Data quality in governance: A definition and a research agenda

CURATE Project
https://curate-project.com/

Data quality in governance:
A definition and a research agenda

CURATE Working Paper No. 1 – November 2022

Matthias Leese
mleese@ethz.ch

This Working Paper explores the concept of data quality as an important yet so far understudied aspect in the social sciences, especially in regard to forms of governance that are predicated on data. To do so, it first provides a definition of data quality and discusses its most pertinent dimensions. The second aim is to outline a research agenda for the study of data quality from a social science perspective. Such a research agenda, so the argument put forward here, must include both the politics of data quality, i.e. the regulatory mechanisms around the setting of standards and procedures, as well as the situated practices of those who are involved in data quality activities on an everyday basis.

"A closer look [..] reveals certain shortcomings that illustrate the need to improve the quality and coverage of the data collected. Tragic cases that have caught the public's attention have highlighted the insufficient availability of important national information in European databases" (Council of the European Union, 2020: 2). This is how the Presidency of the Council of the European Union in September of 2020 framed the current state of play with regard to the centralized databases that are considered key for Justice and Home Affairs in the EU. Looking specifically at the Schengen Information System (SIS II) for law enforcement, border control, and judicial cooperation and the Europol Information System (EIS) for information exchange in criminal investigations, the Presidency came to the conclusion that dedicated action was required for "improving the quantity and quality of the data in the EIS and SIS" (Council of the European Union, 2020: 2) and instructed the Working Party on Justice and Home Affairs Information Exchange (IXIM) to conduct a survey regarding data quality among member states and to set up a European Data Quality Day (EDQD). Since then, the EU has adopted an Implementing Decision on "the automated data quality control mechanisms and procedures, the common data quality indicators and the minimum data quality standards" that are in the future to be applied to all data that are

transferred from the EU Member States to centrally managed Justice and Home Affairs databases (European Union, 2021a, 2021b).

As the Council's problem definition and ensuing remedial actions demonstrate, a so far largely neglected aspect of data has by now been put on the political agenda: their quality. After years of more or less uncurbed enthusiasm about "Big Data", artificial intelligence, and interoperability as enablers of new business models and forms of governance, questions about the actual reliability and trustworthiness of the data that underpin knowledge and action are increasingly becoming pertinent. This seems only logical: as the production and availability of data have been scaled up massively, so have the potential shortcomings within datasets. And while some use-cases for large datasets, for example targeted advertisements based on the analysis of interaction patterns, are not necessarily reliant on high data quality due to their error tolerance (i.e. nobody is harmed if they are shown an advertisement that does not actually correspond with their consumer preferences), the interface between the state and its citizens comes with ultimately high data quality requirements. Faulty data in public administration systems, for example, can mean that citizens might not get the benefits that they are entitled to, and faulty data in databases for Justice and Home Affairs can mean that threats might be wrongly assessed and that innocent people might end up in the cross-hair of security authorities.

This Working Paper has two aims. The first one is to explore the concept of data quality as an important yet so far understudied aspect in the social sciences, especially in regard to forms of governance that are predicated on data. To do so, it first provides a definition of data quality based on computer science and management literature and subsequently discusses the most pertinent dimensions of data quality. The second aim is to outline a research agenda for the study of data quality from a social science perspective. Such a research agenda, so the argument put forward here, must include both the politics of data quality, i.e. the regulatory mechanisms around the setting of standards and procedures, as well as the situated practices of those who are involved in data quality activities on an everyday basis.

Defining data quality

In their essence, data are the translation of empirical phenomena into another form. Humans have throughout history always produced data – and they have increasingly done so in more sophisticated, systematic, formalized, and scalable forms that have enabled novel ways of statistically empowered techniques of state-citizen interaction and governance (Desrosières, 2002; Hacking, 1990). A major caveat is, however, presented by considerations about what the "correct" form of representation would be in the first place. Empirical phenomena can be translated into data in many different ways, each one resulting in idiosyncratic structures, properties, and informational value. Data are, in other words, not self-evident but, as Gitelman and Jackson (2013: 3) have argued, need to be "imagined and enunciated against the seamlessness of phenomena" before they are crafted. Such imagination usually takes place in the context of specific use-cases, meaning that the form and informational value of data are shaped backwards from how they are to be used later on.

