A Semiotic Approach to Investigate Quality Issues of Open Big Data Ecosystems

John Krogstie and Shang Gao

Norwegian University of Science and Technology (NTNU)
{krogstie,shanggao}@idi.ntnu.no

Abstract. The quality of data models has been investigated since the mid-nineties. In another strand of research, data and information quality has been investigated even longer. Data can also be looked upon as a type of model (on the instance level), as illustrated e.g. in the product models in CAD-systems. We have earlier presented a specialization of the general SEQUAL-framework to be able to evaluate the combined quality of data models and data. In this paper we look in particular on the identified issues of 'Big Data'. We find on the one hand that the characteristics of quality of big data can be looked upon in the light of the quality levels of the SEQUAL-framework as it is specialized for data quality, and that there are aspects in this framework that are not covered by the existing work on big data. On the other hand, the exercise has resulted in a useful deepening of the generic framework for data quality, and has in this way improved the practical applicability of the SEQUAL-framework when applied to discussing and assessing quality of big data.

Keywords: Big data, data quality, Semiotic levels**.**

1 Introduction

The term 'big data' and the accompanying area have received increasing interest over the last years. The area is often defined through describing a number of V-s (Volume, Variety … etc). In a way, with an area where a major conference already in the seventies took the name VLDB - Very Large Data Bases one might wonder what is particularly different, and how the notion of data quality in connection to big data is different than data quality in general.

Data quality has for a long time been an established area of research [4] and work on quality assessment in data integration has also appeared as an area recently [24]. A related area that was established in the nineties is quality of models (in particular quality of conceptual data models) [25]. Data can be looked upon as a type of model (on the instance level), as illustrated e.g., in the product models in a CAD-system. Traditionally, one has looked at model quality for models on the M1 (type) level (to use the model-levels found in e.g., MOF). On the other hand, it is clear especially in product and enterprise modelling that there are models on the instance level (M0), an area described as containing data (or objects in MOF-terminology). Also if we look upon administrative data, e.g. data on persons, it is clear that this is an abstraction,

K. Liu et al. (Eds.): ICISO 2015, IFIP AICT 449, pp. 41–50, 2015.
© IFIP International Federation for Information Processing 2015

focusing on certain properties (e.g. name, age) of persons based on the purpose of having the data, not being a mirror of reality capturing all perceivable properties of persons. Thus, our outset is that also data quality can be looked upon relative to more generic frameworks for quality of models.

In this paper we look on characteristics of big data in the light of data quality as conceptualized using the SEQUAL-framework for quality of models. We will in section 2 describe some different work describing big data characteristics. In section 3 we described SEQUAL and its specialization for data quality based on earlier work. The main contribution of this work is section 4, where we position the big data characteristics within this semiotic framework. Section 5 concludes the paper.

2 Background on Big Data

Big Data has by many been 'conceptualized' by letter-magic centred on the letter V. Big Data first related to Volume, but later 3 V, 4 V and 5 V frameworks have been presented. As a start, Beyer and Laney [6] describe 3 areas:

- Volume, referring to the size of the data.
- Variety, referring to the heterogeneity of data representations, data acquisition, and semantic interpretation.
- Velocity, referring to the rate at which new data arrive and the time in which it must be acted upon.

IBM presents four areas[1]. In addition to the three above, they mention veracity, relating to the uncertainty of the data quality. Another source[2] also add a fourth area, namely viability pointing to that it is necessary to filter through all the data and carefully select the attributes and factors that most likely to predict outcomes and matter most to the business.

Finally Advanced Performance Institute [23] summarizes the area with 5 V-s, giving a bit more detailed information on the characteristics. We further exemplify with the current situation in the media industry.

- Volume refers to the large amounts of data generated going from are Terabytes to Zettabytes and more. This makes many datasets too large to store and analyse using traditional database technology. New big data tools use distributed systems so that we can store and analyse data across databases that are potentially spread around the world. In the news area, it means that larger volumes of data are available to be analysed relative to understanding what is happening.
- Velocity refers to the speed at which new data is generated and distributed. One example is a social media messages going viral in minutes. Technology allows us now to analyse the data while it is being generated, without ever putting it into databases. E.g. the SAP HANA architecture is providing an in-memory database solution that can integrate transactional data and analytical queries.

