Methods and Techniques for Data Quality Improvement of (Linked) (Open) Data

Maria Angela Pellegrino

Dipartimento di Informatica, Universitá degli Studi di Salerno, Italy
mapellegrino@unisa.it

Abstract. *Good decisions need good data.* Hence, only by exploiting *good* data it is possible to make effective decisions. The goodness of data is usually related to the task they will be used for. However, it is possible to identify some task-independent quality dimensions which are merely related to the data themselves. In order to improve the *intrinsic* data quality, we propose a *proactive* approach. Our goal is to offer data providers (and consumers) a set of methods and techniques to guide them in assessing and improving the quality of data they are interested in. We mainly focus on Linked (Open) Data. Since the published data might also contain personal data, there is the need to make the data set compliant with the General Data Protection Regulation (GDPR). Therefore, besides quality problems, we are also interested in discovering any privacy breach and - if needed - in proposing corrective actions. The final goal is to give data providers the possibility of publishing *better* data. The proposed approach is pragmatic. Thus, we will not only design but also implement it. We plan to wrap it into a social platform, already used by several public administrations, which enable us to test the applicability of the proposed methods in real settings.

Keywords: Data quality, Privacy breaches, Quality assessment, Quality improvement, Privacy awareness, Data publication

1 Problem statement

Data Quality (DQ) can be defined as "*as the level of compliance of the data with the purpose they will be used for*" [1]. Thus, data quality is defined in terms of *fitness of use*. The exploitation of data *not* ready-for-use may lead to incorrect conclusions and poor decisions. Only by using high-quality data it is possible to achieve effective decision making. Thus, data providers should make an effort to improve the quality of data sets under definition to simplify their exploitation. Moreover, data providers might also deal with data sets containing *personal data*. The General Data Protection Regulation (GDPR) [22] defines which data are considered personal and in which case the individual privacy can be compromised in order to increase the utility for the community. Therefore, data providers have to make data sets compliant with the GDPR before the publication.

The problem we want to face is how to design a unified approach which allows to publish high-quality data while preserving individual privacy. Since we are interested in both Quality and Privacy, we call our approach *Qualicy* aware.
The general workflow to assess and improve quality/privacy aspects should be: 1) choose the quality dimensions of interest, 2a) assess the quality and 2b) detect privacy problems, 3a) improve the overall data set quality and 3b) prevent privacy breaches. By a *reactive* approach, the quality can be improved after the data set publication, for instance when it has to be used in a practical use case. The reactive philosophy can be summarised by "*publish first, refine later*". The alternative is to adopt a *proactive* approach by improving the data set quality as early as possible. Ideally, it might be improved during the publishing phase.
Our proposal is to provide data publishers a set of (semi-)automatic techniques to identify quality problems and improve the overall quality of a data set before its publication. Moreover, we aim to take privacy concerns into account and prevent personal information leakage. Our *pragmatical* approach will be integrated into SPOD (Social Platform for Open Data) which can be used by citizens, Public Administrations (PAs), associations, and every kind of stakeholder in order to produce and consume Open Data (OD) also in Linked format. Our approach must not require technical skills to be compliant with the SPOD audience. Since data can be both in tabular and linked format, we plan to work with data in general and try to define strategies independent of data format. Only when it is necessary we intend to use peculiarities of the specific data format.
In conclusion, we can summarise our goal as *the definition of strategies to assess and improve data quality and manage privacy aspects of (Linked) (Open) Data.* The parentheses delimit the parts which can be omitted. In other words, we plan to work i) with Data in general, ii) with Open Data in tabular format (3-star data according to Tim Berners-Lee's rating system [3]), iii) with Linked Data, iv) also released with the open license as 5-star data [3], i.e. Linked Open Data.

1.1 Data Quality

Several quality dimensions and taxonomies have been defined to evaluate data quality. Ballou and Pazer [2] identify *accuracy*, *completeness*, *consistency*, and *timeliness* as main quality dimensions. Wand et al. [29] classified quality dimensions in *intrinsic*, *accessibility*, *contextual*, and *representational* DQ: data should be i) intrinsically of a good qualitative level and ii) accessible; iii) they should be compliant with the context they will be used for and iv) also the format itself should be qualitatively good. Besides these general definitions, data quality dimensions can been specialised for the Linked (Open) Data (LOD). According to Zaveri et al. [30], *accuracy* and *completeness* belong to the *intrinsic* data quality. They further distinguish *syntactic* and *semantic* accuracy.