The data in the SIS II are, for example, to a large extent coined by the identification and investigation practices of European law enforcement and border control agencies that require truthful and actionable information about people and objects allegedly linked to crime and terrorism. The data stored in the

SIS II accordingly include biographical, biometric, and further descriptive categories (photographs, investigative knowledge) about persons who are for example reported as missing, wanted for judicial assistance in court proceedings, or suspected to be implicated in criminal activities. Moreover, the system contains descriptive data on stolen or allegedly forged items, such as vehicles, fire arms, banknotes, or identity papers. Notably, for all these persons and objects in the SIS II, the system gives concrete instructions as to how to proceed (e.g., observation, arrest, seizure, etc.) in case they are encountered during interactions with state officials, for example at border crossing points or in the context of traffic controls. To do so, the SIS II is set up as a relational database that produces individual records by linking data points that belong to an individual person or object. Importantly, the records created in this way contain particular data points that allow for identification, i.e. that tie a physical person or object to the corresponding administrative record. Such an identifier can, for example, be biographical or biometric data attributed to a person or a serial number or document number pertaining to an object.

The data in the SIS II thus closely correspond with the logics and operational needs of those agencies that are in the EU concerned with Justice and Home Affairs matters, i.e. law enforcement, border control, judicial cooperation, migration management, and asylum. But how do we know whether these data are of high or low quality? The literature usually defines data quality in relative terms by taking on the perspective of "information consumers" (Wang et al., 2002: 4) and asking whether data are "fit for use" (Herzog et al., 2007: 7) vis-à-vis the consumers' needs. Accordingly, a key quality criterion is whether they can be regarded a "trusted source" (McGilvray, 2008: 5) and can be confidently used for a particular task. Notably, not every use-case comes with the same quality requirements. As argued above, in fields such as advertising, misdirected personalized product recommendations might not be cost-effective but also won't cause any actual harm. Untrustworthy data can therefore, to a certain extent, be considered tolerable. Other domains have less error tolerance, or no error tolerance at all. The safety procedures in the operation of a nuclear power plant, for example, are reliant on highly trustworthy data, as decisions made on faulty data could lead to significant hazards. And agencies concerned with EU Justice and Home Affairs need to be able to unanimously identify persons and objects and link them to investigative knowledge and concrete instructions for action based on accurate and complete data in the SIS II and other databases.

Within a relational understanding of data quality, governance, here broadly understood as networked forms of governing and regulation predicated on the interdependence of institution and the continued exchange of resources between them (Rhodes, 1997), arguably presents a unique category of use-cases. The interface between the state, represented through its institutions, and citizens/third-country nationals is predicated on imaginaries of ultimate precision. Public administration and security authorities need to be able to identify individual persons and access available information about them to determine whether they are, for example, eligible for welfare benefits, have paid their taxes, have a valid residence permit, or are subject to an outstanding arrest warrant (Scott, 1998). Low quality data in public records can, in turn, have severe implications for individual lives, as a person could be denied the benefits that they are in fact entitled to, be issued a penalty for allegedly unpaid taxes, or even falsely extradited or arrested and prosecuted. States are thus interested in rendering the data that they use as the basis for decision-making and interventions as reliable as possible. High data

quality, in other words, is imperative in governance contexts.

From an epistemic perspective, data quality is predicated on the assumption that "an information system [is] a representation of a real-world system" (Batini and Scannapieca, 2006: 36). High data quality would in this sense be defined as the most accurate representation of empirical phenomena in a dataset. Low data quality, vice versa, would be defined as a suboptimal form of representation. Notably, the idea of data quality is usually linked to the notion that databases almost by definition do not achieve optimal representation but that they contain factual errors, outdated information, or missing values. There is thus, so the assumption, always room for improvement that can be achieved through a variety of technical and organizational activities and processes such as automated screening for anomalies or the identification and mediation of root causes (e.g., Batini et al., 2009; Lee et al., 2006; Morbey, 2011).

