THE NATIONAL UNIVERSITY
of SINGAPORE

School of Computing
Computing 1, 13 Computing Drive, Singapore 117417

TR 11/21

Survey on Data Quality and Provenance

Martin Schmitz

November 2021

T e c h n i c a l R e p o r t

Foreword

This technical report contains a research paper, development or tutorial article, which has been submitted for publication in a journal or for consideration by the commissioning organization. The report represents the ideas of its author, and should not be taken as the official views of the School or the University. Any discussion of the content of the report should be sent to the author, at the address shown on the cover.

Mohan KANKANHALLI
Dean of School

Survey on Data Quality and Provenance

School of Computing Technical Report, TR 11/21

Martin Schmitz

November 2021

Abstract

This technical report summarizes research on data quality, provenance and truth discovery from the last decades. It examines opportunities to use machine learning methods to enhance data quality and provenance. This report can serve as a starting point to find the key publications of the topics "provenance" and "data quality" and to do further research in those areas in general as well as in combination with machine learning algorithms.

1 Introduction

Data quality is a topic with increasing importance as our world gets more and more dependent on data with the increasing use of machine learning methods and the ongoing digitalization. Data provenance, which is a term to describe the history of a data set, is useful to validate the quality, and truthfulness of data and to investigate dependencies between different data sets. This technical report includes two sections about research on the two areas "data quality" and "provenance" and a conclusion section, where we discuss the field of truth discovery and machine learning methods to improve data quality and provenance.
Section 2 gives an introduction into the topic of data quality and explains the challenges of finding a general-purpose definition for the term. We gather the most important publications from the 1990s to today and show how various authors use different methods to come up with different criteria for data quality. Section 3 deals with the concept of provenance. We first look at the history and definitions of provenance. Then we show different mathematical models to formulate provenance and techniques to include provenance into workflows with databases. In section 4, the conclusion, this report briefly shows important research from the area of truth-discovery and discusses the role of provenance and data quality for this domain. Truth-discovery is a topic where data quality and provenance meet machine learning methods and thus a good starting point for research, which includes the combination of those topics. Finally, we gather thoughts of how one could use modern machine learning models to improve provenance techniques and to enhance data quality.
Each paragraph includes two lists of other publications at the end. Those are the relevant publications of the topic (i.e. provenance or data quality) which are citing the publications referenced in the paragraph and the publications which are cited by the described publications.

2 Data Quality

Data Quality in the 90s Different topics and applications have different needs for their data, which makes it difficult to come up with a single general-purpose definition for data quality. From the 1990s, there have been several attempts to form definitions for data quality, and to create frameworks to practically assign quality measures on data sets. Eppler et al. 2000 [29] gather and summarize the most influential of those data quality (during this time mostly called information quality) frameworks from the 90s. They gather 20 data quality frameworks from 1989-1999 to analyze and evaluate those frameworks. The authors show the final definitions of the gathered publications as well as the scientific approach of how the authors came up with their respective definition for data quality. Some of the frameworks are generic, some are for specific areas. Eppler et al. perform a deeper analysis of 7 of the

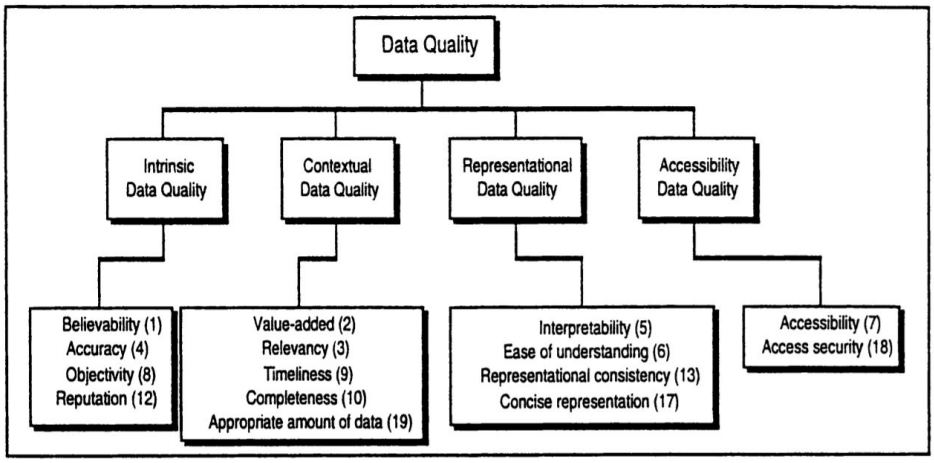

Figure 1: Data Quality Framework after "Fitness for Use" approach by Wang and Strong 1996

found frameworks. They conclude that most frameworks entail a "time" and a "accessibility" dimension to measure data quality. Apart from that, all frameworks contain significant differences in the final choice of attributes, and none of them deal explicitly with trade-offs and other interdependencies between attributes.
Two examples of publications that find data quality attributes are Klein 2002 [45] and Veregin et al. 1999[71]. Klein does a survey on students which results in the 5 categories: "accuracy", "completeness", "relevance", "timeliness", and "amount of data". Veregin defines data quality as following attributes: "accuracy", "precision", "consistency", and "completeness" and defines those attributes for data in the domain of geography. Those publications are just two examples and as explained above a lot of different researchers come up with completely different attributes, based on different approaches and different use cases for their data in different domains.

