Ethical Dimensions for Data Quality

DONATELLA FIRMANI, Roma Tre University
LETIZIA TANCA, Politecnico di Milano
RICCARDO TORLONE, Roma Tre University

CCS Concepts: • **Information systems** → *Information integration*; • **Social and professional topics** → *Codes of ethics*.

Additional Key Words and Phrases: Data Integration, Source Selection, Knowledge Extraction

ACM Reference Format:
Donatella Firmani, Letizia Tanca, and Riccardo Torlone. 2019. Ethical Dimensions for Data Quality. *ACM J. Data Inform. Quality* 1, 1, Article 1 (January 2019), 5 pages. https://doi.org/10.1145/3362121

1 INTRODUCTION

Data quality is a typical ethical requirement: we could never trust a piece of information if it did not have the typical data quality properties. Yet, we can also assert the opposite: that data should conform to a high ethical standard, for it to be considered of good quality. Hence, the satisfaction of the ethical requirements is actually necessary to assert the quality of a dataset, and in this paper we propose to introduce the most common ethical requirements as dimensions of quality, grouped in an *Ethics Cluster*. By now, we are more than aware that the Internet, and the worldwide extent of the usage of IT and computers, have generated a plethora of datasets in all kinds of application areas; this data can correspond to useful information only if it is of good quality, and let us emphasize that it can be profitable to society *only if its usage conforms to ethical principles*. With a somehow more constructive and dynamic viewpoint, in this paper we discuss the dimensions of ethics in connection with the various phases of what we call the *information extraction process* [20], that is, the process of (i) identifying the data sources containing the information of interest, (ii) collecting the corresponding data and integrating them in order to produce a unique dataset, and (iii) applying the appropriate information extraction methods (from the application of a simple query up to a complex statistical, machine learning or data mining analysis). We thus advocate the need to extend the well-established data quality framework in [5] to incorporate ethics explicitly.

2 THE ETHICS CLUSTER

We present the ethics dimensions, based on the main ethical principles in the literature.
Fairness of data is defined as the *lack of bias* [19]. Recently its importance has been acknowledged, due, for instance, to the unsettling consequences of training systems with biased data [11].

Authors' addresses: Donatella Firmani, Roma Tre University, Via della Vasca Navale, 79, Rome, Italy, 00146, donatella.firmani@uniroma3.it; Letizia Tanca, Politecnico di Milano, Via Ponzio 34/5, Milan, Italy, 20133, letizia.tanca@polimi.it; Riccardo Torlone, Roma Tre University, Via della Vasca Navale, 79, Rome, Italy, 00146, torlone@dia.uniroma3.it.

1936-1955/2019/1-ART1 $15.00
https://doi.org/10.1145/3362121

Transparency is the ability to interpret the information extraction process in order to verify which aspects of the data determine its results. In this context, transparency metrics can use the notions of (i) *data provenance* [19, 18], by measuring the amount of meta-data describing where the original data come from; (ii) *explanation* [15], by describing how a result has been obtained.
Diversity is the degree to which different kinds of objects are represented in a dataset. Several metrics are proposed in [9]. Ensuring diversity at the beginning of the information extraction process may be useful for enforcing fairness at the end. The diversity dimension may conflict with established dimensions in the *Trust* cluster of [5], that prioritizes few high-reputation sources.
Data Protection concerns the ways to secure data, algorithms and models against unauthorized access. Defining measures can be an elusive goal since, on the one hand, anonymized datasets that are secure in isolation can reveal sensible information when combined [1], and on the other hand, robust techniques such as ϵ-differential privacy [10] can only describe the privacy impact of specific queries. Data protection is related to the well-established *security* dimension of [5].

3 ETHICAL CHALLENGES IN THE INFORMATION EXTRACTION PROCESS

We highlight some challenges of complying with the dimensions of the Ethics Cluster, throughout the three phases of the information extraction process mentioned in the Introduction.

A. Source Selection. Data can typically come from multiple sources, and it is most desirable that *each of these* complies with the ethics dimensions described in the previous section. If sources do not comply with (some) dimension individually, we should consider that the really important requirement is that *the data that are finally used for analysis or recommendations do.* It is thus appropriate to consider ethics for multiple sources in combination, so that the bias towards a certain category in a single source can be eliminated by another source with opposite bias. While for the fairness, transparency and diversity dimensions this is clearly possible, for the privacy we can only act on the single data sources because adding more information can only lower the protection level, or, at most, leave it as it is. Ethics in source selection is tightly related to the *transparency* of the source, specifically for sources that are themselves aggregators. Information lineage is of paramount importance in this case and can be accomplished with the popular notion of provenance [18]; however, how to capture the most fine-grained type of provenance, namely *data provenance*, remains an open question [12]. A more general challenge is *source meta-data* extraction, especially for interpreting unstructured contents and thus their ethical implications. Finally, we note that also the data acquisition process plays a role, and developing inherently transparent and fair collection and extraction methods is an almost unstudied topic.