In the example of the SIS II, data quality pertains to the question how close the representations of persons and objects in the system come to the actual persons and objects that they relate to. For biometric identifiers such as fingerprints, this means, for instance, that they must meet the required ISO norms for biometric templates, including sufficient resolution of the original image of a fingerprint and the correct enrollment of the image, such that the produced data can be algorithmically compared and matched against other biometric data (Joint Research Centre, 2015). And for biographical data, it means for example that the birth date of a person is correct, and that their address of residence is up-to-date.

Minimizing the delta between desired and achieved representation is, however, not a straightforward task. Shortcomings in data quality can relate to several different dimensions, depending on the type of data, the database structures within which they are housed, the empirical phenomena that they relate to, and not least available knowledge about the characteristics of the "real-world system" that data are supposed to represent. The following section provides an overview of the most prevalent data quality dimensions.

Data quality dimensions

There is little consistency in the literature regarding the exact number and properties of data quality dimensions. Wang et al. (2002), for example, identify 16 different dimensions that fall into four broader categories (intrinsic quality; accessibility; contextual quality; representational quality). Others offer less fine-grained classification systems for data quality dimensions, such as for instance McGilvray (2008: twelve dimensions), Batini and Scannapieca (2006: eight dimensions), or Herzog et al. (2007: seven dimensions). Notwithstanding the differences in classification and corresponding definitions, there are several dimensions of data quality that cut across the literature. These are (1) accuracy; (2) completeness; (3) the trustworthiness of sources; (4) temporality; and (5) accessibility. They will be discussed in turn in the following.

(1) *Accuracy* concerns the question how well empirical phenomena are captured and represented in a dataset. It has several sub-dimensions. Syntactic accuracy concerns the question whether the data value that represents an empirical characteristic is permitted within the logical structure of the dataset. The representation of age in a dataset would, for example, in most cases require a numerical value whereas text would not be permitted. Semantic accuracy, on the other hand, pertains to the question whether the

data representation captures the empirical phenomenon in the most "correct" way possible. Staying within the example of age, for a person who is in fact 27 years old, data indicating they were 37 years old would be considered a violation of semantic accuracy. Another sub-dimension pertaining to accuracy is whether there are more than one representation instances of the same real-world entity in the same dataset. This can for example happen when a person takes on a second, fraudulent identity in order to disguise their real identity in the context of criminal activities. The same person could thus be represented in the database twice: under their real identity and under their fake identity. Semantic accuracy and duplicate records are closely connected to two other dimensions of data quality: completeness and the trustworthiness of sources.

(2) *Completeness* concerns the question whether empirical phenomena are captured and represented in the dataset in their entirety. In relational database structures, individual records are usually considered complete when a valid data value has been assigned to all available categories. In a public register, a person might for example be represented as a combination of name, address, birth date, and telephone number. A complete record would contain syntactically accurate information for all these categories. Completeness comes, however, with some epistemic caveats that relate to assumptions about the "real-world system" that data are supposed to represent. In case data for one or more fields are missing, there are different possibilities as to why this might be the case. Missing data for "telephone number" could mean that the person has no telephone. It could, however, also mean that the person has a telephone but the number is not known. Moreover, it could mean that the number is not known but it is also not known whether the person has a telephone in the first place. Finally, there could be a possibility that the person has more than one phone number, in which case the data model would not be adequately specified to represent the empirical reality it relates to. Completeness is a particularly pertinent issue in databases that deal with uncertainty about the real-world entities that they relate to, such as for example the SIS II and EIS. In domains such as law enforcement or intelligence, there is often a lack of knowledge about the empirical phenomena that are to be captured as data (e.g., suspects, counterfeit documents, criminal networks). Accordingly, it is difficult to assess whether the data that represent them are complete and where the reasons for potential lack of completeness are located.

