Challenges for Value-driven Semantic Data Quality Management

Rob Brennan
Knowledge and Data Engineering Group, ADAPT Centre, School of Computer Science and Statistics, Trinity College Dublin, College Green, Dublin 2, Ireland

Keywords: Data Quality, Semantics, Linked Data, Data Value, Data Management.

Abstract: This paper reflects on six years developing semantic data quality tools and curation systems for both large-scale social sciences data collection and a major web of data hub. This experience has led the author to believe in using organisational value as a mechanism for automation of data quality management to deal with Big Data volumes and variety. However there are many challenges in developing these automated systems and this discussion paper sets out a set of challenges with respect to the current state of the art and identifies a number of potential avenues for researchers to tackle these challenges.

1 INTRODUCTION

Data governance aligns organisational goals with the management of the data assets that data driven enterprises need to leverage (Brous, 2016). Data governance thus provides oversight and goal-setting for data quality management since creating and maintaining "appropriate" or "fit for use" (i.e. high quality) data for users is an organisational imperative (Logan, 2016). To date, bridging the gap between business priorities and data quality management has been hampered by factors such as: the focus of data quality tools on low-level intrinsic quality measures (e.g. syntactic validity) (Zaveri, 2015), the multi-faceted nature of data quality (Radulovic, 2016), an increasingly diverse data technology ecosystem built on siloed tool-chains (Khatri, 2010), and the lack of business-oriented metadata (Schork, 2009). Automated support for data stakeholder or steward governance of data quality is immature due to the lack of standardisation of integration points for big data quality control systems (ISO, 2014). In contrast, for low level data quality metrics, recent advances in semantic data quality analysis show great promise (Bertossi, 2013, Feeney, 2016, Kontokostas, 2014) but methods have not yet emerged to apply them to traditional databases and semi-structured web data, where most data growth is centred. Current work on dataset meta-data standards by the W3C (Maali, 2014) would be a natural basis for business-oriented quality metadata.

Gartner in 2016 have urged for a more business-led data governance discipline, highlighting that "it's all about the business value". Moody previously observed that "100% accurate information is rarely required in a business context" (Moody, 1999) so it is impractical (and unprofitable) to blindly try to achieve "high quality data" across the board (Evan, 2010). Investment in data quality can be seen as insurance against the risk that your data is not "fit for use". In 2013 Tallon spelt out the challenge "Finding data governance practices that maintain a balance between value creation and risk exposure is the new organizational imperative" (Tallon, 2013). Nonetheless few technologists have taken up this challenge (Brous, 2016, Yousif, 2015). In part this is due to a lack of consensus on mathematical models for the estimation of business value (Viscusi, 2014). Using value estimates to drive business processes is common but automated data quality management toolchains based on value is limited to spot cases such as quality assessment (Evan, 2005), file retention management (Wijnhoven, 2014) and data lifecycle management (Chen, 2005).

This paper presents a survey of recent developments relevant to developing a new generation of automated data quality management systems that are capable of dealing with Big Data volumes and variety in a way that minimises costs by using models of the organisational value of data linked to data quality (section 2). A set of three research challenges for value-driven data quality management are then identified along with potential directions for research that will satisfy these

Brennan, R.
Challenges for Value-driven Semantic Data Quality Management.
DOI: 10.5220/0006387803850392
In *Proceedings of the 19th International Conference on Enterprise Information Systems (ICEIS 2017) - Volume 1*, pages 385-392
ISBN: 978-989-758-247-9

challenges (section 3). Finally in section 4 some thoughts are presented on the current outlook for solutions in this space.

