Access Control and Quality Attributes of Open Data: Applications and Techniques

Erisa Karafili[1], Konstantina Spanaki[2], and Emil C. Lupu[1]

[1] Department of Computing, Imperial College London
e.karafili, e.c.lupu @imperial.ac.uk
[2] School of Business and Economics, Loughborough University
k.spanaki@lboro.ac.uk

Abstract. Open Datasets provide one of the most popular ways to acquire insight and information about individuals, organizations and multiple streams of knowledge. Exploring Open Datasets by applying comprehensive and rigorous techniques for data processing can provide the ground for innovation and value for everyone if the data are handled in a legal and controlled way. In our study, we propose an argumentation and abductive reasoning approach for data processing which is based on the data quality background. Explicitly, we draw on the literature of data management and quality for the attributes of the data, and we extend this background through the development of our techniques. Our aim is to provide herein a brief overview of the data quality aspects, as well as indicative applications and examples of our approach. Our overall objective is to bring serious intent and propose a structured way for access control and processing of open data with a focus on the data quality aspects.

Keywords: Data quality · Argumentation Reasoning · Open Data · Data Access.

1 Introduction

Open Data as a term is often questioned in literature and public debates; however the definition of open data addresses two core concepts: the openness and the features of the data (quality aspects) required to be characterized as 'open data'. Open Data are more than public data, they include also private sector data, and therefore their use can be controversial regards the ownership of the data. In our paper we accept the definition of Lindman et al. ([23] p. 740) for the open data as *"data, which is legally accessible through the Internet in a machine-readable format"*. The roots of open data stem from the context of open source [10] in combination with the background of open innovation [6, 22] and open access [37]. The main difference of open source and open data as it is highlighted by Lindman et al. [23, 24] is the fact that *"data is used for storage and, the application is used for different operations based on data"* ([24] p. 1240) where the application is the open source.

We also include here similar assumptions for our examples as in our previous work [20] that the data are open for use/reuse/redistribution and in a form that can be accessible to everyone who has the rights to access them. Data should be accessible, assessable, and reliable and in a state that judgments and sense-making can be made so as value is created [14, 15]. Data should give the opportunity for the audience to make intelligent judgments or assessments once they are scrutinized. Another important aspect of the open data is around their format, so as the data can be used or reused for different purposes, including information and meta-data or linked data. Opportunities and innovation around the concept of open data can bridge the social divide and equality issues by their use from all the social levels [14]. In organizational level large numbers of enterprises have tried to open their data and they have generated new services and applications for individuals. The entrepreneurial movement towards open data shared the view for transparency and accountability of organizational practices and also the collaborative potential around the data with other organizations and individuals for innovative purposes [23, 24] .

However, some major challenges for the supporters of open data come mostly in the form of privacy issues related to personal data (identities, privacy and regulatory landscape). The privacy issues are more prevalent lately in line with the new GDPR regulations. Some other challenges can be considered as the costs of collecting, producing and releasing the datasets, as well as the diversity of interests and behaviour in opening the data for the public use or even the actual use of the data.

The proposed approach extends the concept of open data with a focus on the quality aspects and processing techniques and the associated data sharing agreements (DSAs). The study proposes an expressive policy analysis language for representing the DSAs enriched with data quality attributes to capture specified aspects of the data. The introduced analysis, based on argumentation and abductive approach, permits the construction of correct and efficient DSAs that can be applied in different contexts during the processing of data.

We introduce briefly the related work for data quality and data access and usage control in Section 2. In Section 3 we present the used techniques for representing the data quality properties and permitting the correct access of open data. We show an application of our methodology in Section 4, where we work with immigration movement open data. Finally, we conclude in Section 5 and discuss some future research directions.

