There are different demands in a chain. In chains an extra complexity is added by multiple interests and responsibilities from other organisations. Integrating the data and getting most out of it is vital for the costs, innovation and stability of the energy infrastructure. There are multiple ways to describe the data and data quality. The most prominent problems found in the electricity chain are listed below.

No communication

Parties in the energy chain do not talk a lot about data quality. Contacts between organisations in the chain are rather practical and usually concentrate on electricity-connections. Electricity connections have a European Article Number (EAN) that makes them able to be identified by all parties in the electricity chain. All the different parties have information about a particular EAN. EAN-codes are very important for the functioning of the (Dutch) electricity grid, but the information of the individual parties is not shared.

No incentive to improve

Taking a step back and looking at how data processing can be improved can be beneficial for the different parties. The structure of the Dutch electricity market is currently withholding parties to improve the data. District System Operators (DSOs) for instance execute their task that is demanded from them according to the law. DSOs are for example obliged by law to calculate averages of the electricity use. These electricity usage rates however are inaccurate and differ from the real usage data. DSOs do not gain by improving these data, therefore there is not a big pressing need to improve the data quality of the usage rates. The consequences is that the forecasting of expected energy usage is inaccurate because it is based upon wrong averages. Furthermore, another effect of this is when parties do not take on their responsibility electricity is allocated wrongly. This creates imbalances on the grid.

No access to database

There are different processes in the energy chain that are executed among the parties. Within the electricity chain there is a common database that stores a lot of energy about the connections: the C-AR. The different parties in the electricity chain however do not have access to the same amount of data in the C-AR. The companies that measure energy for instance are not able to change data in the C-AR without permission from the DSOs. A complicating factor is that within the chain a lot of data is only to be created, and maintained by one party. From interviews conducted with different parties can be concluded that for instance DSOs have a powerful role. Their importance is big for the execution of processes since they own the data however not always take responsibility.

No checking

Once data is entered it is rarely checked. This behaviour can be partly explained by the fact that it is rather costly to check data manually. Moreover, not all parties in the electricity chain have the same interest and within the energy sector it is hard to check the data since the EAN-connections are spread over the country.

The assigning of profiles is a process that directly influences the electricity that is allocated on the grid. Furthermore, there are not a lot of checks executed on the data. An example was that the energy company was mistaken in their calculation and charged around 3.000.000 kWh instead of the 3.000 kWh. Only after somebody at another organisation spotted the mistake manually it was corrected. Multiple types of mistakes are in the system because almost no checks are performed once the data is in the system. Incorrect ZIP-codes, house numbers, and meter numbers are not uncommon.

Opportunities

Multiple lessons have been identified that could improve the overall data quality in the electricity chain. Shifting certain responsibilities to other parties that are affected by the incorrect data is one of the possible examples. Even though this may seem obvious still some responsibilities are wrongly allocated. Another example is that data is rarely compared between organisations. For instance, conflicting information about EAN-codes have been found, which leads to incorrect billing.

A dominant way to measure data quality is through Data Quality Dimensions. Data quality is measured separately by organisations in a chain mainly based upon the four dimensions mentioned in paragraph two. However, for chains a broader approach should be take in order to unleash the full potential of the date. Data quality can be good according to one party but might not be sufficient for another party. Therefore, there is a need for more dimensions to be taken into account. Based on the literature and findings from the case study the following dimensions are suggested.

3-1 Reliability

"Reliability of data refers to the extent to which data is collected consistently over time and by different organisations either manually or electronically" (HIQA, 2011). When data is operated in a chain there is a big dependence between the parties. Reliability is an important dimension since a lack of it will not contribute to better data.

3-2 Comparability

"Comparability of data refers to the extent to which data is consistent between organisations over time allowing comparisons to be made." (HIQA, 2011). In the electricity chain there are different organisations having information about the same physical connection. It is therefore important that data can be compared between the parties in order to make sure there are not any differences.

3-3 Accessibility

"Accessibility of data refers to how easily it can be accessed, the awareness of data users of what data is being collected and knowing where it is located." (HIQA, 2011). Within the electricity chain there are multiple types of organisations involved. Access to the data and a spread knowledge of what information is available is therefore very important.

3-4 Credibility

"Credibility is used to evaluate non-numeric data. It refers to the objective and subjective components of the believability of a source or message." (Cai, 2015). When multiple organisations are working with the data it is important that the data that is exchanged is credible.

4 Conclusions

The main recommendation is that the responsibilities for data sometimes should be shifted. Different calculations in the chain are currently located at the wrong place with a party (DSO) that gains nothing from the correct execution. Multiple fields have been identified that could improve the overall data quality in the chain as has been described in the thesis of Pronk (Pronk, 2017).

An improved framework for data quality in the electricity chain was constructed. The presented framework (see figure 2) was derived from literature, interviews with different stakeholders and a case study in the Dutch electricity chain. These elements have been supporting the identification of improvements. The data quality dimensions that are useful for in a chain are

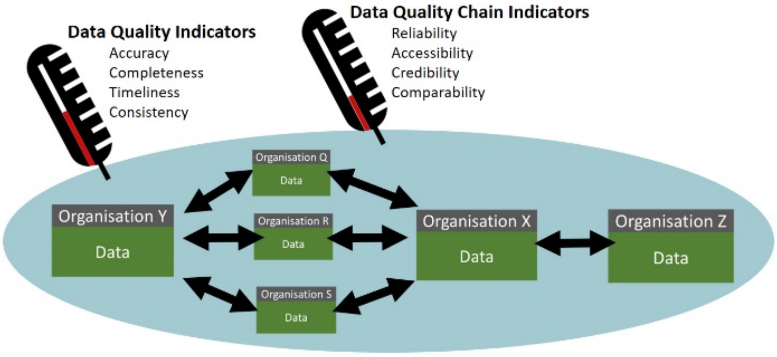


Figure 2:Framework for measuring data quality in a chain

reliability, accessibility, credibility, and comparability.

Reliability is needed since the parties are depending a lot on the data the other parties have provided. Comparability is needed in order to spot differences between different organizations. Accessibility is needed so the different parties have access to the data as well. Finally, credibility is needed since otherwise wrong data is entered into the system.

For the identified problems in the electricity chain these four dimensions were helpful to spot where the lower data quality was caused and for identifying options for improvement of data quality in the chain. Although a single case study is not hard evidence the findings might be well applicable to other sectors and data quality in chains, like production chains and logistics. Further research in these domains is recommended.

References

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A. (2009). Methodologies for data quality assessment and improvement. ACM computing surveys (CSUR), 41(3), 16. ISO 690

Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. Data Science Journal, 14.

Health Information and Quality Authority (Ireland). (2011). International review of data quality.

IBM. (2014). The four v's of big data. Retrieved from: http://www.ibmbigdatahub.com/infographic/four-vs-big-data.

Lee, I. (2017). Big data: Dimensions, evolution, impacts, and challenges. Business Horizons, 60(3), 293-303.

Pronk, M. (2017). Policy recommendations to improve data quality in the electricity chain.

Scannapieco, M., Missier, P., & Batini, C. (2005). Data quality at a glance. Datenbank-Spektrum, 14(January), 6-14. ISO 690

Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. Communications of the ACM, 41(2), 79-82.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. Journal of management information systems, 12(4), 5-33.

Weidema, B. P., & Wesnaes, M. S. (1996). Data quality management for life cycle inventories—an example of using data quality indicators. Journal of cleaner production, 4(3-4), 167-174.