(3) *Trustworthiness of sources* relates primarily to data that have been produced outside the use-case within which they are to be used. Trustworthiness is in this sense not first and foremost about data themselves, but about the social relations between their producers and their consumers. With regard to data acquired from external sources, trust can be built via transparent documentation as to how the data were produced, enabling a process-based assessment of their quality. In other contexts, when data are for instance produced on the basis of the accounts of informants, assessments of trustworthiness become more complicated and idiosyncratic. The trustworthiness of sources is, once more, particularly pertinent in the context of security-related and investigative databases, where data are likely to be mediated by the accounts of witnesses or suspects and the social environments and power relations within which they are located. Data not considered to come from trustworthy sources thus require validation procedures, for example by triangulating multiple sources or multiple accounts, or by cross-checking with data in other databases.

(4) *Temporality* pertains to the time-related features of data. There are several sub-dimensions of temporal data quality. Volatility relates to the question how frequently data

vary over time. In databases for governance and public administration, the majority of data can be assumed to be relatively stable. Birth dates, for example, are not supposed to change at all, while names usually change rather seldom (e.g., after a marriage, divorce, or adoption). Other data might change more frequently, but still not very often, such as for example residence status or home address. Still other data can change at a higher rate, for example information in a criminal database during an ongoing police investigation. In some contexts, data might even change in (near) real-time, for example weather data or traffic information produced by sensors. One caveat in regard to volatility is presented by the fact that there is often no fixed frequency with which data change. Especially changes in the social world tend to occur non-systematically, thus presenting epistemic challenges in volatility that are similar to those found in completeness. In simple terms, even for a recently updated data point, there can hardly be any certainty that the characteristic it relates to in the real world has not changed again in the meantime. Currency addresses this challenge partly by expressing how promptly data are updated once change has been detected. Currency is a well-established issue in public administration databases, where records can be outdated in spite of available updates due to insufficient update procedures. Timeliness, finally, expresses how current data are within a specific use-case context and relates to the requirements that have been specified for the task at hand.

(5) *Accessibility* pertains to the question whether data are readily available for the task at hand or whether they can be easily made available. Whereas the data quality dimensions discussed so far refer primarily to the relation between data and the empirical phenomena that they represent, accessibility concerns the data infrastructures that enable or disable access to data, be it in terms of interfaces and data formats, access rights and legal restrictions, or silo structures where databases are not interconnected. Especially with regard to Justice and Home Affairs databases in the EU, accessibility has been deemed a policy priority and is currently addressed through the development of technical components that render separate systems interoperable and facilitate information exchange between them (Bellanova and Glouftsios, 2022; Leese, 2022b).

It is at this point important to note that dimensions of data quality, while representing distinct epistemic, technical, social, and infrastructural questions, are in many cases interdependent. Notably, as Batini and Scannapieco (2006: 40) argue, "if one dimension is considered more important than the others for a specific application, then the choice of favoring it may imply negative consequences for the other ones." Such trade-offs can, for example, be found in the relations between timeliness and accuracy or timeliness and completeness. The time it takes to address issues relating to invalid or missing values in a dataset, in other words, will almost by definition affect update frequencies and the immediate availability of data for the task at hand. Trade-offs involving timeliness are particularly pertinent in the context of databases for knowledge and intelligence, where data can either be quickly available for analysis but untrustworthy or trustworthy but only available at a later point in time when they have undergone quality control processes (Leese, 2022a). Another key relation exists between the dimensions of completeness and syntactic accuracy, i.e. when a dataset might appear to have no missing values, but the existing values might violate rules for permissible values for a given category.