Data Quality as Fitness for Use One strongly influential group for research in this field is the "Total Data Quality Management group" of MIT University led by Professor Richard Y. Wang. Wang's group distinguishes two types of data quality: objective measurements and subjective perceptions on the data (from stakeholders). Moreover, objective measurements can be task-independent or task-dependent [60]. The books "Journey to Data Quality" from (Yang W. Lee, Leo L. Pipino, James D. Funk, Richard Y. Wang) from 2006 [49] and "Data Quality" (By Richard Y. Wang, Mostapha Ziad, Yang W. Lee) [73] are fundamental works of the topic.
In their publication [72] from 1996, Wang and Strong gather thoughts on how to create a framework that captures the aspects of data quality. They classify research on data quality attributes into 3 categories: (1) intuitive (on researchers experience), (2) theoretical, and (3) empirical. While most approaches are from the 1st category, Wang's team chose the 3rd one and ask data consumers/stakeholders about their preferences.
For their own data quality framework they define "data quality" as "fitness for use". This means, they use the consumer's opinion as the metric to measure data quality. They conduct a survey and a sorting study to produce a data quality framework to capture dimensions that are important to their stakeholders. Figure 1 shows the result of 15 weighted attributes for data quality categorized in 4 categories. Categorizing data quality attributes into 2 layers (with categories and subcategories) is common between many data quality frameworks. (Note that Wang's data quality framework is one of the frameworks, which are compared in the previously noted summary of data quality frameworks by Eppler et al. 2000 [29])
In their publication Pipino et al. 2002 [60], Wang's team focuses on the Development of data measurement metrics. Different companies and projects should choose their data quality measures differently, according to their tasks. The authors provide a large set of possible data quality attributes and propose different functional forms to measure data quality on those attributes, like simple ratio, min or max operations and weighted average of attributes.

To summarize their work, data quality is decided by the needs of the stakeholders and data users. Thus, for every project one has to define data quality differently in line with the different stakeholders and their interests.

Recent works on Data Quality The publication "data quality for big data assessments" by Cai et al. 2015 [17] gathers more recent data quality publications. The authors list studies and conclude that existing work focuses on two areas: web data quality and data quality in specific domains (i.e. biology, medicine, geophysics, telecommunications, scientific data, etc). They also highlight that big data is becoming more and more important in data quality research.
One example of work focusing on big data is Katal et al. 2013 [44]. According to this work, big data comes with the 4Vs: Volume, Velocity, Variety, and Value which lead to a set of different challenges for data quality. Like many other data quality frameworks, the authors define a 2 layer framework, consisting of categories and subcategories for data quality. For example "availability" is a category with subcategories "accessibility", "timeliness", and "authorization". The authors claim they use data quality dimensions commonly accepted and widely used as big data quality standards and redefined their basic concepts based on actual business needs. They come up with the categories "Availability", "Usability", "Reliability", "Relevance", "Presentation Quality" and define those and the subcategories. They also propose a pipeline for data quality assessment.
In conclusion, there are different data quality benchmarks and metrics for different industries and even for a single industry, it is hard to find a general data quality definition. Also over time, the industries change and different data aspects increase or decrease in overall importance.

3 Provenance

Definition Provenance, pedigree, lineage, traceability, and attribution are all terms to describe the origin and history of data. In some publications those terms are referred to as slightly different terms, while sometimes they are referred to as synonyms. A simple definition for the terms would be "the history of a data set". The first publication to address this issue of "where does the data originate from" and "which intermediate data sources were used to arrive at that data" was Wang et al. 1990 [74]. In this first publication the authors describe provenance as "data source tagging" and "intermediate source tagging" problems. They formally define the problems and lay the first theoretical foundation for them.
The term data lineage first appeared in [46] in 1991. Data lineage is first described as an output record of all the contributory inputs of a data set. It means keeping the history of the data including origin and changes inside the metadata. The term provenance was first introduced in 2000 by Bunemal et al. [13]. Buneman describes provenance as an open problem in scientific databases, where users should know about the origin and perform modifications on the data they use. DeLusignan et al. 2011 [24] explains differences between the terms in more detail and gathers various definitions from different authors and discusses them.
In publications around the year 2000, the term lineage is mostly used, while recent publications usually use the term provenance as noted by Senerellart 2018 [63]. As it is the more modern term, we will further use the term provenance in this report. Provenance plays an important role in the area of data quality as information about the origin and history of data is crucial to measure the quality. Also, an accurate history of a data set enhances data quality as it boosts some attributes like credibility and transparency of the data.
Some authors distinguish multiple sub-categories of provenance, like why-provenance, how-provenance, and where-provenance. The book "Provenance in Databases. Why? How? Where?" from Cheney et al. 2009 [18] gives a deeper explanation of those terms and an exhaustive overview of provenance. Moreover, they explain the mathematical background and gather the research up to 2009 on the topic. However, similar to the problem with data quality explained above, different projects and branches have different criteria for what they exactly need. Therefore it is not easy to agree on a standard provenance technique or representation, as different areas might have different needs. Thus, it is a challenge for a suitable provenance standard to evolve.
In contrast to the data quality problem, the problem with provenance is a more mathematical problem. It is not just a matter of how to represent the history but also which mathematical constructs are used for the underlying operations. Additionally, it is a technical problem, for example to get provenance

