Linking Trust to Data Quality

Davide Ceolin, Valentina Maccatrozzo, and Lora Aroyo

VU University Amsterdam
de Boelelaan, 1081a
1081HV Amsterdam, The Netherlands
{d.ceolin, v.maccatrozzo, lora.aroyo}@vu.nl

Abstract. Trust in data is a user-oriented and subjective phenomenon that hooks on specific data qualities. This paper lays the foundations for studying the existence of a correlation between data qualities and trust. In particular, if these data qualities can be measured by relying on the data only, then it would be possible to infer the likelihood of a given piece of data to be trusted based on a specific measurement made on the data itself. Here we provide a categorization of data quality dimensions as intrinsic (i.e., measurable by analyzing the data only) and extrinsic, and we model it using the W3C Data Quality Vocabulary. We show how it is possible to align credibility, an extrinsic data quality, with trust. This alignment will serve as a basis for a future investigation about the existence of a correlation between trust and intrinsic data qualities.

1 Introduction

Data quality is a crucial matter because it provides the basis for users to decide whether they can safely and reliably use data in their decision and productive processes. Often, users decide to trust or not (Web) data based on the value of specific data and metadata qualities, like accuracy, precision, and others. Trust is a complex phenomenon that implies a user (trustor) attitude towards third party (trustee, which can be also a piece of data), followed by a trust action (i.e., the actual action to trust or not the trustee) in a specific context. Due to its subjective nature, trust attitudes and actions are challenging to estimate and predict. However, we hypothesize that there exist intrinsic data qualities (i.e., data qualities which, to be measured, do not require external input) that correlate with trust, and the goal of this paper is to lay the foundations for investigating this aspect in depth. The contribution of this paper is twofold: (1) a categorization of data quality dimensions as intrinsic and extrinsic, and (2) an alignment between trust and data quality, through the extrinsic data quality credibility. By approximating a data quality from a set of subjective trust judgments, we prepare the ground for studying the existence of a correlation between intrinsic data qualities and trust.

For the definitions of data quality we refer to the ISO model 25012 [5], and we extend it to include additional data quality dimensions specifically used in the Semantic Web (e.g., for information retrieval or recommendation purposes).

Regarding the modeling of data quality, we refer to the W3C Data Quality Vocabulary (DQV) [1]. For the modeling of trust, we refer to the ontology for trust in data outlined in a previous work of ours [3], which models trust by extending the trust ontology provided by Alnemr et al. [2] with elements of the theory of O'Hara [11].

The rest of the paper is structured as follows: Section 2 presents related work. Sections 3 and 4 discuss the categorization of data quality categories, and Section 5 presents an alignment of data quality and trust in data based on such categorization. Lastly, Section 6 presents a final discussion and future work.

2 Related Work

The relation between data quality and trust has been investigated from different points of view, for instance in the medical domain [8], e-commerce [12], and on scientific data [4]. Also, Wang and Strong [14] draw a link between trust and data quality in an e-commerce setting. Strong et al. [13] analyze, instead, the importance and the impact of data quality in different contexts, consequently linking it to trust.

The ISO Model 25012 defines fourteen different data quality dimensions. The starting point for our categorization are those from ISO itself [5] and Natale [9]. We differentiate from them because our categorization relies on the provenance of the data measurements.

Lee et al. [7] define an information quality categorization that contains also the 'Intrinsic' category. This is defined as the category of information qualities that "implies that information has quality in its own right". This category includes also accuracy. However, we classify accuracy as an extrinsic data quality, because we consider that accuracy, to be measured, needs external references, i.e., we base our distinction on provenance. We will motivate this aspect more deeply in Section 4. Also, we focus on data, and not on information quality.

Naumann [10] defines four categories of information quality: Content-related, Technical, Intellectual and Instantiation-related. This categorization is orthogonal with respect to the one we propose here. In fact, this categorization applies to information qualities (and, again, not data qualities), and is determined on the specific aspect evaluated by the quality (e.g., information content). Our categorization is determined by the kind of characteristics we need to consider in order to measure a given quality, i.e., on the provenance of the data quality measurement.

3 Intrinsic Data Quality Dimensions

Intrinsic data quality dimensions are those dimensions that can be measured by means of metrics that depend entirely on the data characteristics. Any metric could, in principle, be enriched with information from the external world (e.g., for calibration). However, the peculiarity of this dimension is that it is measurable by means of metrics that apply on the data measured only. The SPARQL query

reported in listing 1.1 outlines the characteristics of this category of dimensions using DQV.

```
PREFIX prov:<http://www.w3.org/ns/prov#>
PREFIX dqv:<http://www.w3.org/ns/dqv#>
PREFIX daq:<http://purl.org/eis/vocab/daq#>

SELECT ?category
WHERE {?category a daq:Category .
        ?dimension dqv:hasCategory ?category .
        ?metric daq:hasDimension ?dimension .
        ?qualityMeasure daq:hasMetric ?metric .
        ?qualityMeasure daq:computedOn ?dataset .
        ?qualityMeasure prov:wasDerivedFrom ?dataset .
        ?qualityMeasure prov:wasDerivedFrom ?external .
        FILTER NOT EXISTS {
         ?qualityMeasure prov:wasDerivedFrom ?external
         FILTER(?external != ?dataset) })} .
```

Listing 1.1. SPARQL 1.1 query to retrieve intrinsic data quality categories. The variable ?external is empty for intrinsic data qualities because it indicates any input other than the dataset itself.