[1] http://www.ibmbigdatahub.com/infographic/four-vs-big-data
[2] http://www.pros.com/big-vs-big-data/

- Variety refers to the different types of data that one might want to look at in concert. In the past one mainly focused on structured data that fitted into tables or relational databases. A large amount of the world's data is unstructured (text, images, video, voice, etc.), kind of data that is particularly relevant in the media industry. With big data technology one can now analyse and bring together data of different types such as messages, social media conversations, photos, sensor data, video and voice recordings. Note that the variety aspect is not particular to Big Data, the issues is also found within large organizations in their attempt to address data integration [19, 24] internally or in collaboration with business partners.
- Veracity refers to the messiness or trustworthiness of the data. With many forms of big data, quality including accuracy are less controllable (just think of Twitter posts with hash tags, abbreviations, typos and colloquial speech as well as the reliability and accuracy of the content). In the media industry the veracity issue has become increasingly acute with the mix of journalistic and public sources of news.
- Value. Having access to big data is no good unless one can turn it into some value. This can be said to relate to Viability as described above. To have value in the news area, it is important to be able to report quickly based on what is happening. In newspapers we find for instance special systems such as Quake-Bot used by LA Times for more or less auto-generating news-articles based on available data. It can also be an issue as for the rights of reusing existing data to create value from the data.

On our own account, we add the need for visualization. To be able to get value from the data, it must be abstracted and visualized in an appropriate way to make the data useful, applying and extending techniques in the area of information visualization [35].

3 SEQUAL Data Quality Framework

SEQUAL [17] is a framework for assessing and understanding the quality of models and modelling languages. It builds on early work on quality of model [22, 26], but has been extended based on theoretical results [27, 28, 31] and practical experiences [17, 21] with the original framework. It has earlier been used for evaluation of modelling and modelling languages of a large number of perspectives, including data quality [18, 19], data modeling [16], ontologies [11], process modeling [1, 15], enterprise modeling [20], topological modeling (maps) [29] and goal-oriented modelling [13, 14]. Quality has been defined referring to the correspondence between statements belonging to the following sets:

- **G,** the set of goals of the modelling task.
- **L,** the language extension.
- **D**, the domain, i.e., the set of all statements that can be stated about the situation.
- Domains can be divided into two parts, exemplified by looking at a software requirements specification:

  - Everything the computerized information system (CIS) is supposed to do. This is termed the primary domain.
  - Constraints on the model because of earlier baselined models such as system level requirements specifications, enterprise architecture models, statements of work, and earlier versions of the requirement specification. This is termed the modelling context. In relation to data quality, the underlying data model is part of the modelling context when it is defined.

- **M**, the externalized model itself.
- **K**, the explicit knowledge relevant to the domain of the audience.
- **I**, the social actor interpretation of the model
- **T**, the technical actor interpretation of the model
- The main quality types following the steps of the so-called semiotic ladder [9] are:
- Physical quality: The basic quality goal is that the externalized model **M** is available to the relevant social and technical actors (and not to others).
- Empirical quality deals with comprehension and predictable error frequencies when a model **M** is read by different social actors
- Syntactic quality is the correspondence between the model **M** and the language extension **L**.
- Semantic quality is the correspondence between the model **M** and the domain **D**.
- Perceived semantic quality is the similar correspondence between the social actor interpretation **I** of a model **M** and his or hers current knowledge **K** of domain **D**.
- Pragmatic quality is the correspondence between the model **M** and the actor interpretation **(I** and **T)** and application of it.
- The goal defined for social quality is agreement among actor's interpretations.
- The deontic quality of the model relates to that all statements in the model **M** contribute to fulfilling the goals of modelling **G**, and that all the goals of modelling **G** are addressed through the model **M**.

3.1 Data Quality Relative to the SEQUAL Quality Types

We here discuss means within each quality level, positioning the areas that are specified by Batini et al. [4] and Price et al. [10, 11]. These are emphasised using italic. This overview is largely taken from [18].

Physical Data Quality: Aspects of persistence, data being accessible (Price) for all (accessibility (Batini)), currency (Batini) and security (Price) cover aspects on the physical level. This area can be looked upon relative to measures of persistence, currency and availability that apply also to all other types of models.