2 Related Work

2.1 Data quality

Data quality as a concept was initially presented through the data manufacturing analogy. Data as a "raw material" was initially introduced by Brodie [5] through the analogy between product manufacturing and data manufacturing process when data quality was a major concern in transforming data to valid information and knowledge [1, 11, 28]. Studies as those of Fox et al. [11] and Wang

and colleagues [31, 35, 36] focused on the analysis of 'data quality', in terms of dimensions, attributes, as well as the upcoming issues and research areas. The so called 'dimensions' for data quality representation as they were presented by Fox et al. [11]: accuracy, freshness, completeness, and consistency. In our study, we will base our proposed approach on the dimensions and quality aspects as they were described in the literature of data quality.

Some of the most indicative studies around the quality aspects have developed the concept of data manufacturing analogy in order to find out the path for better data quality [13, 25, 27, 34] and they designed frameworks that describe and track data manufacturing processes [2, 29, 34, 35]. A simple framework of input-process-output describing the similarities between the two manufacturing processes is proposed by Wang et al. [35] and calls for continuously defining, measuring, analysing, and improving data quality. Mostly, the data manufacturing analogy was focusing on data quality and the ways to ensure that we can trust the data we use in manufacturing processes. Recent studies in data quality research, apply the data manufacturing analogy, in order to explain the tailoring techniques and the potential of data marketplaces within the context of supply chain [12, 30].

2.2 Data access and usage control

Due to the increasing connectivity between users, there is a parallel increase of the associated security breaches and attacks. Protecting and securing the environment where the data is transferred/stored/used or even re-used [19] remains a major challenge for all interested parties. Data-centric security solutions have dominant position in the literature [3, 26, 33] and specifically the protection of the data transfers and transactions. Data-centric security solutions present two main challenges associated with the access and the usage control of the data. Both of them, have been widely studied and the research has developed multiple solutions for solving such problems [9, 21].

Before creating, sharing and using the data, the data subject, controller and processor should agree regarding the different rules that describe how the data should be treated, called the *data sharing agreements*, denoted by DSAs [32]. The DSAs describe not only the agreements between the data subject, controller, and processor, but also the compliance of the different business and regulatory contexts for data sharing. Thus, the DSAs require an expressive language to represent the agreements. Given the heterogeneous nature of the agreements, various conflicts can be generated, especially between legal and business rules, or legal rules and user requirements. The authors in previous studies [17, 20] propose a policy language that represents complex agreements, and an analysis process for capturing the conflicts and solving them. The approach is based on abductive [16] and argumentation based reasoning [4, 8], as this technique can facilitate decision making mechanisms under conflicting knowledge.

3 Methodology

The methodology presented in this study assumes that data are processed by different entities. Data Sharing Agreements (DSAs) are established between the entities for data processing and are composed of various constraints and rules. As in previous work [20], an expressive policy analysis language is used for representing the DSAs. Data quality is the main focus of the data processing mechanism; therefore the policy language is enriched to capture various data quality properties like accessibility, timeliness and accuracy. The used policy language permits the analysis of the various policies and the detection of the rising conflicts, redundancies or the missing cases. We follow argumentation and abductive reasoning to build our proposed methodology to capture and solve conflicts between context dependent rules. The introduced analysis permits the construction of precise DSAs that can be applied in various contexts during data processing phase.

3.1 A policy language for DSAs and data quality

To represent the various data sharing agreements that incorporate extensive types of information, we use a policy analysis language [17]. The used policy language is an extension of the one introduced by Craven et al. [7], where the used extension supports efficiently the data access of open data and the data quality properties. Let us introduce briefly the policy language, which represents the requirements of accessing, using and sharing the data.

The policy language is composed of rules that are predicates and domain descriptions, and represent the authorization and obligation rules. The first three predicates are authorization rules and have in their structure a specific subject, as well as specified targets, and actions, while the last one is a domain description predicate.

$$permitted(Sub, Tar, Act, T)$$
$$denied(Sub, Tar, Act, T)$$
$$obl(Sub, Tar, Act, T_s, T_e, T)$$
$$holdsAt(Predicate, T)$$

The above predicate represents correspondingly: the permission *permit* for a given subject Sub to perform an action Act to a target object Tar at the instant of time T; the prohibition *denied* for a given subject Sub to perform an action Act to a target object Tar at the instant of time T; the obligation *obl* for a given subject Sub to perform an action Act during the period of time from T_s to T_e; the domain description predicate $holdsAt$ means that a given property/predicate $Predicate$ is true in a given instance of time T.