information while querying a sample in a large heterogeneous database. The following paragraphs give an overview of the first provenance techniques, more recent publications and the mathematical view on provenance as well as applications for specific areas.
Citing: Groth et al. 2004 [37], Moreau et al. 2011 [55], Lanter 1991 [46], Karvounarakis et al. 2012 [43], Hartig 2009 [39], Green et al. 2007 [35], Ives et al. 2008 [41] Deutch et al. 2014 [26], Davidson et al. 2008 [23], [13], Amsterdamer et al. [6] [4], Amarilli et al. 2015 [2], Amarilli et al. 2016 [3]
Cited By: Moreau 2010 [54], Meliou et al. 2010 [52], Amsterdamer et al. 2011 [4], Missier et al. 2013 [53], Senellart et al. 2018 [65], Buneman 2019 [16], Ramusat 2018 [61]

First Provenance Techniques The publication Wang et al. 1990 [74] first addressed the issue of finding the source of a data sample. Their "polygen model" is a direct extension of the relational model for multiple databases. Given a query, their model makes it possible to not only get the query result but also the source of the data (they define the source of a data sample as the name of the database from which a data sample is taken). Therefore they translate a polygen query into a set of local queries using their proposed query translation mechanism. Moreover, they define a formal algebra for their polygen queries.
Lee et al. 1998 [47] extend this work to mediation networks. Those networks consist of various data sources including relational databases but also semi-structured data sources like web data. The authors refer to the problem of provenance as "ownership attribution". They define "attribution" as an association between a value and a source of a data sample and offer a general framework to solve this attribution problem for queries in heterogeneous data sources. Additionally, they introduce a formal algebra for attribution.
In follow-up work, Lee et al. 1999 [48] introduce one of the first practical tools for extracting provenance metadata while querying databases. The tool is called 'CI', a corporate information integrator, and applies XML as meta-data syntax. The prototype is designed to gather corporate information to help data analysts in their daily activities at a company. Many other works discussed in the next paragraphs also use XML or RDF based modeling for provenance information.
Citing: Cui and Widow 2000 [21]
Cited By: Cheney et al. 2009 [18], Cui and Widow 2003 [20], Buneman et al. 2006 [12], Cui and Widow 2000 [19], Tan 2004 [70]

Provenance Overview There are various attempts to create metadata standards for provenance representation. The publication "taxonomy to understand and compare provenance techniques" Simmhan et al. 2005 [66] gives an overview and taxonomy of provenance. They summarize various applications for provenance. They classify provenance techniques into two groups: annotation and inversion techniques. Annotation techniques store the metadata of transformations on the data set in some way. Inversion techniques on the other hand keep the data in a way that all modifications can be inverted and the user can access the raw data or middle steps of it.
In [68] the same authors perform a survey on 9 provenance systems. They compare the provenance systems based on why they record provenance, what they describe and how they represent and store provenance information. The publication offers a summary of provenance systems from up to 2005.
Another term to note is the concept of "phantom lineage". Widom 2004 [75] proposes phantom lineage as lineage or provenance about missing or deleted data and as an addition to historical provenance. The same publication also proposes a database system called "Trio". Trio manages accuracy and provenance of Inexact (uncertain, probabilistic, fuzzy, approximate, incomplete, and imprecise) databases. They design a query language as an extension to SQL.
To perform operations on annotations on databases, various publications introduce add-ons to the algebraic structure of those annotations. Senellart 2018 [63] gives a provenance overview from a mathematical point of view and describes different provenance formalisms like boolean provenance and provenance semirings, including provenance in probabilistic databases. The next paragraph focuses on the mathematical modelling of provenance and the underlying add-ons to relational algebra. Up to today the most commonly used mathematical representation of Provenance is the representation of provenance as semirings.
Citing: Romeu 1999 [62], Guptill et al. 2013 [38], Woodruff et al. 1997 [76], Bose 2004 [10], Cui and Widow 2000 [19], Widom 2004 [75]
Cited By: Muniswamy et al.2006 [57], Simmhan et al. 2006 [67], Dai and Lin 2008 [22], Glavic et al.