2 BACKGROUND EXPERIENCES AND RELATED WORK

The challenges identified in this paper come from the author's experiences developing semantic data quality tools (Feeney, 2014, Feeney, 2017) and applying them to large, international data collection efforts like large, international social science datasets (Brennan, 2016, Turchin, 2015) or major linked data hubs (Meehan, 2016). Interactions with dataset stakeholders over a number of years have suggested that despite the advantages of semantic data quality approaches, e.g. more expressive schema (Mendel-Gleason, 2015), that the intrinsic quality metrics that the majority of such tools focus on (Zaveri, 2015) are not the locus of business or organisational value in the datasets (Evan, 2005). Even the best data quality processes and tools require human oversight to be most effective (Brous, 2016) and as the number and variety of datasets increases (especially in a world dominated by Big Data) it is not scaleable to try and improve data quality uniformly (Evan, 2010). Instead some means must be developed to focus the attention of automated tools on the places where they can do the most good with the least investment of effort. In a holistic approach that goes beyond the intrinsic or universal quality measures that the stakeholders have already rejected, this naturally leads to a desire for unlocking the greatest value in an organisation's data assets. This poses the questions of how can data value be located and estimated.

Unfortunately, (a) defining what is "appropriate data" at a business level is both a hard problem and not adequately addressed by current approaches; and (b) many dimensions of data value are expressed in extrinsic data quality measures that depend on metadata, provenance or usage information that do not exist within the dataset itself. This leads to the conjecture that bottom-up data quality tools that focus on dataset-centric intrinsic quality metrics such as consistency or integrity will improve over time but that this progress is only incremental and a step-change in the effectiveness of data quality governance requires new methods to driect and monitor data quality methods and tools based on the organisational value of data.

Given the explosion in data we are witnessing, current approaches will not scale to meet the demand for data that is fit for use. Even with some progress on tools extrinsic measures like availability of licensing information (Neumaier, 2016), the real gains in application of data quality tools will be at the interface between addressing business needs (Schork, 2009), supporting domain experts rather than information architects (Mosley, 2010) and methods to focus on the available tools and people on the most relevant data quality issues rather than wasting effort on uniform metric improvements that might not even feed into business goals (Evan, 2010). There is a direct parallel between this situation and the author's track record on bridging the gaps in human involvement in semantic mapping processes (Conroy, 2009, Debruyne, 2013) in contrast to the majority of the research in ontology matching which focuses on improving low-level matching algorithms (Shvaiko, 2013). Another important influence on the challenges identified for data quality governance is recent work on semantic mapping lifecycle governance that uses W3C PROV as an underlying basis to capture human decision-making in a machine-readable way (Debruyne, 2015).

Previous work on metadata and business value as drivers for data quality has focused on pre-semantic technology for data warehousing (Helfert, 2002, Shankaranarayan, 2003), organisational decision support systems (Evan, 2005, Evan, 2010, Schork, 2009, Tallon, 2007, Viscusi, 2014) as opposed to our challenges for tool automation. There are however related active research topics like file-retention strategies based on file metadata (Wijnhoven, 2014) and autonomic data lifecycle management (Chen, 2005).

A survey of the literature on data governance, management and lifecycles finds that while data quality is widely regarded as critical (Brous, 2016, Khatri, 2010, Tallon, 2013, Weber, 2009), that most current processes are human rather than machine oriented (Aiken, 2016, Mosley, 2010). In part this may be because there are a wide variety of data lifecycle models but no clear standards (ISO, 2014). This is influenced by the diversity of data storage and structuring technologies currently in vogue, from traditional RDBMS to NoSQL, linked data and data warehouses to data lakes. Nonetheless the need for more automated data quality management is manifest and key to this is how goals are set for these systems to enable planning, monitoring and enforcement (Logan, 2016). Underpinning any automated decision-making will be rich data quality criteria and the oversight of domain experts such as data stewards. These quality criteria must capture and even predict the links between data assets, processes, tools, users and data value.

3 THE CHALLENGES

The overall vision of these challenges is to work towards business-driven, automated semantic data quality management directed by data value estimates that becomes more effective over time. Semantics are at the heart of our approach since they provide (a) a formal specification model (b) the basis for rich data quality methods and tools and (c) an effective data quality integration and interchange technology when instantiated as enterprise Linked Data.