The used policy language can represent the permission, denial and obligation concepts for the DSAs, e.g., the owner of the data can access to her/his own data.

$$permitted(Sub, Data, access, T) \leftarrow holdsAt(owner(Sub, Data), T).$$

In the above formula, the preconditions are on the right side while the conclusion is on the left side of the arrow. Thus, if it is true (represented by using

the *holdsAt* predicate) that subject Sub is the owner of the data, described as $owner(Sub, Data)$, at the instant of time T, then the subject is permitted to access the $Data$ at the instant of time T.

Following the above example, let us introduce how a prohibition predicate can be constructed, e.g., if the subject is not the data owner then s/he cannot access the data,

$$denied(Sub, Data, access, T) \leftarrow \textbf{not}\ holdsAt(owner(Sub, Data), T).$$

An obligation force the data user to perform certain actions to the data. The enforcement is made possible by using the sticky policy mechanism, where the rules are attached to the data and enforce to them various constraints, e.g., temporal, geographical etc. Continuing with the example, an obligation can use temporal constraints where it asks for a particular data to be deleted after a certain amount of time from when the data was accessed.

$$obl(Sub, Data, delete, T, T', T) \leftarrow holdsAt(access(Sub, Data), T), \\ T' = T + 6.$$

In the above case the subject has the obligation to delete the data from the moment s/he accessed them T, until after 6 years from that instant of time, and the obligation is enforced from the moment s/he accessed the data T.

Representation of the quality and sharing aspects

The sharing and usage of data raises issues around the description of other properties related to the quality of the collected data. The data quality is an important factor when we are working with data consumers[3], where data quality is defined as data that fit the data consumers' requirements. The used policy language permits to represent various *data quality* properties.

An important data quality property is *accessibility*. Our methodology ensures that data accessibility respects the imposed constraints e.g., security, legal and business constraints. The *permitted*, *denied* and *obl* regulation rules enforce a correct data accessibility, by permitting the allowed users to access the data, prohibiting the users that do not satisfy the needed requirements for accessing the data and putting obligations on the users about the use/access and sharing of the rules.

When collecting data with the purpose on releasing them as open data, not all part of the data can become public. Our methodology permits to classify the data depending on their level of privacy. An example would be the data collected from individuals for statistical purpose. When the data are being public, their private information are not released to the public.

$$denied(Sub, Data, access, T) \leftarrow holdsAt(private(Data), T), \\ holdsAt(member(Sub, public), T).$$

$$permitted(Sub, Data, access, T) \leftarrow holdsAt(private(Data), T), \\ holdsAt(member(Sub, staff), T).$$

[3] The data consumers are called the entities that use/share/access the data.

The above predicates state that if the subject is a member of the public, then s/he cannot access the private data, while if the subject is a member of the staff, thus working with the data, then s/he is permitted to access the data.

Accuracy is another data quality attribute that our methodology is able to represent. When data are collected, e.g., by an human actor or IoT devices, an obligation for satisfying a particular accuracy level when collecting and storing the data is enforced.

$$accuracy(Data, T, level) \leftarrow holdsAt(collect(Device, Data), T), \\ holdsAt(capacity(Device, level), T).$$

In the above predicate, a certain level of accuracy is ensured, when the device that collects the data, use the capacity of that level of accuracy. The accuracy can also be restricted/manipulated in order to give different level of accuracy to different data consumers, or depending on the type of data. An example would be the visual data collected by drones in certain area. The image quality (the accuracy of the images) can be restricted when is released publicly, in case of sensitive data.