Syntactic accuracy. A value is syntactically accurate when it is valid, i.e. it belongs to the set of acceptable values according to the domain of interest [12]. Therefore, *the syntactic accuracy (also called syntactic validity) is the degree of*

conformity to the syntactic rules determined by the modelled domain.
The metrics identified for the syntactic validity are

- detecting the explicit definition of the allowed values for a certain data type,
- detecting the compliance of values with syntactic rules (e.g. patterns),
- detecting the presence of outliers,
- detection of typos in literals.

Semantic accuracy. According to Zaveri et al. [30], the *semantic accuracy is defined as the degree to which data values correctly represent real-world facts.* For instance, supposing that the flight between Paris and New York is A123, while in a data set the same flight instance is represented as A231. In this case, the instance is semantically inaccurate since the flight ID does not represent its real-world state [30]. The metrics identified for semantic accuracy are:

- detection of outliers by using distance-based methods,
- detection of inaccurate values comparing values of different properties,
- detection of inaccurate classifications and labelling.

Completeness. Fürber et al. [12] classified completeness into i) schema completeness, ii) column completeness, iii) population completeness, and iv) interlinking completeness. *The Schema completeness is the degree of completeness of the ontology*, i.e. there are no relevant classes and properties not represented in the ontology. *The column completeness can be defined as the number of missing values for a specific property/column. The population completeness is the percentage of the coverage of all the real-world objects of a particular type represented in the data sets. The interlinking completeness (specific for LOD) refers to the degree to which the instances contained in the data set are interlinked.*

2 Relevancy

By providing methods to publish high-quality data, the effort and the time needed to make data ready-for-use will be reduced. Since we want to improve the (Linked) (Open) Data quality, the problem is relevant for all data publishers, contributors, and consumers. Moreover, the proposed approach will be wrapped into SPOD which is already adopted by several users, such as our national Public Administration and cultural associations. Therefore, on one side they can benefit from our results; on the other side, they can also be involved in the evaluation phase of our approach in order to assess its applicability in real settings.

3 Related work

Linked (Open) Data quality assessment. SWIQA [12] is a quality assessment framework which relies solely on Semantic Web technologies, without any external source. As our proposed approach, SWIQA may be used both by data

consumers to find high-quality data sources and by data owners to evaluate the quality of their own data. They selected quality dimensions which rely only on the data source, without caring about the specific task they will be used for. Thus, they aim to provide an *objective* - i.e. task independent - quality assessment. If we limit ourselves to the intrinsic DQ, they cover syntactic and semantic accuracy, completeness, timeliness, and uniqueness. In general, they consider a wider range of metrics. They evaluate the metrics based on the Closed Worlds Assumption (CWA), i.e. everything that is not known can be assumed as false. This hypothesis is due to the metric definitions. However, typically the Semantic Web assumes an open world, i.e. everything we do not know is not defined yet. The CWA might be not always applicable since LOD suffer from incompleteness. Sieve [19] is based on the opposite assumption: it considers data quality strictly dependent on the task. Therefore, the user can customise the settings by specifying metrics, scoring functions, and aggregation functions in an XML file. It evaluates both the semantic accuracy and the completeness of the queried LOD. About how to assess data quality and display results, Langer et al. [15] report a clear workflow to evaluate a set of metrics and report results. Looking at the provided results, users can also change quality desiderata. It implies a cyclic process in order to define/evaluate/refine quality metrics. This theoretical workflow is implemented into SemQuire [15] which is focused on quality assessment. SHACL Shapes Constraint Language[1] is a W3C standard to validate LOD against a set of conditions. It is useful for different purposes, e.g. data integration.
Other interesting works can be found in a survey written by Zaveri et al. [30]. Cited work focus only on data quality assessment without considering the improvement step. Moreover, they do not provide a privacy-aware process.