Empirical Data Quality: This is addressed by understandable (Price). Since data can be presented in many different ways, this relates to how the data is presented and visualized. How to best present different data depends on the underlying data-type. There are a number of generic guidelines within data visualization and related areas that can be applied, and we will only mention a few of these here. For computer-output specifically, many of the principles and tools used for improving human computer interfaces are relevant at the empirical level [33]. For visual presentation of data, one can also base the guidelines on

work in cognitive psychology and cartography with the basis that data is meant to be useful in connection to communication between people.

Syntactic Data Quality: From the generic SEQUAL framework we have one main syntactic quality characteristics, syntactical correctness. This means that all statements in the model are according to the syntax and vocabulary of the language

Syntax errors are of two kinds:

- Syntactic invalidity, in which graphemes not part of the language are used.
- Syntactic incompleteness, in which one lack constructs or information to obey the language's grammar

Conforming to metadata (Price) including that the data conform to the expected data type of the data (as described in the data model) are part of syntactic data quality. This will typically be related to syntactic invalidity when e.g. the data is of the wrong data-type.

Semantic Data Quality: When looking upon semantic quality relative to the primary domain of modelling, we have the following properties:

- Completeness in SEQUAL is covered by completeness (Batini), mapped completely (Price)**,** and mapped unambiguously (Price)**.**
- Validity in SEQUAL is covered by accuracy (Batini), both syntactic and semantic accuracy as they have defined it, the difference between these is rather to decide on how incorrect the data is, phenomena mapped correctly (Price), properties mapped correctly (Price) and properties mapped meaningfully (Price). Since the rules of representation are formally given, consistency (Batini)/mapped consistently (Price) is also related to validity. The use of metadata such as the source of the data is an important mean to support validity evaluation of the data.

Properties related to the model context are related to the adherence of the data to the data model. One would expect for instance that

- All tables of the data model should include tuples
- Data is according to the constraints defined in the data-model

The possibility of ensuring high semantic quality of the data is closely related to the semantic quality of the underlying data model. When looking upon semantic quality of the data model relative to the primary domain of modelling, we have the following properties: Completeness (Moody and Batini) (number of missing requirements) and integrity (Moody) (number of missing business rules) relates to completeness.

Completeness (Moody) (number of superfluous requirements) and integrity (Moody) (number of incorrect business rules) relates to validity. The same applies to Batini's points on correctness with respect to model and correctness with respect to requirements**.**

Pragmatic Data Quality: Pragmatic quality relates to the comprehension of the model by participants. Two aspects can be distinguished:

- That the interpretation by human stakeholders of the data is correct relative to what is meant to be expressed. In addition to the data it will often be useful to have different meta-data represented (making it easier to understand the intent behind the data).
- That the tool interpretation is correct relative to what is meant to be expressed.

Starting with the human comprehension part, pragmatic quality on this level is the correspondence between the data and the audience's interpretation of it. Moreover, it is not only important that the data has been understood, but also who has understood (the relevant parts of) the data.

The main aspect at this level is interpretability (Batini), that data is suitably presented (Price) and data being flexibly presented (Price). Allowing access to relevant metadata (Price) is an important mean to achieve comprehension.

Social Data Quality: The goal defined for social quality is agreement. Relative agreement means that the various sets to be compared are consistent -- hence, there may be many statements in the data representation of one actor that are not present in that of another, as long as they do not contradict each other. The area quality of information source (Batini) touches important means for the social quality of the data, since a high quality source will increase the probability of agreement. Another term found in data quality literature on this aspect is provenance.

In some cases one need to combine different data sources. This consists of combing the data-models, and then transferring the data from the two sources into the new schema. Schema integration techniques [10] are specifically relevant for this area.

Deontic Data Quality: A number of aspects are on this level relating to the goals of having the data in the first place. Aspects do decide volatility (Batini) and timeliness (Batini)/ timely (Price) needs to relate to the goal of having and distributing the data. The same is the case for type-sufficient (Price), the inclusion of all the types of information important for its use. For anything but extremely simple and highly inter-subjectively agreed domains, total validity, completeness, comprehension, and agreement as described above under semantic, pragmatic and social quality cannot be achieved. Hence, for the goals on these levels to be realistic, they have to be relaxed, by introducing the idea of feasibility. The time to terminate a modelling activity is thus not when the model is ``perfect'' (which will never happen), but when it has reached a state where further modelling is regarded to be less beneficial than applying the model in its current state. Accordingly, a relaxed kind of these goals being dependent on human judgement can be defined, which we term feasible validity, feasible completeness, feasible comprehension, and feasible agreement. Feasibility thus introduces a trade-off between the value and drawbacks for achieving a given model quality. When we structure different aspects according to these levels, one will find that there might be conflicts between the levels (e.g., what is good for semantic quality might be bad for pragmatic quality and vice versa).