2007 [33]

Provenance as Semirings Green et al. 2007 [35] [34] introduces the idea to represent provenance annotations as commutative semirings of polynomials.
They analyze various attempts to store provenance as algebraic notations on samples in databases and aim to generalize those techniques. They state a semiring is the most generic algebraic structure for annotations on relations and it can generalize a lot of previously proposed techniques. Commutative semirings are defined as structures $(K, +, *, 0, 1)$ such that $(K, +, 0)$ and $(K, *, 1)$ are commutative monoids. The authors choose $K = N[x]$ the polynomials and show that they are as general as any other possible K. They give the following definition for their Provenance semiring:

> Let X be the set of tuple ids of a (usual) database instance I. The positive algebra provenance semiring for I is the semiring of polynomials with variables (a.k.a. indeterminates) from X and coefficients from N, with the operations defined as usual: $(N[X], +, *, 0, 1)$. [34]

The authors point out that they are talking about polynomials in commutative variables, so their operations are "the same as in middle-school algebra, except that subtraction is not allowed". Besides the mathematical formulation of provenance as semirings, the authors develop a technique to keep query records with Datalog (a database query language in a syntax similar to the logic programming language Prolog). This formulation of semirings is used as a base for most of the recent provenance publications. The following publications extend this concept:
Fink et al. 2012 [31] compiles semiring expressions into so-called decomposition trees for more efficient query evaluations. Geerts et al. 2010 [32] provides further mathematical thoughts on semirings for provenance and extend Green's work to record the provenance of data with different kinds of annotations, while Amsterdamer 2011 [5] shows the limitations of the semiring approach. Senellart et al. 2019 [64] extend the theory of provenance semirings from relational databases to XML, graph, and triple store (knowledge base) databases. Amarilli et al. 2015 [2] present a provenance framework for trees and treelike instances, by describing a linear-time construction of a "circuit provenance representation for monadic second-order logic queries" and connect their work to the semiring definition. **Citing:** Buneman et al. 2001 [14], Cui and Widom 2000 [21] Benjelloun et al. 2005 [7] **Cited By:** Davidson et al. 2008 [23], Cheney et al. 2009 [18], Moreau et al. 2010 [54], Buneman et al. 2007 [15], Deutch 2014 [26], Deutchet al. 2015 [25], Senellart et al. 2018 [63]

Provenance for Specific Domains As explained above, similar to data quality in general, different projects and branches have different criteria for what they need as provenance. In this paragraph, we list some examples of publications, which propose provenance techniques for a given field of applications and try to find a standard provenance technique for their area. Note that the following publications are not an exhaustive list but just a small subset of provenance applications for specific domains:
Myers et al. from 2003 [58] propose a provenance framework for the area of multiscale chemical science. They build a derivation graph for users to inspect. Zhao 2004 [82] describes an attempt to build provenance for bioinformaticians. The authors formalize semantics for provenance logs and a provenance model based on a resource description framework (RDF) database. Bose et al. 2004 [10] propose a provenance model for earth science research on the example of satellite-derived ocean color data. They also try to derive a general provenance model, storing metadata as XML or RDF graphs. Jagadish et al. 2004 [42] uses provenance to estimate data quality and data reliability in the domain of biological data.
Citing: [36], Szomszoret al. 2003 [69], Zhao et al. 2003 [81], Myers 2003 [58] **Cited By:** Simmhan 2005 [66], Bose et al. 2005 [11], Moreau et al. 2010 [54], Buneman et al. 2006 [12], Hasan et al. 2009 [40], Muniswamy et al. 2009 [56]

4 Conclusion

In this section we discuss opportunities to use machine learning to support data quality and provenance techniques. This section starts with a brief overview of machine learning techniques for enhancing data quality. Afterwards, we gather key publications from the area of "truth discovery", which relies on

data quality and provenance and can be formulated as an optimization problem and as such machine learning can be used to solve it. Finally, we give an outlook of how machine learning could be integrated in future provenance research.

Machine Learning for Data Quality It is obvious that data quality plays an important role to enhance the performance of machine learning models. On the other side one can also use machine learning to enhance data quality. Machine learning can be applied in the data collection step, to enhance data quality. Maffettone 2021 [30] applies reinforcement learning to decide which data samples are collected and added to a data set and which not. Machine learning can be applied to modify data sets to enhance their quality, for example through eliminating weak samples in a data set like Wang et al. 2020 [1] or through the optimal chain of preprocessing steps like Learn2Clean [8] using reinforcement learning.
Also, as noted by Eppler et al. 2000 [29] data quality frameworks focus on attributes with linear scoring. Interdependencies between attributes or non-linear dependencies are not modeled by current data quality frameworks. Non-linear functions like neural networks could be trained to estimate a quality score of a data set.