Linked (Open) Data quality improvement. In his survey, Hadhiatma [13] underlines the need for a framework which helps in improving the LOD data quality. They count several approaches which exploit inductive learning methods in order to enrich and complete LOD. Among them, Paulheim [23] defined an algorithm able to detect co-occurrences and patterns in DBpedia types. Sleeman and Finin [26] worked on a labelled training set to predict the type of instances. In general, machine learning, statistical methods, and external knowledge are the mainly employed methods to detect patterns and find missing information [13]. DaCura [8] is a framework developed to help data set curators. Because their users may not have technical skills, we share the same audience. The framework is made up of a collection of tools able to detect and curate quality problems over the evaluation of linked data sets. Therefore, it is used both to assess and to improve the data set quality. DaCura and our proposed approach share the idea that the quality should already be affected in the definition stage. Moreover, the process has to be cyclic. If we consider only the intrinsic DQ, Freeney et al. [8] address both the accuracy and the completeness quality metrics. In general, they consider a wider set of metrics, including several metrics which we are not

[1] https://www.w3.org/TR/shacl/

considering at the moment. On the other side, our goal is to consider both the quality and privacy awareness - which is completely absent in DaCura.

Privacy awareness in LOD. A typical content-based data leakage prevention system (DLPS) works by monitoring sensitive data mainly by using regular expressions, data fingerprinting and statistical analysis. Regular expressions are normally used under a certain rule such as detecting social security numbers and credit card numbers. Dataguise, a leader in data privacy protection and compliance, will demonstrate how DgSECURE is supporting enterprise administrators as the basis for secure data analytics, application testing and development, and the general protection of sensitive data across enterprise cloud repositories. DgSECURE [6] enables you to discover, count, and report on sensitive data assets through a sophisticated regular expression (regex) pattern builder; it combines structured, semi-structured, or unstructured content and it finds sensitive data - such as credit card numbers, SSN, names, email addresses. For what concerns the anonymization, it is well-consolidated approach [16,18,28] in relational data. However, its counterpart on LOD is still under development [17,31]. The main concern is that both the de-anonymization techniques and LOD base their strength on interlinking. However, researchers working on heterogeneous graph de-anonymization are trying to reuse and adapt approaches already used in a homogeneous graph, e.g. social networks. These approaches are mainly based on clustering and graph modification [33]. One of the considered approaches is k-RDF-Neighbourhood Anonymity [14] which adapts the k-Neighbourhood [32] algorithm to LOD released as RDF graphs. It is rare to find a technique able to manage both the graph structure and the attributes attached to each node. k-Neighbourhood is able to manage both structural and attribute aspects.

4 Research questions

Our research questions (RQs) deal with both the assessment and the enhancement steps and both considering quality dimensions and the avoidance of privacy breaches. The RQ related to the assessment steps can be summarised as follows:

RQ1. *To what extent data quality and privacy concerns can be assessed independently of the data format? In which case - if any - is there the need to consider the original data format?* In order to define only once the quality dimensions and reuse it both for OD and LOD, we want to investigate if the quality dimensions - in particular focusing on *accuracy* and *completeness* - can be defined independently of the data format without loosing in precision. The same consideration holds for privacy aspects.
RQ2. *Can (automatic) data type inference be useful (in terms of effectiveness and efficiency) in (linked) (open) data quality assessment?*
RQ3. *Can (automatic) data type inference be useful (in terms of effectiveness and efficiency) in discovering privacy breaches?*

The main RQ related to the improvement phase can be summarised as follows: ***RQ4.*** *How to improve data quality while preventing privacy breaches?*

Research questions are presented in the same order in which they are considered during my Ph.D. It also justifies the different degree of refinement of the RQ. During this year (which is my first year of Ph.D.), I will mainly focus on quality and privacy assessment, while in the following years I will focus, first, on how to improve data quality and, then, how to manage privacy leakages. Therefore, the fourth question will be further refined in the future.

5 Hypotheses

At this stage of the work, we are able to hypothesise results only about the quality and privacy assessment. ***H1*** is related to ***RQ1***, while ***H2*** is related to ***RQ2*** and ***RQ3***.