An example of an intrinsic dimension is *efficiency*. The ISO model 25012 defines efficiency as "*The degree to which data has attributes that can be processed and provide the expected levels of performance by using the appropriate amounts and types of resources in a specific context of use.*" If we analyze a given dataset, we can have a rough measure of efficiency by computing the Kolmogorov complexity [6] or by simply computing the ratio of null cells contained in the dataset or whether the range allowed by the datatypes used allows larger values than the range of the values actually stored. More refined efficiency values could be computed by comparing these values with those from a corpus of datasets, but still, a measure of efficiency could be computed by analyzing a dataset alone.

An example of intrinsic data quality measure often employed in semantically-enabled recommender systems is *diversity*. Suppose we have an RDF dataset consisting of LOD patterns linking movies. The start and the end of these patterns are movies, and the paths that link them are LOD semantic patterns. Using these patterns we can cluster movies, and evaluate them in terms of diversity. Diversity could be measured in syntactic terms, or in a more elaborated way by making use of semantic similarity (sim) or of other types of measure, like in the case of the following diversity measure we developed:

$$Div(p_1, p_2) = \frac{(1 - sim(genre(p_1), genre(p_2))) + (1 - sim(topic(p_1), topic(p_2)))}{2}$$

This data quality measure requires the data and the related metadata (genre, topic) in order to be evaluated. Of course, it is possible to extend it and improve it, also with external inputs, but the formula above allows computing item diversity and requires no external information, except the items metadata.

4 Extrinsic Data Quality Dimensions

We define extrinsic data quality dimensions as those data quality dimensions that can only be measured by means of metrics that make use of a necessary external input. Listing 1.2 shows a SPARQL query that identifies extrinsic data quality dimensions. Consider, for instance, *accuracy*, which is defined by ISO 25012 as "*The degree to which data has attributes that correctly represent the true value of the intended attribute of a concept or event in a specific context of use*". To evaluate the accuracy of a piece of data, we need to know the truth value of a given concept in a given context. Without this information, it is impossible to estimate this quality value.

For instance, to determine if:

```
exMuseum:ParisPainting ex:depicts dbpedia:Paris
```

is accurate, we need to check if the first URI actually corresponds to a painting, if it depicts a city and if that city is Paris, the same city referred to by the object of the triple. If we do not have access to this information, we cannot compute the metric value. For this reason, although it is true that the quality regards property of the data itself, we classify it as an extrinsic one.

```
PREFIX prov:<http://www.w3.org/ns/prov#>
PREFIX dqv:<http://www.w3.org/ns/dqv#>
PREFIX daq:<http://purl.org/eis/vocab/daq#>

SELECT ?category
WHERE { ?category a daq:Category .
        ?dimension dqv:hasCategory ?category .
        ?metric daq:hasDimension ?dimension .
        ?qualityMeasure daq:hasMetric ?metric .
        ?qualityMeasure daq:computedOn ?dataset .
        ?qualityMeasure prov:wasDerivedFrom ?dataset .
        ?qualityMeasure prov:wasDerivedFrom ?external .
        FILTER ( ?external != ?dataset) .} .
```

Listing 1.2. SPARQL 1.1 query to retrieve extrinsic data quality categories. ?external is an input entity which is meant to be different from the dataset itself.

5 Aligning Data Quality and Trust in Data

As stated in Section 1, trust is a complex phenomenon that is described by the trustor attitude towards the trustee, along with his or her actual decision to trust the trustee or not. For instance, in a previous work [3], we defined the trust attitude that a given user has with respect to an RDF triple $\langle spo \rangle$ as the belief that, given the subject s and the property p, an object o exists:

$$TrustAttitude_{trustor}(o|s,p) = Belief_{trustor}(\exists(o) : \langle (s), I(o) \rangle \in IEXT(I(p))$$

In other words, the trust attitude can be considered as the belief of the trustor in the *accuracy* (or truthfulness) of the triple. Such an attitude is determined on subjective bases but we hypothesize that it correlates with specific data or metadata qualities. This means that we suppose that users obtain evidence against or in favor of the trustworthiness of a piece of data by analyzing specific data or metadata qualities. For instance, a user can have a positive trust attitude towards a piece of data because she trusts its creator. Another user might trust it because she measured the efficiency of that data, and based on her evidence, that efficiency value correlates with accuracy. So, the data characteristics that can influence the trust attitude are several and depend both on the user (trustor) and on the context. Nevertheless, some data are more likely to be trusted than other, and we suppose that this is because some data qualities correlate with trustworthiness. Thus, we can link trust attitude to *Credibility*, which is defined in the ISO 25012 model as "*The degree to which data has attributes that are regarded as true and believable by users in a specific context of use.*", as follows:

$$Credibility(o|s,p) \approx min_{trustor \in Context}(TrustAttitude_{trustor}(o|s,p)) \quad (1)$$