The notion of data *freshness* is part of the timeliness as a data quality aspect. Data freshness is the degree data represent reality in the required point in time. In our methodology it is represented as a predicate that expresses that the data represented by Tar are fresh at the instant of time T: $freshness(Tar, T)$.
An example of freshness is that data is collected in the last 10 minutes.

$$freshness(Data, T) \leftarrow holdsAt(collect(Device, Data), T'), \\ T \leq T' + 10.$$

3.2 Analysis and conflict resolution

Given the heterogeneity of the rules that compose the DSAs is natural to have conflicts between rules. There exists a conflict between the DSAs rules when when an action is both permitted or denied on the same instant of time.

$$permitted(Sub, Tar, Act, T) \qquad denied(Sub, Tar, Act, T)$$

Another type of conflict is when an action is denied and obliged to occur at the same instant of time.

$$obl(Sub, Tar, Act, T_s, T_e, T) \qquad denied(Sub, Tar, Act, T) \qquad T_s < T$$

The conflicts exists not only between exactly matching entities, but also when a certain entity is subset of another one. Going back to the previous example of public data, suppose Bob is a member of the public $member(Bob, public)$ and **not** $member(Bob, staff)$ therefor, he should not access the private data, because of the following rule.

$$r_1 : denied(Sub, Data, access, T) \leftarrow holdsAt(private(Data), T), \\ holdsAt(member(Sub, public), T).$$

As *Bob* is the owner of the data $owner(Bob, Data)$, he should have access to his own data, even the private ones, because of the following rule

$$r_2 : permitted(Sub, Data, access, T) \leftarrow holdsAt(owner(Sub, Data), T).$$

The two above rules are in conflict between each other.

To capture the above conflicts and conflicts similar to the ones shown previously, we introduce an analysis process to the DSAs [17], based on the abductive reasoning, that identifies conflictual policy regulation rules. The analysis is able to identify gaps between rules as well as redundancies. Once the conflicts are identified, we use a conflict resolution based on the argumentation reasoning, that solve the conflicts by introducing priorities between them. In specific, for the previous example of *Bob* permitted and denied to access the private data, rule r_1 and r_2 are in conflict, then we decide that r_2 is stronger than r_1, denoted by $r_2 > r_1$ for the case when the data subject is the owner of the data.

4 Use Case: Immigration Movement Open Data

In this section we show our proposed methodology applied in a realistic case scenario taken from the *immigration movement open data*. The open immigration data are gathered from governmental and humanitarian organisations from refugees camps, as well as the immigrants while trying to proceed with their travel document applications. The collected data enclose various and multiple properties, e.g., age, country of origin, type of immigration (e.g., political, economic), education level, legal/illegal immigration. The properties of the data, as well as their accuracy and timeliness, compose aspects around the quality of the data. Revealing immigration data to the general public by humanitarian, statistical, and governmental entities can be important. On the other hand, it is crucial to privatize and sanitize the data, before they are made public. Exposing to the public all these types of data (privatized and sanitized) can sometimes be dangerous or beyond the human rights, e.g., revealing sensitive information about the refugees camps in unstable or conflicting geopolitical areas. Hence for publicizing immigration movement data, we can divide the data in three categories, that represent also their quality: *basic*, *medium*, and *detailed* data.

The *basic* data are open data of the immigration movement. Such data are published every year, and support the categorization of immigrants in three age groupings (i.e., children, adults, and elder people), the sum of these groupings give the total number of immigrants per country.

The *medium* data can be accessed only by national statistics agencies. Such data can be updated monthly, and have the exact age of the immigrants, the type of immigration (e.g., political, economic), the country of origin, education level, legal/illegal immigration.

The *detailed* data can be accessed only by governmental and UN entities. Such data have the same properties as the medium one but are updated weekly, instead of monthly.