4 Applying SEQUAL on Big Data Characteristics

On a high level, we can position the big data characteristics described in section 2 in relation to SEQUAL data quality framework in the following way:

Physical Quality: Volume is particularly relevant on this level, since it can be hard to have access to all relevant data at the same time. That it is the right (most current) data that is accessible is influenced by the velocity of data change. Supporting provenance, it might also be necessary to store the full chain of the data revisions, and not only the last version. In general, provenance meta-data should be represented independent on the technologies used for data storage, e.g. by using PROV3. An area that is not so much discussed in the big data literature is aspects of security, although in particular the use of Big Data-oriented techniques on personal data is rife with privacy-challenges. If people are more aware of this, potentially more people will make it more difficult for those working with Big Data techniques to get access to all the data that is of interest, indicating a need to also be open on how Big Data (e.g. location data) is meant to be used in a preferably anonymous manner in doing analysis [7].

Empirical Quality: Visualization of data is typically done on the basis of guidelines from cognitive psychology, and applying these is a good approach for achieving pragmatic quality (i.e. that people understand what the accumulated data mean). Note that guidelines for aesthetics are partly incompatible, and one has to make choices based on the usage and interpreters of the representation. In connection to maps [32] states that "different combinations, amounts of application, and different orderings of these techniques can produce different yet aesthetically acceptable solutions". Since the data visualization often must be auto-generated (to address issues of velocity), aspects described under this level is even more important for pragmatic quality than for traditional models developed mostly manually by human modelers.

Syntactic Quality: Variety comes into play here, since not all data sources have a strictly defined meta-model with a pre-defined syntax. This means that to match the different data-sources, certain presumptions have to be made on the structure and contents of data, i.e. one need to instil structure if it is not there, and in some cases meaning (see semantic quality) to data based on at best qualified guesses.

Semantic Quality: Whereas traditional data quality aspects such as completeness, accuracy and consistency are not covered specifically in Big Data - literature, the area veracity points to data quality more generally. A reason for the wanted variety is to ensure completeness since not all relevant data is to be found in one data source. Variety on the other hand brings traditional challenges in data integration quality [24] matching data on different level of abstraction and preciseness. When data is retained from sensor networks, one might experience issues of redundancy (e.g. reporting location every second from an object that is not moving). Such redundancies should be filtered out as should erroneous readings due to noise, e.g. indicating than an object suddenly moved a large distance in a short time, and this filtering should be done in the right order. To avoid issues of poor physical quality, one might often abstract the data, in which case it is important that the resulting dataset keep the important characteristics of the original dataset [34]. This points to an interesting side of big data not experienced in traditional modeling and data representations, namely that the modeling is partly done by algorithms, and not only be humans.

Pragmatic Quality: Relates both to machine understanding of data sources, and of human understanding of the results. As for machine understanding, the issues here is

[3] http://www.w3.org/TR/prov-dm/

very different for different types of data (e.g. between structured and unstructured data such as text, video etc.). When it comes to human understanding of the results, this is primarily supported by taking empirical quality into account when devising visualizations of the data. Another approach that can be used is to provide personalized output, in which case it might be important to make the user-model used in the personalization controllable for the user [3]

Social Quality: Provenance relating to the trustworthiness of the source as part of veracity is central at this level. Also in combination with variety (including data coming from a variety of sources that is evolving in an uncoordinated fashion by autonomous agents constituting parts of a digital ecosystem), we get apparently new issues, since some sources might be more trustworthy than other. Also internally in organizations this might be an issue, e.g. matching personal data in spreadsheets, with data from enterprise systems such as a PLM-system [19]. Since these sources are in the same organization though, the possibility to enforce compliance is larger than in a Big Data setting. Due to velocity aspects, one might need to quickly and automatically deduce a trust-level using a trust model [2] based on existing meta-data on the data-source.