Truth Discovery In settings with different databases the same query might get different answers from different data sources. Information in web data specifically, but also in other data sources, can be unstructured and inaccurate, outdated or simply wrong. Finding out which information is true and which is false is called truth discovery or truth finding. Inconsistencies between data sources are very common and finding out which information is true and which is false is a complicated problem as shown by Li et al. 2015 [51]. Naive techniques use a majority vote: If most databases give the same answer and some give another answer, then the majority is most likely correct. However, there are many cases where this leads to wrong results. There are many attempts to resolve those conflicts between data sources, called data-fusion techniques. Most of them extend the voting scheme with a trustworthiness score for each of the sources. Many extensions also assign a confidence value to the specific returned data values. Li et al. 2015 [51] lists and compares the most prominent of those data-fusion approaches.
One example of those approaches is Truthfinder by Yin et al. 2008 [77]. Here, the trustworthiness of a website is higher if it provides many pieces of true information. When an information is shared by many websites and the websites are trustworthy, then it is more likely that the information is true. Other data-fusion approaches are based on other heuristics to figure out the trustworthiness of a data source, like bayesian measures, web-links or information retrieval techniques. Web-link and information retrieval techniques are good to find truth in categorical data, while bayesian probabilistic models like Zhao et al. 2012 [80] are used for truth discovery with numerical data. Li et al. 2014 [50] add optimizations for databases with different sizes as they claim that small databases should not be trusted too much. This is because the true trustworthiness value is harder to estimate, because of the small sample size. Pasternack 2010[59] proposes an approach to find the truth using prior knowledge.

Machine Learning for Truth Discovery Truth discovery can be formalized as an optimization problem and as such machine learning algorithms can be applied. A truth discovery algorithm optimizes the amount of correct choices between a set of answers from different data sets. As shown above, a common idea is to optimize a trustworthiness score of the data sources and the confidence values of data samples from the sources. Yin et al. 2011 [78] explores semi-supervised learning for finding trustworthiness scores of data sources. They use ground-truth data and compare it to the data of the sources to be able to identify how trustworthy the different data sources are. Their method is based on studies on semi-supervised graph learning like Zhou et al. 2004 [83].
Another idea is to use machine learning algorithms to evaluate the evidence or confidence of data samples. Yu et al. 2014 [79] explore truth discovery with neural networks and knowledge bases. They compute a confidence score for every returned data sample by searching for evidence, using machine learning based language models. After finding an answer to a query in a database the language models search in the database for evidence to support their claim.
Some publications assign scores for different data quality attributes to the data sources. Dong et al. 2009 [28] evaluates coverage, exactness and freshness of their data sources. Optimizing those scores

could be another possible task for machine learning algorithms.

Machine Learning for Provenance Provenance can help to evaluate the credibility of a data source. Also, it can be used to find out if the data sets have the same source for their answer or if data is copied between data sources. Provenance can help to evaluate how trustworthy and how recent a database is.
Some truth discovery publications try to figure out provenance information or more specifically dependencies between data sets directly from the data sources. Dong et al. 2009 [27], [28] develop a hidden markov model that decides whether a source is a copier of another source and identifies the specific moments at which it copies. They combine this model with a bayesian model, which assigns the final truth value. Blanco 2010 [9] proposes another probabilistic approach to deal with this problem. Similarly to those probabilistic models one could use machine learning models to find those dependencies and provenance information between data sets.
Finally, Buneman 2019 [16] gathers many thoughts about the future of provenance research in general but also specifically about the future of provenance on machine learning, mainly focusing on how provenance can help machine learning to get a better interpretability.

Acknowledgement

This research is funded by the National Research Foundation (NRF), Prime Minister's Office, Singapore, and the French National Centre for Scientific Research (CNRS) under the Campus for Research Excellence and Technological Enterprise (CREATE) DesCartes Programme.

References

[1] Weak supervision for fake news detection via reinforcement learning. 34:516–523, Apr. 2020.

[2] Antoine Amarilli, Pierre Bourhis, and Pierre Senellart. Provenance circuits for trees and treelike instances. In *International Colloquium on Automata, Languages, and Programming*, pages 56–68. Springer, 2015.

[3] Antoine Amarilli, Pierre Bourhis, and Pierre Senellart. Tractable lineages on treelike instances: Limits and extensions. In *Proceedings of the 35th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, pages 355–370, 2016.

[4] Yael Amsterdamer, Susan B Davidson, Daniel Deutch, Tova Milo, Julia Stoyanovich, and Val Tannen. Putting lipstick on pig: Enabling database-style workflow provenance. *arXiv preprint arXiv:1201.0231*, 2011.

[5] Yael Amsterdamer, Daniel Deutch, and Val Tannen. On the limitations of provenance for queries with difference. In *TaPP*, 2011.

[6] Yael Amsterdamer, Daniel Deutch, and Val Tannen. Provenance for aggregate queries. In *Proceedings of the thirtieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 153–164, 2011.