H1. We hypothesise that it is possible to define approaches which work directly on values (or their collection) without caring about the original data set format. ***H2.*** We consider our automatic *data type inference* (which will be detailed in section 6) a suitable method to address both quality problems and privacy concerns. We defined and implemented an approach to automatically infer the type exclusively working on values. Inferred data types are consequently used i) to give an insight about quality aspects and ii) to detect if privacy breaches occurred. The performance and the scalability of this approach have been tested on open data sets organised in tabular format. In the near future, we aim to verify if we gain the same (positive) results also on LOD. In particular, we plan to verify if it returns correct results and if it is the most efficient way to manage it. If so, it is a first step in defining promising techniques that are independent of the original data format. Consequently, we should verify if the same consideration can be expanded to other assessment and enhancement approaches.

6 Preliminary results

We designed and implemented an approach [9] to assess the quality level and the occurrence of privacy leakage working on the actual content of each value. Our approach infers not only basic data types (such as string, number, date) but also *meta data types* inspired by the GDPR. The novelty of our approach does not lie in the type inference step, but in the exploitation of inferred types both to assess quality aspects and to detect privacy breaches. About privacy breaches, we check if i) a *content privacy breach* occurred: we verify if a description (i.e. any field classified as string without a refined meta data type) contains any structured sensitive information, such as phone number, IBAN, SSN; ii) a *structural privacy breach* occurred: by considering meta data types attached to the columns, we verify if data provider badly designed the data set by forcing users to fill in cells with personal information. More in detail, our approach works as follows:

- it takes as input a data set seen as a collection of columns
  * for each column seen as a collection of values
    1. for each value, the *data type* is *inferred* according to its content. Our approach attaches to each value a *basic data type* - such as number, string, and date - and (if possible) a *meta data type* - such as province, municipality, ZIP code, SSN, IBAN, email, address, surname, name and so on. The latter is used to capture the semantic of the value. By default, each value is a string as basic data type and it has no meta data type;
    2. for string values (i.e. for the values which the type inference approach fails in refining the meta data type), the proposed approach checks if a *typo* occurs. In other words, it verifies if by replacing, adding, removing or swapping letters a known meta data type is matched;
    3. if also the typo check fails, the proposed approach verifies if the value contains a structured personal data, e.g. if it contains an IBAN, an email, an address. In that case a *content privacy breach* occurred;
  * to each column we assign the most frequent data types among its values;
  * for each collection, the *completeness* and the *accuracy* are computed;
- once a data type has been attached to each column, the type inference module checks if a *structural privacy breach* occurred. By structural privacy breach we mean both the presence of information which exposes individually personal details - e.g. SSN - or the co-occurrence of quasi-identifier, i.e. bits of information which identifies unequivocally an individual - e.g. the co-occurrence of date of birth, gender and ZIP code.

Both the correctness and the scalability of this approach have been evaluated on open data sets [9]. The approach is completely independent of the data format: starting from a tabular or a linked data set, it is possible to work on each value and to assess the quality and privacy aspects by our approach. We already provided SPOD with a prototype of this approach. Moreover, SPOD is also enhanced with a component [7] to query LOD by SPARQL and organise the results into a tabular format. The results of a SELECT query can be always organised in tabular format. Therefore, we plan to verify the applicability of our approach also to LOD by organising queried data in a tabular view.

7 Approach

Our pragmatic approach aims to help data providers 1a) both in assessing data quality problems and 1b) in identifying privacy leakages and 2) in providing effective and efficient strategies in solving detected problems. It is interesting to notice that by improving a quality dimension, the other ones could be compromised. For instance, by making data compliant with the GDPR, the completeness could be compromised: to anonymise ZIP codes we might omit the last two digits. In this way, the completeness (and also the accuracy) is affected. This symbiosis of causes and solutions of quality and privacy aspects should be taken into account when defining a data quality assurance process. The entire process can be implemented as a cyclic approach, summarised as follows:

1. data quality assessment
   (a) definition of the quality dimensions to assess
   (b) measurement of the chosen quality dimensions
   (c) representation of the results of the measurements
2. data quality improvement
3. check if the improvements negatively affected the other quality dimensions (also called validation)

In the validation step all the measurements for all the considered quality dimensions must be repeated. Thus, the validation step matches the measurement step. The approach can be graphically represented by Figure 1.