We are able to deal with this division of data quality, by using the argumentation reasoning approach. Before introducing the rules that describe how the access to data is made by agents with different roles, we define the *freshness* of data for this use case as below.

$$\begin{aligned} Fresh(Data, T) \leftarrow\ & holdsAt(update(Data), T_i), \\ & holdsAt(update(Data), T_j), \\ & \textbf{not}\ holdsAt(update(Data), T_k), \\ & T_j < T_k < T_i,\ T_i - T_j \leq 7, \\ & T \geq T_i,\ T - T_i \leq 7. \end{aligned}$$

The above predicate, states that $Data$ is $Fresh$ at the instant of time T, if it was updated at the instant of time T_i, (where T is bigger than T_i of maximum 7 days), the previous time when the data was updated is T_j (where T_j is at most 7 days before T_i) and there was no update made between T_i and T_j.

The data quality is divided into three categories, depending on the three types of public where these categories are released. Given an $Agent$ that wants to access the data, it can be a general entity/individual denoted by $Public(Agent)$ (in this case the data is open access), a statistical entity $Statistic(Agent)$, or an individual/entity part of the UN or Governmental Entities, denoted by $UN/Gov(Agent)$. Depending on the role of the agent a freshness restriction is made $Cast\_Fresh$, as described below.

$$\begin{aligned} Cast\_Fresh(Agent, Data\_In, Data\_Out) \leftarrow\ & Public(Agent), \\ & Fresh(Data\_In, T), \\ & holdsAt(update(Data\_Out), T'), \\ & T - T' \leq 365, T' = X. \end{aligned}$$

$$\begin{aligned} Cast\_Fresh(Agent, Data\_In, Data\_Out) \leftarrow\ & Statistic(Agent), \\ & Fresh(Data\_In, T), \\ & holdsAt(update(Data\_Out), T'), \\ & T - T' \leq 30, T' = Y. \end{aligned}$$

The predicate $Cast\_Fresh$ depends on the type of $Agent$. Thus, given the fresh data $Data\_In$, it gives to the $Public$ the data collected at most 1 year ago $Data\_Out$, and to the $Statistic$ institute the data collected at most 1 month ago, where X and Y are fixed time range correspondingly representing when the data are released every year, e.g., X=[01/01-07/01], while Y when the data are released every month, e.g., $Y = [01 - 07]$ of every month.

We make use of another predicate, $Cast$, that removes some of the immigration data properties. Thus given $Cast(Data\_In, Data\_Out)$ where $Data\_In$ is the fresh data that has different properties, the result of this restriction/alteration is $Data\_Out$ that does not have any more the following properties: immigration type, education level, country of origin, legal/illegal immigration, and the age property is aggregated into three categories, i.e., children, adult, elder. For constructing the $Cast$ predicate, we use a similar mechanism as the one introduced in [18].

We can now represent the rules that describe who can access the data. In case the original data, $Data'$, are not sanitized/privatized ($Anonym(Data', Data)$), then nobody can access the data, as described in rule (1). Rule (2) describes that agents that are from the UN or/and governments can access the data unaltered and the data should be fresh and sanitized. In case the agent is not part of UN/Gov then it cannot access the data, even though the data is sanitized, as described in rule (3).

$$denied(Agent, Data, access, T) \leftarrow \textbf{not } Anonym(Data', Data) \tag{1}$$

$$\begin{aligned} permitted(Agent, Data, access, T) \leftarrow\ & UN/Gov(Agent),\ Fresh(Data', T), \\ & Anonym(Data', Data) \end{aligned} \tag{2}$$

$$\begin{aligned} denied(Agent, Data, access, T) \leftarrow\ & \textbf{not } UN/Gov(Agent), \\ & Anonym(Data', Data) \end{aligned} \tag{3}$$

$$\begin{aligned} permitted(Agent, Data, access, T) \leftarrow\ & Statistic(Agent),\ Fresh(Data', T), \\ & Cast\_Fresh(Agent, Data', Data''), \\ & Anonym(Data'', Data) \end{aligned} \tag{4}$$

$$\begin{aligned} permitted(Agent, Data, access, T) \leftarrow\ & Public(Agent),\ Fresh(Data', T), \\ & Cast\_Fresh(Agent, Data', Data''), \\ & Cast(Agent, Data'', Data'''), \\ & Anonym(Data''', Data) \end{aligned} \tag{5}$$