[7] Omar Benjelloun, Anish Das Sarma, Alon Halevy, and Jennifer Widom. Uldbs: Databases with uncertainty and lineage. Technical report, Stanford, 2005.

[8] Laure Berti-Equille. Learn2clean: Optimizing the sequence of tasks for web data preparation. 05 2019.

[9] Lorenzo Blanco, Valter Crescenzi, Paolo Merialdo, and Paolo Papotti. Probabilistic models to reconcile complex data from inaccurate data sources. In Barbara Pernici, editor, *Advanced Information Systems Engineering*, pages 83–97, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[10] Rajendra Bose and James Frew. Composing lineage metadata with xml for custom satellite-derived data products. In *Proceedings. 16th International Conference on Scientific and Statistical Database Management, 2004.*, pages 275–284. IEEE, 2004.

[11] Rajendra Bose and James Frew. Lineage retrieval for scientific data processing: a survey. *ACM Computing Surveys (CSUR)*, 37(1):1–28, 2005.

[12] Peter Buneman, Adriane Chapman, and James Cheney. Provenance management in curated databases. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data*, pages 539–550, 2006.

[13] Peter Buneman, Sanjeev Khanna, and Wang-Chiew Tan. Data provenance: Some basic issues. In *International Conference on Foundations of Software Technology and Theoretical Computer Science*, pages 87–93. Springer, 2000.

[14] Peter Buneman, Sanjeev Khanna, and Tan Wang-Chiew. Why and where: A characterization of data provenance. In *International conference on database theory*, pages 316–330. Springer, 2001.

[15] Peter Buneman and Wang-Chiew Tan. Provenance in databases. In *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 1171–1173, 2007.

[16] Peter Buneman and Wang-Chiew Tan. Data provenance: What next? *ACM SIGMOD Record*, 47(3):5–16, 2019.

[17] Li Cai and Yangyong Zhu. The challenges of data quality and data quality assessment in the big data era. *Data science journal*, 14, 2015.

[18] James Cheney, Laura Chiticariu, and Wang-Chiew Tan. *Provenance in databases: Why, how, and where*. Now Publishers Inc, 2009.

[19] Yingwei Cui and Jennifer Widom. Practical lineage tracing in data warehouses. In *Proceedings of 16th International Conference on Data Engineering (Cat. No. 00CB37073)*, pages 367–378. IEEE, 2000.

[20] Yingwei Cui and Jennifer Widom. Lineage tracing for general data warehouse transformations. *the VLDB Journal*, 12(1):41–58, 2003.

[21] Yingwei Cui, Jennifer Widom, and Janet L Wiener. Tracing the lineage of view data in a warehousing environment. *ACM Transactions on Database Systems (TODS)*, 25(2):179–227, 2000.

[22] Chenyun Dai, Dan Lin, Elisa Bertino, and Murat Kantarcioglu. An approach to evaluate data trustworthiness based on data provenance. In *Workshop on Secure Data Management*, pages 82–98. Springer, 2008.

[23] Susan B Davidson and Juliana Freire. Provenance and scientific workflows: challenges and opportunities. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1345–1350, 2008.

[24] Simon De Lusignan, S-T Liaw, Paul Krause, Vasa Curcin, M Tristan Vicente, Georgios Michalakidis, Lars Agreus, Peter Leysen, Nicola Shaw, and Kumara Mendis. Key concepts to assess the readiness of data for international research: Data quality, lineage and provenance, extraction and processing errors, traceability, and curation. *Yearbook of medical informatics*, 20(01):112–120, 2011.

[25] Daniel Deutch, Amir Gilad, and Yuval Moskovitch. Selective provenance for datalog programs using top-k queries. *Proceedings of the VLDB Endowment*, 8(12):1394–1405, 2015.

[26] Daniel Deutch, Tova Milo, Sudeepa Roy, and Val Tannen. Circuits for datalog provenance. In *ICDT*, volume 3, page 2014. Citeseer, 2014.

[27] Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. Integrating conflicting data: the role of source dependence. *Proceedings of the VLDB Endowment*, 2(1):550–561, 2009.

[28] Xin Luna Dong, Laure Berti-Equille, and Divesh Srivastava. Truth discovery and copying detection in a dynamic world. *Proc. VLDB Endow.*, 2(1):562–573, August 2009.

[29] Martin J Eppler and Dörte Wittig. Conceptualizing information quality: A review of information quality frameworks from the last ten years. *IQ*, 20(0):0, 2000.

[30] Phillip M Maffettone et al. Gaming the beamlines—employing reinforcement learning to maximize scientific outcomes at large-scale user facilities. *Mach. Learn.: Sci. Technol.*, 2(025025), 2021.

[31] Robert Fink, Larisa Han, and Dan Olteanu. Aggregation in probabilistic databases via knowledge compilation. *arXiv preprint arXiv:1201.6569*, 2012.