In this manner, *Credibility* is represented as a quality of the triple, which is bound to the trust attitude that each user in a given context has in the data. By having at our disposal the trust attitude value on a given piece of data provided by a lot of users, we can approximate its *Credibility*. This would provide us the basis for exploring in depth which data qualities correlate with trust in data. Equation (1) can be generalized as follows:

$$Credibility(o|s,p) \approx Agg_{trustor \in Context}(TrustFunction_{trustor}(o|s,p)) \quad (2)$$

In fact, several aggregation functions (*Agg*) and several trust functions (*TrustFunction*) are utilizable in this context. In Equation (1) we refer to the trust attitude (i.e., the "intention to trust" [3]), but we could aggregate (e.g., using the *average* function) the trust actions (i.e., whether users do actually trust the triple or not [3]) or a combination of the two. The specific implementation of Equation (2) depends on the data available and on the requirements faced. However, the importance of this formula lies in the fact that it allows to bind subjective trust evaluations to the approximation of an extrinsic data quality. By comparing these approximations with the values of intrinsic data qualities (which require no other input than the data themselves to be computed), we can check if any correlation exists.

6 Discussion and Future Work

In this paper, we lay the foundations for the study of the correlation between intrinsic data qualities and trust in data. We do so by first providing a categorization of data qualities as intrinsic and extrinsic. This categorization is modeled using the W3C Data Quality Vocabulary. Each category is defined by means of

a SPARQL query. Then, we approximate an extrinsic data quality (*Credibility*) as an aggregation of subjective trust judgments in the data. This allows us to link subjective and user-oriented evaluations to data qualities. This step is crucial for the final goal of the work introduced by this paper, that is studying the correlation between intrinsic data qualities and trust. If this correlation exists, then by observing the values of intrinsic data qualities we could infer priors or probabilistic estimates of the trust that users have in the corresponding data.

We plan to investigate this matter thoroughly, both with respect to the classification of data qualities and regarding the links between data quality and trust. In particular, we aim at studying the correlation between *Credibility* and intrinsic data qualities using different implementations of Equation 2. Finally, we will extend this study to other extrinsic data qualities, besides *Credibility*.

Acknowledgements This work is funded by the Dutch national research funding programme COMMIT.

References

1. R. Albertoni, C. Guéret, and A. Isaac. Data quality vocabulary. `http://www.w3.org/TR/2015/WD-vocab-dqv-20150625/`, 2015.
2. R. Alnemr and C. Meinel. From reputation models and systems to reputation ontologies. In *IFIPTM*, volume 358, pages 98–116. Springer, 2011.
3. D. Ceolin, A. Nottamkandath, W. Fokkink, and V. Maccatrozzo. Towards the definition of an ontology for trust in (web) data. In *URSW*, pages 73–78, 2014.
4. M. Gamble and C. Goble. Quality, trust, and utility of scientific data on the web: Towards a joint model. In *WebSci '11*, pages 15:1–15:8, 2011.
5. ISO/IEC 25012:2008. *Software engineering - Software product Quality Requirements and Evaluation (SQuaRE) - Data quality model.* ISO, 2008.
6. A. Kolmogorov. On Tables of Random Numbers. *Theor. Computer Science*, 207(2):387395, 1998.
7. Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang. Aimq: a methodology for information quality assessment. *Information & Management*, 40(2):133–146, 2002.
8. G. Lippi, K. Becan-McBride, D. Behúlová, R. A. Bowen, S. Church, J. Delanghe, K. Grankvist, S. Kitchen, M. Nybo, M. Nauck, N. Nikolac, V. Palicka, M. Plebani, S. Sandberg, and S. A. Preanalytical quality improvement: in quality we trust. *Clin. Chem. and Lab. Med.*, 51(1):229–241, 2012.
9. D. Natale. Complexity and data quality. In *CHItaly*, 2011.
10. F. Naumann. *Quality-Driven Query Answering for Integrated Information Systems.* 2001.
11. K. O'Hara. A General Definition of Trust. Technical report, University of Southampton, 2012.
12. D. Ribbink, A. C. van Riel, V. Liljander, and S. Streukens. Comfort your online customer: quality, trust and loyalty on the internet. *Managing Service Quality: An International Journal*, 14(6):446–456, 2004.
13. D. M. Strong, Y. W. Lee, and R. Y. Wang. Data quality in context. *Commun. ACM*, 40(5):103–110, 1997.
14. R. Y. Wang and D. M. Strong. Beyond accuracy: What data quality means to data consumers. *J. Manage. Inf. Syst.*, 12(4):5–33, 1996.