Our vision first requires the development by the community of a deeper understanding of how to define data asset value in a generic and formal way. Then it will be possible to specify value-driven data quality criteria, i.e. ways to express the links between data value, metadata, processes and data quality methods or tools. Then, this knowledge can be codified in machine-processable formal models that support semantic data interchange about data quality and value in the organisation. This in turn will enable the development of improved predictive models and intelligent adaptive data quality systems for value-driven data quality in digital enterprises. In a heterogeneous environment the semantic data interchange models could provide the basis of multi-vendor interoperability and tool-chain integration. Each challenge is discussed in a sub-section below, along with potential approaches for addressing the challenge.

3.1 Mature Models of Data Value

Although there is a lot of discussion about "data value", "information asset valuation", "data as an asset" and infonomics at the moment, most of this discussion is industry-led and does not focus on formal models of the difficult topic of exactly how information should be valued. Moody and Walsh defined seven "laws" of information that explained its unique behaviour and relation to business value (Moody and Walsh, 1999) but even that work does not define the concrete measurement techniques or metrics. Moody identifies three methods of data valuation – utility, market price and cost (of collection) – and concludes that utility is in theory best but impractical and thus cost-based estimation is the most effective method. Unfortunately sunk cost is not a strong candidate for directing future quality management activities in an agile environment. Most research on information value merely seeks to identify dimensions or characteristics without defining a mathematical theory of data value.

One guiding research question should be: *What are the fundamental processes and attributes driving changes in data value over time, and how does this give rise to patterns of data value development and diffusion*? This would lead to a formal model with strong explanatory or predictive properties. More importantly it could act as a baseline for future research in this under-specified area.

3.1.1 Potential Approaches

There has been no work to date on formal knowledge models of the data value domain which limits the application of intelligent systems to data value management or profiling. The ultimate goal of such models would be predicting value as well as assessing it but so far this has proved very context-dependent (e.g. the value of a specific dataset is a function of current business goals) and thus hard to formulate general models. However it is likely that, as with data quality, there are both extrinsic and intrinsic measures of value and by calculating the intrinsic measures it may be possible to estimate the extrinsic proportion of value

Another area that must be addressed is long-term validation of models against known data and business lifecycles. It is possible that usage-based models of data value, such as already deployed for file management (Wijnhoven 2014) could be applied to data assets as a whole. This also corresponds to the concept of economic value in usage. Such usage-based models may be based on system logs or provenance information.

3.2 Linking Data Quality to Data Value

This challenge is about defining value-based data quality criteria that enable unified quality governance of datasets and systems. Formal models of data quality criteria that enable us to better understand, evaluate and predict the links between dataset production costs, utility, usage patterns, quality metrics, metadata, topic domains, workflows, provenance, and value would enable new value-driven approaches to data quality governance and new insights into the location of data value within an organisation. These machine-processable models would form the basis for sharing knowledge about data quality and value throughout the data quality ecosystem and enable automation of quality management tasks such as data quality metric selection, tool selection and orchestration, process or workflow configuration and quality task pritorisation, These models would also support new intelligent data quality applications such as data asset value profiling or improvement, and decision support for data quality process design.

Many of the data value dimensions identified in the literature overlap with data quality dimensions. For example Ahituv (Ahituv, 1989) suggests: timeliness (dimensions: recency, response time, and frequency), contents (dimensions: accuracy, relevance, level of aggregation and exhaustiveness), format (dimensions: media, color, structure, presentation), and cost. Compare these with Zalveri et al's recent survey of Linked Data quality dimensions: accessibility, contextual, dataset dynamicity, intrinsic, representation and trust (Zalveri 2015).

3.2.1 Potential Approaches

Linked Data could be used as a unifying technical foundation for quality criteria specification, dataset description (via metadata), dataset usage logs and provenance, and process or tool integration (through the specification of exchange formats and lightweight REST-based interfaces).