B. Data Integration. Ensuring ethics in the selection step is not enough: even if the collected data satisfy the ethical requirements, not necessarily their integration does [1]. Data integration usually involves three main steps: (i) schema matching, i.e. the alignment of the schemata of the data sources (when present), (ii) identification of the items stored in different data sources that refer to the same entity (also called record linkage or entity resolution), and (iii) construction of an integrated database over the data sources, obtained by merging their contents (also called data fusion). Each step is prone to different ethical concerns, as discussed below.

Schema Matching. Groups treated *fairly* in the sources can become over- or under-represented as a consequence of the integration process, possibly causing, in the following steps, unfair decisions. Similar issues arise in connection with *diversity*.

Entity Resolution. Integrating sources that, in isolation, protect identity (e.g. via anonymization) might generate a dataset that violates privacy: an instance of this is the so-called linkage attack [1]. We refer the reader to [21] for a survey of techniques and challenges of privacy-preserving entity resolution in the context of Big Data.

Data Fusion. Data disclosure, i.e., violation of *data protection*, can happen also in the fusion step if privacy-preserving noise is accidentally removed by merging the data. Fusion can also affect *fairness*, when combining data coming from different sources leads to the exclusion of some groups.

In all the above steps *transparency* is fundamental: we can check the fulfilment of the ethical dimensions only if we can (i) provide explanations of the intermediate results (ii) describe the provenance of the final data. Unfortunately, this can conflict with *data protection* since removing identity information can cause lack of transparency, which ultimately may lead to unfair outcomes.

As source selection, also the integration process – especially the last two steps, where schema information is not present – can benefit from the existence of meta-data, allowing to infer contextual meanings for individual terms and phrases. Fair extraction of meta-data is an exciting topic, as stereotypes and prejudices can be often found into automatically derived word semantics.

C. Knowledge Extraction. An information extraction process presents the user with data organized as to satisfy their information needs. Here we highlight some ethical challenges for a sample of the many possible information extractions operations.

Search and Query. These are typical data selection tasks. Diversifying the results of information retrieval and recommendation systems has traditionally been used to minimize dissatisfaction of the average user [4]. However, since these search algorithms are employed also in critical tasks such as job candidate selection or for university admissions, *diversity* has also become a way to ensure the fairness of the selection process [9]. Interestingly, if integrated data are unfair and over-represent a certain category, diversity can lead to data exclusion of the same category.

Aggregation. Many typical decision-support queries, such as GROUP BY queries, might yield biased result, e.g. trends appearing in different groups of data can disappear or even be reversed when these groups are combined, leading to incorrect insights. The work of [17] provides a framework for incorporating *fairness* in aggregated data based on independence tests, for specific aggregations. A future work is to detect bias in combined data with full-fledged query systems.

Analytics. Data are typically analyzed by means of statistical, data mining and machine learning techniques, providing encouraging results in decision making, even in data management problems [14]. However, while we are able to understand statistics and data mining models, when using techniques such as deep learning we are still far from fully grasping how a model produces its output. Therefore, explaining systems has become an important new research area [16], related to the fairness and transparency of the training data as well as of the learning process.

4 RESEARCH DIRECTIONS

In the spirit of the responsible data science initiatives towards a full-fledged data quality perspective on ethics (see, for instance, redasci.org and dataresponsibly.github.io), the key ingredient is *shared responsibility*. Like for any other engineering product, responsibility for data usage is shared by a contractor and a producer: only if the latter is able to provide a quality certification for the various ethical dimensions, the former can share the responsibility for improper usage. Similarly, producers should be aware of their responsibility when quality goes below the granted level.

While such guarantees are available for many classical dimensions of quality, for instance *timeliness*, the same does not hold for most of the ethical dimensions. Privacy already has a well defined way for guaranteeing a privacy level by design: (i) in the knowledge extraction step, thanks to the notion of *ϵ-differential privacy* [10], and (ii) in the integration step (see [21] for a survey). The so-called *nutritional labels* [13] mark a major step towards the idea of a quality certificate for fairness and diversity in the source selection and knowledge extraction steps, but how to preserve these properties throughout the process remains instead an open problem. Transparency is perhaps the hardest dimension to guarantee, and we believe that the well-known notion of *provenance* [12] can provide a promising starting point. However, the rise of machine learning and deep learning techniques also for some data integration tasks [8] poses new and exciting challenges in tracking the way integration is achieved [22]. Summing up, recent literature provides a variety of methods for verifying/enforcing ethical dimensions. However, they typically apply to the very early (such as, collection) or very late steps (such as, analytics) of the information extraction process, but very few works study how to preserve *ethics by design* throughout the process.