[32] Floris Geerts and Antonella Poggi. On database query languages for k-relations. *Journal of Applied Logic*, 8(2):173–185, 2010.

[33] Boris Glavic, Klaus R Dittrich, A Kemper, H Schöning, T Rose, M Jarke, T Seidl, C Quix, and C Brochhaus. Data provenance: A cctegorization of existing approaches. *BTW'07: Datenbanksysteme in Buisness, Technologie und Web*, (103):227–241, 2007.

[34] Todd J Green, Grigoris Karvounarakis, Zachary G Ives, and Val Tannen. Update exchange with mappings and provenance. 2007.

[35] Todd J Green, Grigoris Karvounarakis, and Val Tannen. Provenance semirings. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 31–40, 2007.

[36] Mark Greenwood, Carole Goble, Robert Stevens, Jun Zhao, Matthew Addis, Darren Marvin, Luc Moreau, and Tom Oinn. Provenance of e-science experiments-experience from bioinformatics. 2003.

[37] Paul Groth, Michael Luck, and Luc Moreau. A protocol for recording provenance in service-oriented grids. In *International Conference on Principles of Distributed Systems*, pages 124–139. Springer, 2004.

[38] Stephen C Guptill and Joel L Morrison. *Elements of spatial data quality*. Elsevier, 2013.

[39] Olaf Hartig. Provenance information in the web of data. In *LDOW*, 2009.

[40] Ragib Hasan, Radu Sion, and Marianne Winslett. Preventing history forgery with secure provenance. *ACM Transactions on Storage (TOS)*, 5(4):1–43, 2009.

[41] Zachary G Ives, Todd J Green, Grigoris Karvounarakis, Nicholas E Taylor, Val Tannen, Partha Pratim Talukdar, Marie Jacob, and Fernando Pereira. The orchestra collaborative data sharing system. *ACM Sigmod Record*, 37(3):26–32, 2008.

[42] H. V. Jagadish and Frank Olken. Database management for life sciences research. *SIGMOD Rec.*, 33(2):15–20, June 2004.

[43] Grigoris Karvounarakis and Todd J Green. Semiring-annotated data: queries and provenance? *ACM SIGMOD Record*, 41(3):5–14, 2012.

[44] Avita Katal, Mohammad Wazid, and Rayan H Goudar. Big data: issues, challenges, tools and good practices. In *2013 Sixth international conference on contemporary computing (IC3)*, pages 404–409. IEEE, 2013.

[45] Barbara Klein. When do users detect information quality problems on the world wide web? 2002.

[46] David P Lanter. Design of a lineage-based meta-data base for gis. *Cartography and Geographic Information Systems*, 18(4):255–261, 1991.

[47] Thomas Lee, Stéphane Bressan, and Stuart E Madnick. Source attribution for querying against semi-structured documents. In *Workshop on Web Information and Data Management*, pages 33–39. Citeseer, 1998.

[48] Thomas Lee, Melanie Chams, Robert Nado, Michael Siegel, and Stuart Madnick. Information integration with attribution support for corporate profiles. In *Proceedings of the eighth international conference on Information and knowledge management*, pages 423–429, 1999.

[49] Yang W Lee, Leo Pipino, James D Funk, and Richard Y Wang. *Journey to data quality*. MIT press Cambridge, 2006.

[50] Qi Li, Yaliang Li, Jing Gao, Lu Su, Bo Zhao, Murat Demirbas, Wei Fan, and Jiawei Han. A confidence-aware approach for truth discovery on long-tail data. *Proc. VLDB Endow.*, 8(4):425–436, December 2014.

[51] Xian Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, and Divesh Srivastava. Truth finding on the deep web: Is the problem solved?, 2015.

[52] Alexandra Meliou, Wolfgang Gatterbauer, Katherine F Moore, and Dan Suciu. The complexity of causality and responsibility for query answers and non-answers. *arXiv preprint arXiv:1009.2021*, 2010.

[53] Paolo Missier, Khalid Belhajjame, and James Cheney. The w3c prov family of specifications for modelling provenance metadata. In *Proceedings of the 16th International Conference on Extending Database Technology*, pages 773–776, 2013.

[54] Luc Moreau. *The foundations for provenance on the web*. Now Publishers Inc, 2010.

[55] Luc Moreau, Ben Clifford, Juliana Freire, Joe Futrelle, Yolanda Gil, Paul Groth, Natalia Kwasnikowska, Simon Miles, Paolo Missier, Jim Myers, et al. The open provenance model core specification (v1. 1). *Future generation computer systems*, 27(6):743–756, 2011.

[56] Kiran-Kumar Muniswamy-Reddy, Uri Jacob Braun, David A Holland, Peter Macko, Diana Maclean, Daniel Wyatt Margo, Margo I Seltzer, and Robin Smogor. Layering in provenance systems. In *Proceedings of the 2009 USENIX Annual Technical Conference (USENIX'09)*. USENIX Association, 2009.