Studying data quality

Data quality, as has become apparent, is a multi-faceted concept that will take on different shapes in relation to the use-cases of data and the quality requirements (including the prioritization of particular quality dimensions) that have been specified for these use-cases. Moreover, data quality is almost always understood as a process that does not stop at the point of diagnosis but implies actions to reduce the distance between achieved and desired forms of representation. The literature accordingly tends to foreground practical ways in which data quality can be assessed and improved. In the words of Sadiq (2013: ix), data quality should be considered as a holistic set of activities that include "organizational aspects, i.e. strategies to establish people, processes, policies, and standards required to manage data quality objectives; architectural aspects, i.e. the technology landscape required to deploy developed processes, standards, and policies; and computational aspects which relate to effective and efficient tools and techniques for data quality." There are important implications from such a perspective as to how data quality should be studied from a social science perspective.

On the organizational level, data quality can be considered a political/policy-making question. Decisions need to be made about the desired level of trustworthiness for a particular use-case, assessment procedures and tools need to be agreed on, and standards need to be defined. Moreover, actors need to be coordinated and resources need to be budgeted. In governance and public administration, as the initially discussed example of data quality in the SIS II and EIS demonstrates, such politics of data quality involve the coordination of multiple levels of data production and consolidation (i.e. data being produced locally/nationally and then aggregated to the EU level) with multiple possible intervention points for data quality activities, requiring decisions as to who should be responsible for which data quality aspects and where and when corresponding measures should be carried out. In the example of EU Justice and Home Affairs databases, it appears as though we are currently witnessing the emergence of a top-down regulation approach that defines quality thresholds that data from the Member States must meet to become part of centralized systems or else be rejected (European Union, 2021a: 21). The EU would in this sense be the standard setter but actual data quality activities would be relegated to the national level. Politics of data quality can presumably be encountered in all domains of governance and public administration. To study them, we need to pay attention to the political discourses and policy-making processes around databases and data exchange.

On the architectural and computational level, data quality might initially appear as a technical or bureaucratic aspect that is in fact in practice usually relegated to the back office activities of IT personnel or data scientists and analysts. While often overlooked, their work has in recent years been foregrounded as an important element in how data "enact" worlds, i.e. how they bring into being the phenomena they relate to in specific ways and render them amenable to interventions (Mol, 2002). Critical Data Studies scholars have in this sense claimed that close analytical attention needs to be paid to the practices – i.e. the "embodied, materially mediated arrays of human activity centrally organised around shared practical understanding" (Schatzki, 2001: 11) – of those who deal with data on an everyday basis. Engaging with their rationales and their routines can provide us with an important complimentary perspective that foregrounds how high-level considerations around data quality are implemented and realized (Ruppert and Scheel, 2021). Moreover, data practices must be understood

as taking place within complex assemblages that involve social as well as technical components, including the likes of infrastructures, budgets, legal frameworks, or economic considerations (Kitchin and Lauriault, 2014). Their study thus requires an approach that can account for the socio-technically mediated forms of human-data interaction that shape and reshape the form and informational value of data. Suitable methods to do so include ethnographic approaches such as (participant) observation or qualitative interviews that capture the experiences and viewpoints of practitioners.

Conclusions

In conclusion, despite its growing importance in governance contexts, data quality so far remains an understudied issue area in the social sciences. One of the reasons for this might be its seemingly technical character. However, as discussed throughout this Working Paper, not only does data quality have implications for the interface between the state and citizens/third-country nationals and corresponding governance capacities. Notably, as has been shown, data quality involves numerous epistemic challenges and choices how to address them. These choices correspond closely with the use-cases of data and the strategic decisions that prioritize particular data quality dimensions and define standards, assessment methods, and procedures.

A research agenda on data quality, so the argument put forward here, must, however, not only analyze the politics of data quality, but must also include the situated practices of those who "do" data quality on an everyday basis. Paying attention to their tacit assumptions, routines, and implicit ways of saying and doing provides an important complementary perspective. In doing so, the study of data quality ties in with an emerging literature at the intersection of Critical Data Studies and Science and Technology Studies that empirically investigates the often surprising and not straightforward ways in which data come to matter in the world.

References

Batini C, Cappiello C, Francalanci C and Maurino A (2009) Methodologies for Data Quality Assessment and Improvement. *ACM Computing Surveys* 41(3): Article 16.