5 RELATED WORKS AND CONCLUDING REMARKS

An early attempt to consider ethics as a data quality dimension can be found in the book by Batini and Scannapieco [5] already mentioned in this paper. Some dimensions in their *Trust* cluster including, *believability*, *reliability*, and *reputation*, are indeed related to ethical principles. However, they capture how much information derives from an authoritative source, irrespective of the ethical implications of its usage. One of the first attempts to consider the ethical principles systematically in data analysis is in [19]. The work [20] advocates the injection of ethical principles into the whole information extraction process, by properly amalgamating and resolving contrasts among various ethical requirements. The idea of achieving data transparency using blockchain technology, thus without violating privacy requirements, is put forward in [6] and an initial road map for critical research challenges in data transparency is articulated in [7]. The need for high-quality, *unbiased* data for building Artificial Intelligence (AI) systems and more robust data analysis practices is acknowledged in [13] and in the AI ethics guidelines recently published by the EU commission [2]. Finally, [3] brings regulatory frameworks such as European Union's General Data Protection Regulation (GDPR) to the attention of the data management community.

Conclusions. Data quality is defined in [5] as a multifaceted concept, including some dimensions implicitly related to ethical aspects. However, an explicit and holistic vision is still needed to make ethics a first-class citizen for data quality. In this paper, we discussed the challenges of introducing an ethics cluster, arguing that applying an ethical quality control during the three phases of the information extraction process is a necessary step to allow stakeholders to enact law regulations and ensure that individuals can equally benefit from modern data processing techniques.

REFERENCES

[1] Online; accessed 25 Nov. 2019. URL: research.neustar.biz/2014/09/15/riding-with-the-stars-passenger-privacy-in-the-nyc-taxicab-dataset.

[2] Online; accessed 25 Nov. 2019. URL: ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[3] Serge Abiteboul and Julia Stoyanovich. "Transparency, Fairness, Data Protection, Neutrality: Data Management Challenges in the Face of New Regulation". In: *JDIQ* 11.3 (2019), p. 15.

[4] Rakesh Agrawal et al. "Diversifying search results". In: *Proceedings of WSDM*. ACM. 2009, pp. 5–14.

[5] Carlo Batini and Monica Scannapieco. *Data and Information Quality - Dimensions, Principles and Techniques*. Springer, 2016.

[6] Elisa Bertino, Ashish Kundu, and Zehra Sura. "Data Transparency with Blockchain and AI Ethics". In: *JDIQ* 11.4 (2019), p. 16.

[7] Elisa Bertino et al. "Redefining Data Transparency: A Multidimensional Approach". In: *Computer* 52.1 (2019), pp. 16–26.

[8] Xin Luna Dong and Theodoros Rekatsinas. "Data integration and machine learning: A natural synergy". In: *Proceedings of SIGMOD*. ACM. 2018, pp. 1645–1650.

[9] Marina Drosou et al. "Diversity in big data: A review". In: *Big data* 5.2 (2017), pp. 73–84.

[10] Cynthia Dwork. "Differential privacy". In: *Encyclopedia of Cryptography and Security* (2011), pp. 338–340.

[11] Luciano Floridi et al. "AI4People – An ethical framework for a good AI society: opportunities, risks, principles, and recommendations". In: *Minds and Machines* 28.4 (2018), pp. 689–707.

[12] Melanie Herschel, Ralf Diestelkämper, and Houssem Ben Lahmar. "A survey on provenance: What for? What form? What from?" In: *The VLDB Journal* 26.6 (2017), pp. 881–906.

[13] Sarah Holland et al. "The dataset nutrition label: A framework to drive higher data quality standards". In: *arXiv:1805.03677*.

[14] Sidharth Mudgal et al. “Deep learning for entity matching: A design space exploration”. In: *Proceedings of SIGMOD*. ACM. 2018, pp. 19–34.

[15] Emilee Rader, Kelley Cotter, and Janghee Cho. “Explanations as mechanisms for supporting algorithmic transparency”. In: *Proceedings of CHI*. ACM. 2018, p. 103.

[16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Why should I trust you? Explaining the predictions of any classifier”. In: *Proceedings of KDD*. ACM. 2016, pp. 1135–1144.

[17] Babak Salimi, Johannes Gehrke, and Dan Suciu. “Bias in OLAP queries: Detection, explanation, and removal”. In: *Proceedings of SIGMOD*. ACM. 2018, pp. 1021–1035.

[18] Yogesh L Simmhan, Beth Plale, and Dennis Gannon. “A survey of data provenance in e-science”. In: *ACM Sigmod Rec.* 34.3 (2005), pp. 31–36.

[19] Julia Stoyanovich, Serge Abiteboul, and Gerome Miklau. “Data, responsibly: Fairness, neutrality and transparency in data analysis”. In: *Proceedings of EDBT*. OpenProceedings.org. 2016, pp. 718–719.

[20] Letizia Tanca et al. “Ethics-aware Data Governance (Vision Paper)”. In: *Proceedings of SEBD*. CEUR-WS.org. 2018, p. 49.

[21] Dinusha Vatsalan et al. “Privacy-preserving record linkage for big data: Current approaches and research challenges”. In: *Handbook of Big Data Technologies*. Springer, 2017, pp. 851–895.

[22] Xiaolan Wang, Laura Haas, and Alexandra Meliou. “Explaining data integration”. In: *Data Engineering Bulletin* 41.2 (2018), pp. 47–58.