Rule (4) represents that *Statistic* institutes can access the data with a minor freshness restriction, maximum 30 days. Rule (5) represents that the generic *Public* can access the data with a freshness restriction of maximum 1 year, and the data quality is restricted. In both rules the data is sanitized.

While constructing the rules using our argumentation framework the various conflicts are detected. Rule (3) is in conflict with rules (4) and (5). In this case, as they are special cases of rule (3), we give priority to rule (4) and (5) over rule (3), denoted by (4) > (3) and (5) > (3).

5 Conclusion and Future Work

In our paper, we explained in brief our proposed approach for data processing and access control. Theoretically the approach is structured using the data quality background applied for the context of open data and open knowledge databases. Through our case scenario we present our method and example applications. However, through our presentation we identified additional issues pertinent to

data quality which were not discussed in this study and should be further investigated. In this study we did not focus or discuss a specific conceptualization for data quality in open data context, although it is of great importance and one of our future goals. Another important topic that should be discussed in future research is the measurement of data quality and also a particular focus on multi-source and multi-format data and the supply chains around them. Further directions also could include topics as how the open data could create value for individuals, governments and organizations, in terms of financial growth, well-being, innovation, ethics and also sustainability.

Acknowledgments

Erisa Karafili was supported by the European Union's H2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 746667.

References

1. S. E. Arnold. Information manufacturing: the road to database quality. *Database*, 15(5):32–39, 1992.
2. D. Ballou, R. Wang, H. Pazer, and G. K. Tayi. Modeling information manufacturing systems to determine information product quality. *Management Science*, 44(4):462–484, 1998.
3. J. Bayuk. Data-centric security. *Computer Fraud & Security*, 2009(3):7–11, 2009.
4. A. Bondarenko, P. M. Dung, R. A. Kowalski, and F. Toni. An abstract, argumentation-theoretic approach to default reasoning. *Artif. Intell.*, 93:63–101, 1997.
5. M. L. Brodie. Data quality in information systems. *Information & Management*, 3(6):245–258, 1980.
6. H. Chesbrough. The era of open innovation. *MIT Sloan Management Review*, 44(3):35–41, 2003.
7. R. Craven, J. Lobo, J. Ma, A. Russo, E. C. Lupu, and A. K. Bandara. Expressive policy analysis with enhanced system dynamicity. In *Proceedings of the 2009 ACM Symposium on Information, Computer and Communications Security, ASIACCS*, pages 239–250, 2009.
8. P. M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.*, 77(2):321–358, 1995.
9. D. F. Ferraiolo and D. R. Kuhn. Role-based access controls. In *15th National Computer Security Conference*, 1992.
10. B. Fitzgerald. The transformation of open source software. *MIS Q. Manag. Inf. Syst.*, 30:587–598, 2006.
11. C. Fox, A. Levitin, and T. Redman. The notion of data and its quality dimensions. *Information processing & management*, 30(1):9–19, 1994.
12. B. T. Hazen, C. A. Boone, J. D. Ezell, and L. A. Jones-Farmer. Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics*, 154:72–80, 2014.