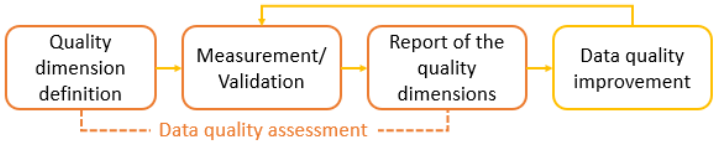

Fig. 1. Schema of the proposed approach.

Quality assessment. We decided to focus on accuracy and completeness.

Syntactic accuracy. We plan to:

- apply our type inference approach on LOD. Then, we want to compare inferred data types with data types specified in the queried LOD. For instance, supposing to test all the values of the `dbo:birthDate` property. We can verify if they are correctly recognised as dates.
- enhance the type inference approach to recognise patterns which are attached to syntactic rules reported in the queried LOD. For example, supposing that a relation has a data type pattern not supported by our type inference approach (e.g. date_time), we can add the regex to recognise it and identify any syntactical wrong values;
- exploit either clustering algorithms or statistical approaches to detect outliers. At this moment, we are following the same approach described by Fleischhacker et al. [10] and we are comparing DBscan, IQR, and Z-score in order to verify which is the most accurate and efficient technique;
- compare the actual typo detection approach (part of the type inference process) with clustering algorithms. The main drawback of the first approach is the scalability: for each string for which a typo is hypothesised, it computes *all* the words by adding, removing, swapping or replacing a letter against the original word. Obviously, this naïve approach explodes if we consider more than one error. Our hypothesis is that a clustering algorithm achieves

better results, gaining also in efficiency. The main difficulty is in detecting clustering algorithms able to deal with strings. At this moment, we are comparing k-means and the agglomerative clustering to identify the most accurate and efficient approach. An alternative is to exploit word or graph embedding techniques to convert words (or the whole graph) into vectors and use them to feed in clustering algorithms. Right now we are evaluating the performance of KGloVe [5] and RDF2Vec [24] upon the clustering task.

Semantic accuracy. We plan to check if a set of values is semantically accurate, by forecasting the values by other properties and then compare the predicted against the actual ones. We aim to exploit either link prediction or external resources. External resources are strictly dependent on the tested source. For example, in order to validate data in DBpedia, we can use other well known Knowledge Bases, such as Wikidata or Freebase;

Completeness. We plan to compare the *column completeness* calculated by the type inference module against the one calculated directly on the graph. We will further consider how to evaluate the other completeness dimensions.

Detection of privacy breaches. GDPR categorises as personal data "*any information relating to an identified or identifiable natural person; an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, ...* ". While some bits of information may not be uniquely identifying individuals on their own, they can be potentially identifying individuals when combined with other attributes [21,27]. The combination of these attributes is defined as *quasi-identifier*. [22]. Both the occurrence of personally identifiable information (PII) and quasi-identifier are detected by our data type inference approach. It can be further enhanced to recognise a wider range of *structural privacy breaches*. We also deal with *content privacy breach*. As an alternative, we could exploit sentiment analysis techniques or classification algorithms to distinguish sensitive and not-sensitive information.

Data quality improvement. To improve the data quality, we plan to perform *data enrichment*. It can be realised by exploiting clustering algorithms in order to identify new classes and grouping. To address the completeness requirement and enrich the data set, we plan to apply link prediction techniques.
By *data cleansing* approaches, we aim to recognise erroneous data and clean them. Applying Machine Learning (ML) approaches to LOD raises several difficulties: LOD lack of negative examples [25], and performing the feature extraction phase on graphs is particularly expensive. Besides applying ML algorithms directly on LOD, entities (and relations) can be vectorised by graph embedding techniques [4,5,24]. The obtained vectors will then be fed in ML algorithms.

Privacy leakage avoidance. The privacy-preserving data publishing [11] guarantees methods and tools to publish data set by reaching a good trade-off between privacy preservation and the overall utility of the published data set. The

anonymization techniques - also suggested by the GDPR - hide personal data based on the idea that they should not be involved in statistical analyses. A naïve solution is the removal of PII, e.g. SSN, name, and surname. However, because of the power of modern re-identification algorithms [20], removing PII data does not guarantee that the remaining data does not identify individuals. In order to make data sets compliant with the GDPR, we want to investigate the k-RDF-Neighbourhood Anonymity [14].