Batini C and Scannapieca M (2006) *Data Quality: Concepts, Methodologies and Techniques*. Berlin / Heidelberg: Springer.

Bellanova R and Glouftsios G (2022) Formatting European Security Integration Through Database Interoperability. *European Security* 31(3): 454-474.

Council of the European Union (2020) 10422/20: Questionnaire Preparing the First European Data Quality Day (EDQD), 9 September. Brussels.

Desrosières A (2002) *The Politics of Large Numbers: A History of Statistical Reasoning*. Harvard: Harvard University Press.

European Union (2021a) Commission Implementing Regulation (EU) 2021/2224 of 16 November 2021 laying down the details of the automated data quality control mechanisms and procedures, the common data quality indicators and the minimum quality standards for storage of data, pursuant to Article 37(4) of Regulation (EU) 2019/818 of

the European Parliament and of the Council. Brussels: Official Journal of the European Union.

European Union (2021b) Commission Implementing Regulation (EU) 2021/2225 of 16 November 2021 laying down the details of the automated data quality control mechanisms and procedures, the common data quality indicators and the minimum quality standards for storage of data, pursuant to Article 37(4) of Regulation (EU) 2019/817 of the European Parliament and of the Council. Brussels: Official Journal of the European Union.

Gitelman L and Jackson V (2013) Introduction. In Gitelman L (ed.) *"Raw Data" is an Oxymoron*. Cambridge/London: MIT Press, 1-14.

Hacking I (1990) *The Taming of Chance*. Cambridge: Cambridge University Press.

Herzog T N, Scheuren F J and Winkler W E (2007) *Data Quality and Record Linkage Techniques*. New York: Springer.

Joint Research Centre (2015) JRC Science for Policy Report: Fingerprint Identification Technology for its Implementation in the Schengen Information System II (SIS-II). Brussels: European Commission.

Kitchin R and Lauriault T P (2014) Towards Critical Data Studies: Charting and Unpacking Data Assemblages and Their Work. The Programmable City Working Paper 2.

Lee Y W, Pipino L L, Funk J D and Wang R Y (2006) *Journey to Data Quality*. Cambridge / London: MIT Press.

Leese M (2022a) Enacting Criminal Futures: Data Practices and Crime Prevention. *Policing and Society* online first: 10.1080/10439463.2022.2112192.

Leese M (2022b) Fixing State Vision: Interoperability, Biometrics, and Identity Management in the EU. *Geopolitics* 27(1): 113-133.

McGilvray D (2008) *Executing Data Quality Projects: Ten Steps to Quality Data and Trusted Information*. San Francisco: Morgan Kaufmann Publishers.

Mol A (2002) *The Body Multiple: Ontology in Medical Practice*. Durham: Duke University Press.

Morbey G (2011) *Data Quality for Decision Makers: A Dialog Between a Board Member and a DQ Expert*. Wiesbaden: Springer Gabler.

Rhodes R A W (1997) *Understanding Governance: Policy Networks, Governance, Reflexivity and Accountability*. Buckingham / Philadelphia: Open University Press.

Ruppert E and Scheel S (eds.) 2021. Data Practices: Making Up a European People, London / New York: Goldsmiths Press/MIT Press.

Sadiq S (ed.) 2013. *Handbook of Data Quality: Research and Practice,* Heidelberg / New York / Dordrecht / London: Springer.

Schatzki T R (2001) Introduction: Practice Theory. In Schatzki T R, Knorr Cetina K & von Savigny E (eds.) *The Practice Turn in Contemporary Theory*. London / New York: Routledge, 10-23.

Scott J C (1998) *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven/London: Yale University Press.

Wang R Y, Ziad M and Lee Y W (2002) *Data Quality*. New York / Boston / Dordrecht / London / Moscow: Kluwer Academic Publishers.

Suggested citation: Leese M (2022) Data quality in governance: A definition and a research agenda. CURATE Working Paper No. 1, November 2022. doi:10.3929/ethz-b-000581434.