13. Y. Huh, F. Keller, T. Redman, and A. Watkins. Data quality. *Information and Software Technology*, 32(8):559–565, 1990.
14. M. Janssen, Y. Charalabidis, and A. Zuiderwijk. Benefits, adoption barriers and myths of open data and open government. *Inf. Syst. Manag*, 29:258–268, 2012.
15. M. Janssen and J. van den Hoven. Big and open linked data (BOLD) in government: A challenge to transparency and privacy? *Gov. Inf. Q.*, 32:363–368, 2015.
16. A. C. Kakas, R. A. Kowalski, and F. Toni. Abductive logic programming. *J. Log. Comput.*, 2(6):719–770, 1992.
17. E. Karafili and E. C. Lupu. Enabling data sharing in contextual environments: Policy representation and analysis. In *Proceedings of the 22nd ACM on Symposium on Access Control Models and Technologies, SACMAT 2017, Indianapolis, IN, USA, June 21-23, 2017*, pages 231–238, 2017.
18. E. Karafili, E. C. Lupu, A. Cullen, B. Williams, S. Arunkumar, and S. B. Calo. Improving data sharing in data rich environments. In *2017 IEEE International Conference on Big Data, BigData 2017, Boston, MA, USA, December 11-14, 2017*, pages 2998–3005, 2017.
19. E. Karafili, H. R. Nielson, and F. Nielson. How to trust the re-use of data. In *Security and Trust Management - 11th International Workshop, STM*, pages 72–88. Springer, 2015.
20. E. Karafili, K. Spanaki, and E. C. Lupu. An argumentation reasoning approach for data processing. *Computers in Industry*, 94:52–61, 2018.
21. A. Lazouski, F. Martinelli, and P. Mori. A prototype for enforcing usage control policies based on XACML. In *Trust, Privacy and Security in Digital Business - 9th International Conference, TrustBus. Proceedings*, pages 79–92, 2012.
22. U. Lichtenthaler. Open innovation: Past research, current debates, and future directions. *Acad. Manag. Perspect.*, 25:75–93, 2011.
23. J. Lindman, T. Kinnari, and M. Rossi. Industrial open data: Case studies of early open data entrepreneurs. In *47th Hawaii International Conference on System Sciences*, pages 739–748, 2014.
24. J. Lindman, M. Rossi, and V. Tuunainen. Open data services: Research agenda. In *46th Hawaii International Conference on System Sciences*, page 1239?1246, 2013.
25. S. T. March and A. R. Hevner. Integrated decision support systems: A data warehousing perspective. *Decision Support Systems*, 43(3):1031–1043, 2007.
26. M. C. Mont and S. Pearson. Sticky policies: An approach for managing privacy across multiple parties. *Computer*, 44:60–68, 2011.
27. T. C. Redman. The impact of poor data quality on the typical enterprise. *Communications of the ACM*, 41(2):79–82, 1998.
28. T. C. Redman and A. Blanton. *Data quality for the information age*. Artech House, Inc., 1997.
29. B. Ronen and I. Spiegler. Information as inventory: a new conceptual view. *Information & management*, 21(4):239–247, 1991.
30. K. Spanaki, Z. Gürgüç, R. Adams, and C. Mulligan. Data supply chain (dsc): research synthesis and future directions. *International Journal of Production Research*, pages 1–20, 2017.
31. D. M. Strong, Y. W. Lee, and R. Y. Wang. Data quality in context. *Communications of the ACM*, 40(5):103–110, 1997.
32. V. Swarup, L. Seligman, and A. Rosenthal. Specifying data sharing agreements. In *7th IEEE International Workshop on Policies for Distributed Systems and Networks (POLICY)*, pages 157–162, 2006.

33. C. Wang, Q. Wang, K. Ren, and W. Lou. Privacy-preserving public auditing for data storage security in cloud computing. In *INFOCOM, 2010 Proceedings IEEE*, pages 1–9, 2010.
34. R. Y. Wang. A product perspective on total data quality management. *Communications of the ACM*, 41(2):58–65, 1998.
35. R. Y. Wang, M. P. Reddy, and H. B. Kon. Toward quality data: An attribute-based approach. *Decision Support Systems*, 13(3):349–372, 1995.
36. R. Y. Wang and D. M. Strong. Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4):5–33, 1996.
37. J. Willinsky. *The Access Principle: The Case for Open Access to Research and Scholarship*. The MIT Press, 2006.