Deontic quality is closely related to what is meant under the point value, are we able to utilize the data for our purpose? Viability is a sub-point of this relating to the discussion of feasible quality in SEQUAL. Based on the goal of data-use, and also partly dependent on data sources to be matched, different weight might be put on different quality levels. A framework for personalization of big data quality deliberations is found in [8], using scientific data as a case.

5 Conclusions and Further Work

We observe on a high level that the V-characteristics of big data can be mapped and understood relative to the levels of model and data quality described in the SEQUAL framework, although the mapping is not 1-to-1. The focus areas of big data do not describe to the same detail though many of the aspects found relevant in work of information and data quality. Issues with variety in Big Data can often be related to issues found in data integration quality in general, but we have also identified new issues from Big Data research when combining the effect of several V-s.

We notice that the treatment of quality in the Big Data area is so far relatively shallow. A first step is to a larger extent describe quality of Big Data relative to traditional data quality. Future work on our general approach on data quality will be to device more concrete guidelines and metrics and evaluate the adaptation and use of these empirically in case studies, also studies dealing with big data - issue e.g. within the media domain. An important aspects is how to perform trade-offs between the different data quality types Some generic guidelines for this exist in SEQUAL [17], which might be specialised for data quality. We will also look at newer work [5, 12] in the area in addition to the ones we have mapped so far. Due to the rapid changes to big data compared to conceptual models indicates that guidelines for achieving and keeping model quality might need to be further adapted to be useful when achieving and keeping quality of big data ecosystems, especially since a lot of what is normally modelled by humans in a big data scenario are modelled through the use of complex algorithms.

References

1. Aagesen, G., Krogstie, J.: Analysis and design of business processes using BPMN. In: vom Brocke, J., Rosemann, M. (eds.) Handbook on Business Process Management. Springer (2010)
2. Artz, D., Gil, Y.: A survey of trust in computer science and the semantic web. Web Semantics: Science. Services and Agents on the World Wide Web 5(2), 58–71 (2007)
3. Asif, M., Krogstie, J.: Externalization of User Model in Mobile Services. International Journal of Interactive Mobile Technologies (iJIM) 8(1), 4–9 (2014)
4. Batini, C., Scannapieco, M.: Data Quality: Concepts, Methodologies and Techniques. Springer (2006)
5. Batini, C., Cappiello, C., Francalanci, C., Maurino, A.: Methodologies for data quality assessment and improvement. ACM Comput. Surv. 41(3) (2009)
6. Beyer, M. A., Laney, D.: The importance of'big data': a definition. Stamford. Gartner, CT (2012)
7. Biczok, G., Martinez, S.D., Jelle, T., Krogstie, J.: Navigating Mazemap: Indoor human mobility, spatio-logical ties and future potential. PERMODY IEEE (2014)
8. Embury, S.M., Missier, P., Sampaio, S., Greenwood, R.M., Preece, A.D.: Incorporating domain-specific information quality constraints into database queries. Journal of Data and Information Quality (JDIQ) 1(2), 11 (2009)
9. Falkenberg, E.D., Hesse, W., Lindgreen, P., Nilsson, B.E., Oei, J.L.H., Rolland, C., Stamper, R.K., Assche, F.J.M.V., Verrijn-Stuart, A.A., Voss, K.: A Framework of information system concepts - IFIP WG 8.1 Task Group FRISCO (1996)
10. Francalanci, C., Pernici, B.: View integration: A survey of current developments. Technical Report 93-053, Politecnico de Milano, Milan, Italy (1993)
11. Hella, L., Krogstie, J.: A Structured Evaluation to Assess the Reusability of Models of User Profiles. Paper presented at the EMMSAD Hammamet, Tunis, 7-8/6 (2010)
12. Jiang, L., Barone, D., Borgida, A., Mylopoulos, J.: Measuring and Comparing Effectiveness of Data Quality Techniques. In: van Eck, P., Gordijn, J., Wieringa, R. (eds.) CAiSE 2009. LNCS, vol. 5565, pp. 171–185. Springer, Heidelberg (2009)
13. Krogstie, J.: Using Quality Function Deployment in Software Requirements Specification. Paper presented at the Fifth International Workshop on Requirements Engineering: Foundations for Software Quality (REFSQ 1999), Heidelberg, Germany, June 14-15 (1999)
14. Krogstie, J.: Integrated Goal, Data and Process Modeling: From TEMPORA to Model-Generated Work-Places. In: Johannesson, P., Søderstrøm, E. (eds.) Information Systems Engineering From Data Analysis to Process Networks, pp. 43–65. IGI (2008)
15. Krogstie, J.: Quality of Business Process Models. Proceedings PoEM 2012, Rostock Germany LNBIP (2012)
16. Krogstie, J.: Quality of Conceptual Data Models. Proceedings 14th ICISO, Stockholm Sweden (2013)
17. Krogstie, J.: Model-based development and evolution of information systems: A quality approach. Springer, London (2012)
18. Krogstie, J.: A Semiotic Framework for Data Quality. Proceedings EMMSAD 2013, Valencia, Spain (June 2013)
19. Krogstie, J.: Evaluating Data Quality for Integration of Data Sources. In: Proceedings PoEM 2013, Riga, Latvia, pp. 39–53 (2013)
20. Krogstie, J., Arnesen, S.: Assessing Enterprise ModelingLanguages using a Generic Quality Framework. In: Krogstie, J., Siau, K., Halpin, T. (eds.) Information Modeling Methods and Methodologies. Idea Group Publishing (2004)