Non-functional requirements. Besides the functional requirements, the proposed approach has to address the following non-functional requirements:

- ***reversibility*** of the actions in order to provide data providers the possibility to perform the undo of every action;
- ***traceability*** of the actions. This requirement is inspired by the potential occurrence of several actors involved in the definition and maintenance of data sets under definition. Therefore, there is the necessity to keep track of the performed actions and their owner;
- ***efficiency***;
- ***scalability*** since the data could increase dramatically;
- ***interactivity*** since the data publishing and quality improvement could involve several actors which have to work collaboratively.

8 Evaluation plan

To assess our approach we plan to evaluate the *scalability* by considering data sets of increasing size. About the *performances* we will consider both the time and the space used. Moreover, we plan to evaluate the *correctness* by i) manually checking the results, ii) by using data set as a gold standard, iii) by comparing them with results obtained by other tools iv) or by evaluating the same metrics upon a different data format. For instance, the correctness calculated through the type inference approach described above can be validated against the value calculated upon the graph. To verify the *usability* and *applicability* of our approach, we plan to involve SPOD users in order to check how it performs in real settings. The research described here is conducted in strict cooperation with our PA and their ICT department. Therefore, they are interested in testing our results and verify if they can be practically exploited in their every-day work.

9 Reflections

To the best of our knowledge, quality aspects and privacy concerns are rarely managed simultaneously, both in OD and LOD. Therefore, our goal is to fill up this gap by proposing a framework which helps data providers and consumers in assessing and improving data quality, while preventing personal information leakage. Moreover, the features offered by this framework will be integrated into SPOD in order to reach a wider range of users both to help them in providing

data of *better* quality and to test the applicability of our proposal. Since this is my first year of Ph.D., I plan to work on the assessment phase and study how privacy concerns can be managed by the end of this year. I will dedicate the next year to the quality improvement both studying the most reliable solutions and by developing our own approach. To avoid reinventing the wheel, each step is preceded by a study phase. I plan to reuse the most promising approaches used in literature and defining our own approach to feel any gap. The third - and last - year is dedicated to the evaluation and improvement of the proposed approach by the collected considerations. The main novelty of our approach is to provide a unique interface to manage both quality and privacy concerns.

Acknowledgement

I would like to thank my supervisor Prof. Vittoro Scarano for his support.

References

1. DAMA international. The DAMA guide to the data management body of knowledge, https://dama.org/content/body-knowledge, last access April 15th, 2019
2. Ballou, D.P., Pazer, H.L.: Modeling data and process quality in multi-input, multi-output information systems. Manage. Sci. **31**(2), 150–162 (1985)
3. Berners-Lee, T.: 5-star open data, https://5stardata.info/en/, last access 04-2019
4. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 26, pp. 2787–2795. Curran Associates, Inc. (2013)
5. Cochez, M., Ristoski, P., Ponzetto, S.P., Paulheim, H.: Global RDF vector space embeddings. In: The Semantic Web - 16th International Semantic Web Conference, Proceedings, Part I. pp. 190–207 (2017)
6. Dataguise: DGSecure (2018), https://www.dataguise.com/detect/, last access 01-2019
7. Donato, R.D., Garofalo, M., Malandrino, D., Pellegrino, M.A., Petta, A., Scarano, V.: Linked data queries by a trialogical learning approach. In: 23rd IEEE International Conference on Computer Supported Cooperative Work in Design (2019)
8. Feeney, K., O'Sullivan, D., Tai, W., Brennan, R.: Improving curated web-data quality with structured harvesting and assessment. International Journal on Semantic Web and Information Systems **10**, 35–62 (2014)
9. Ferretti, G., Malandrino, D., Pellegrino, M.A., Pirozzi, D., Renzi, G., Scarano, V.: A non-prescriptive environment to scaffold high quality and privacy-aware production of open data with AI. In: Digital Government Society. Dg.O. (2019)
10. Fleischhacker, D., Paulheim, H., Bryl, V., Völker, J., Bizer, C.: Detecting errors in numerical linked data using cross-checked outlier detection. In: The Semantic Web - ISWC. pp. 357–372 (2014)
11. Fung, B.C.M., Wang, K., Chen, R., Yu, P.S.: Privacy-preserving data publishing: A survey of recent developments. ACM Comput. Surv. **42**(4), 14:1–14:53 (2010)
12. Fürber, C., Hepp, M.: SWIQA - a semantic web information quality assessment framework. In: ECIS Proceedings (2011)