[57] Kiran-Kumar Muniswamy-Reddy, David A Holland, Uri Braun, and Margo I Seltzer. Provenance-aware storage systems. In *Usenix annual technical conference, general track*, pages 43–56, 2006.

[58] James D Myers, Carmen M Pancerella, Carina S Lansing, Schuchardt Karen L, Didier Brett T, and C Ashish, N. Goble. Multi-scale science: Supporting emerging practice with semantically derived provenance. 10 2003.

[59] Jeff Pasternack and Dan Roth. Knowing what to believe (when you already know something). In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 877–885, 2010.

[60] Leo L. Pipino, Yang W. Lee, and Richard Y. Wang. Data quality assessment. *Commun. ACM*, 45(4):211–218, April 2002.

[61] Yann Ramusat, Silviu Maniu, and Pierre Senellart. Semiring provenance over graph databases. In *10th {USENIX} Workshop on the Theory and Practice of Provenance (TaPP 2018)*, 2018.

[62] Jorge Luis Romeu. Data quality and pedigree. *Material Ease*, 1999.

[63] Pierre Senellart. Provenance and probabilities in relational databases. *ACM SIGMOD Record*, 46(4):5–15, 2018.

[64] Pierre Senellart. Provenance in databases: Principles and applications. In *Reasoning Web. Explainable Artificial Intelligence*, pages 104–109. Springer, 2019.

[65] Pierre Senellart, Louis Jachiet, Silviu Maniu, and Yann Ramusat. Provsql: Provenance and probability management in postgresql. *Proceedings of the VLDB Endowment (PVLDB)*, 11(12):2034–2037, 2018.

[66] Yogesh L. Simmhan, Beth Plale, and Dennis Gannon. A survey of data provenance in e-science. *SIGMOD Rec.*, 34(3):31–36, September 2005.

[67] Yogesh L Simmhan, Beth Plale, and Dennis Gannon. A framework for collecting provenance in data-centric scientific workflows. In *2006 IEEE International Conference on Web Services (ICWS'06)*, pages 427–436. IEEE, 2006.

[68] Yogesh L Simmhan, Beth Plale, Dennis Gannon, et al. A survey of data provenance techniques. *Computer Science Department, Indiana University, Bloomington IN*, 47405:69, 2005.

[69] Martin Szomszor and Luc Moreau. Recording and reasoning over data provenance in web and grid services. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 603–620. Springer, 2003.

[70] Wang Chiew Tan. Research problems in data provenance. *IEEE Data Eng. Bull.*, 27(4):45–52, 2004.

[71] Howard Veregin. Data quality parameters. *Geographical information systems*, 1:177–189, 1999.

[72] Richard Y Wang and Diane M Strong. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems*, 12(4):5–33, 1996.

[73] Richard Y Wang, Mostapha Ziad, and Yang W Lee. *Data quality*, volume 23. Springer Science & Business Media, 2006.

[74] Y Richard Wang, Stuart E Madnick, et al. A polygen model for heterogeneous database systems: The source tagging perspective. 1990.

[75] Jennifer Widom. Trio: A system for integrated management of data, accuracy, and lineage. Technical report, Stanford InfoLab, 2004.

[76] Allison Woodruff and Michael Stonebraker. Supporting fine-grained data lineage in a database visualization environment. In *Proceedings 13th International Conference on Data Engineering*, pages 91–102. IEEE, 1997.

[77] Xiaoxin Yin, Jiawei Han, and Philip S. Yu. Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering*, 20(6):796–808, 2008.

[78] Xiaoxin Yin and Wenzhao Tan. Semi-supervised truth discovery. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, page 217–226, New York, NY, USA, 2011. Association for Computing Machinery.

[79] Dian Yu, Hongzhao Huang, Taylor Cassidy, Heng Ji, Chi Wang, Shi Zhi, Jiawei Han, Clare Voss, and Malik Magdon-Ismail. The wisdom of minority: Unsupervised slot filling validation based on multi-dimensional truth-finding. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1567–1578, 2014.

[80] Bo Zhao and Jiawei Han. A probabilistic model for estimating real-valued truth from conflicting sources. *Proc. of QDB*, 1817, 2012.

[81] Jun Zhao, Carole Goble, Mark Greenwood, Chris Wroe, and Robert Stevens. Annotating, linking and browsing provenance logs for e-science. In *Proc. of the Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data*, volume 176. Citeseer, 2003.

[82] Jun Zhao, Carole Goble, Robert Stevens, and Sean Bechhofer. Semantically linking and browsing provenance logs for e-science. In Mokrane Bouzeghoub, Carole Goble, Vipul Kashyap, and Stefano Spaccapietra, editors, *Semantics of a Networked World. Semantics for Grid Databases*, pages 158–176, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[83] Dengyong Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in neural information processing systems*, pages 321–328, 2004.