21. Krogstie, J., Dalberg, V., Jensen, S.M.: Process modeling value framework. In: Manolopoulos, Y., Filipe, J., Constantopoulos, P., Cordeiro, J. (eds.) Selected papers from 8th International Conference, ICEIS 2006. LNBIP, vol. 3, pp. 309–321. Springer, Heidelberg (2008)
22. Lindland, O.I., Sindre, G., Sølvberg, A.: Understanding Quality in Conceptual Modelling,". IEEE Software 11(2), 42–49 (1994)
23. Marr, B.: Big Data (2014), http://www.ap-institute.com/
24. Martin, N., Poulovassillis, A., Wang, J.: A Methodology and Architecture Embedding Quality Assessment in Data Integration. ACM Journal of Data and Information Quality 4(4) (2012)
25. Moody, D.L.: Metrics for Evaluating the Quality of Entity Relationship Models. In: Ling, T.-W., Ram, S., Li Lee, M. (eds.) ER 1998. LNCS, vol. 1507, pp. 211–225. Springer, Heidelberg (1998)
26. Moody, D.L., Shanks, G.G.: What Makes a Good Data Model? Evaluating the Quality of Entity Relationship Models. In: Loucopoulos, P. (ed.) ER 1994. LNCS, vol. 881, pp. 94–111. Springer, Heidelberg (1994)
27. Moody, D.L.: Theoretical and practical issues in evaluating the quality of conceptual models: Current state and future directions. Data and Knowledge Engineering 55, 243–276 (2005)
28. Nelson, H.J., Poels, G., Genero, M., Piattini, M.: A conceptual modeling quality framework. Software Quality Journal 20, 201–228 (2012)
29. Nossum, A., Krogstie, J.: Integrated Quality of Models and Quality of Maps. In: Halpin, T., Krogstie, J., Nurcan, S., Proper, E., Schmidt, R., Soffer, P., Ukor, R. (eds.) Enterprise, Business-Process and Information Systems Modeling. LNBIP, vol. 29, pp. 264–276. Springer, Heidelberg (2009)
30. Price, R., Shanks, G.: A Semiotic Information Quality Framework. In: IFIP WG8.3 International Conference on DecisionSupport Systems (DSS 2004), Prato, Italy, 1-3, pp. 658–672 (2004)
31. Price, R., Shanks, G.: A semiotic information quality framework: Development and comparative analysis. Journal of Information Technology 20(2), 88–102 (2005)
32. Shekhar, S., Xiong, H.: Encyclopedia of GIS. Springer (2008)
33. Shneiderman, B.: Designing the User Interface: Strategies for Effective Human- Computer Interaction, 2nd edn. Addison Wesley, Reading (1992)
34. Wad, C.: QoS: Quality Driven Data Abstraction for Large Databases. Worcester Polytechnic Institute (2008)
35. Ware, C.: Information Visualization. Morgan Kaufmann (2000)