13. Hadhiatma, A.: Improving data quality in the linked open data: a survey. Journal of Physics: Conference Series **978**, 12–26 (2018)
14. Heitmann, B., Hermsen, F., Decker, S.: k - RDF-neighbourhood anonymity: Combining structural and attribute-based anonymisation for linked data. In: Proceedings of the 5th Workshop on Society, Privacy and the Semantic Web - Policy and Technology co-located with 16th ISWC (2017)
15. Langer, A., Siegert, V., Göpfert, C., Gaedke, M.: Semquire - assessing the data quality of linked open data sources based on DQV. In: Current Trends in Web Engineering. pp. 163–175 (2018)
16. Li, N., Li, T., Venkatasubramanian, S.: t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. IEEE 23rd International Conference on Data Engineering (2007)
17. Liu, K., Terzi, E.: Towards Identity Anonymization on Graphs. Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data (2008)
18. Machanavajjhala, A., Gehrke, J., Kifer, D., Venkitasubramaniam, M.: L-diversity: privacy beyond k-anonymity. 22nd International Conference on Data Engineering (2006)
19. Mendes, P.N., Mühleisen, H., Bizer, C.: Sieve: Linked data quality assessment and fusion. In: Proceedings of the Joint EDBT/ICDT Workshops. pp. 116–123 (2012)
20. Narayanan, A., Shmatikov, V.: Robust de-anonymization of large sparse datasets. In: 2008 IEEE Symposium on Security and Privacy (sp 2008). pp. 111–125 (2008)
21. Narayanan, A., Shmatikov, V.: Myths and fallacies of "personally identifiable information". Commun. ACM **53**(6), 24–26 (2010)
22. Parliament, E.: General data protection regulation (2018), https://eur-lex.europa.eu/eli/reg/2016/679/oj, last access 04-2019
23. Paulheim, H.: Browsing linked open data with auto complete. In: Semantic Web Challenge (2012)
24. Ristoski, P., Paulheim, H.: RDF2Vec: RDF graph embeddings for data mining. In: The Semantic Web - 15th International Semantic Web Conference, Proceedings, Part I. pp. 498–514 (2016)
25. Simperl, E., Norton, B., Acosta, M., Maleshkova, M., Domingue, J., Mikroyannidis, A., Mulholland, P., Power, R.: Using Linked Data Effectively. The Open University, Milton Keynes (2013)
26. Sleeman, J., Finin, T.: Type prediction for efficient coreference resolution in heterogeneous semantic graphs. In: IEEE Seventh International Conference on Semantic Computing. pp. 78–85 (2013)
27. Sweeney, L.: Simple demographics often identify people uniquely (2000), http://dataprivacylab.org/projects/identifiability/
28. Sweeney, L.: K-anonymity: A Model for Protecting Privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst. (2002)
29. Wand, Y., Wang, R.Y.: Anchoring data quality dimensions in ontological foundations. Commun. ACM **39**, 86–95 (1996)
30. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: A survey. Semantic Web **7**(1), 63–93 (2016)
31. Zheleva, E., Getoor, L.: Preserving the Privacy of Sensitive Relationships in Graph Data. Privacy, Security, and Trust in KDD (2008)
32. Zhou, B., Pei, J.: Preserving Privacy in Social Networks Against Neighborhood Attacks. Proceedings of the 24th International Conference on Data Engineering, ICDE (2008)
33. Zhou, B., Pei, J., Luk, W.: A Brief Survey on Anonymization Techniques for Privacy Preserving Publishing of Social Network Data. SIGKDD Explor. Newsl. (2008)