This parallels the work being carried out at the W3C on dataset-level metadata and data quality metric vocabularies and then reused within the H2020 ALIGNED project to describe the combined software and data engineering of data-intensive systems. By creating standardised, reusable semantic specifications it is possible to build models of a domain (such as data value) and relate it to component models describing, for example, data quality, data lifecycles, business context and governance roles or processes. Then each sub-model becomes a basis for data collection and exchange about a specific dimension of data value, e.g. usage patterns. The upper or combined model then becomes the basis for data fusion to determine overall value.

3.3 Methods to Apply Semantic Data Quality Tools to Heterogeneous Data

Effective data driven enterprises require data that is "fit for use" and must employ active data quality management of mixed data ecosystems (e.g. relational databases, linked data and semi-structured data like csv, json and xml) while linking quality actions to business value. Despite recent progress and the emergence of both commercial and research tools (Zaveri, 2015), semantic data quality tool researchers must address the fact that the majority of the world's data is not stored in RDF graphs. This limits the applicability and impact of their tools and methods.

3.3.1 Potential Approaches

Support for dealing with the diversity of real world data infrastructure could be provided by four management capabilities: unified dataset metadata agents, ontology-based data access for quality tools, and a unified PROV-based log service. Together these semantic approaches could use the power of RDF-based data to span multiple local schemata, provide formal models of semi-structured data mappings (R2RML-F), reuse rich metadata specifications and unified data access for multiple storage technologies. However RDF would only be required at the data quality management, metadata and semi-structured data mapping integration points – existing access to data silos based on end-user applications or technologies would be unaffected. This is important to both: (1) be able to deploy these solutions for real-world data sources and (2) to ensure that the systems are flexible enough to cope with diverse data ecosystems rather than being tailored to an idealised "green-field" deployment of semantic web technology. This approach follows the W3C's Data Activity which envisages complimentary, connecting pipelines of diverse data formats and technologies to provide information services.

3.4 Automated Techniques for Value-Driven Data Quality Management

Satisfying these requirements in the time of the Big Data deluge requires a shift away from human-centric processes and requires us to develop new automated techniques for data quality monitoring, analysis, and enforcement that assure business value while minimising human effort. This especially applies to data quality where the rationale that human oversight is required for the highest quality data processes and the limited capabilities of many traditional data quality tools leads to heavy use of manual effort in the data quality domain.

3.4.1 Potential Approaches

Automated quality management requires making and implementing data quality decisions about data assets, e.g. selecting quality metrics or processes, generating quality reports, orchestrating quality processes or tools (Khatri 2010). This can use machine reasoning, inference or statistical approaches based on leveraging knowledge models of: the data quality domain (Radulovic, 2016), threats to data quality and catalogues of best practice (Foley, 2011), how data value can be expressed in dataset metadata (Helfert, 2002, Wijnhoven, 2014) and extrinsic data quality metrics (Viscusi, 2014). Hence the presence of formal models linking data quality and value enable automated decision-making or decision support for data owners such as Data Stewards.

Specific technology performance curves could be developed for data quality processes and tools that, in conjunction with the knowledge models, support data quality planning and prediction. Success will be measured by statistical analysis of the model's predictions using historical studies of data quality (hindcasting).

By building on Chen's work on value-based autonomic data lifecycle management (Chen, 2005) and Even et al.'s approach to value-based data quality management for organisational decision support (Evan, 2010), it may be possible to support automated selection, prioritisation and orchestration of data quality tools.

3.5 Standardised Data Quality Management Architecture

A common architecture and standardised integration reference points will make the latest advances in semantic data quality tools available to traditional databases and semi-structured web data via ontology-based data access (challenge 3), common metadata standards and semantic mappings. The reference architecture should enable a semi-supervised data quality control loop (challenge 4). The behaviour of the control loop goals would be set by data quality criteria expressed in terms of a governance model that links value and data quality (challenge 2). This will address the gap in the state in the art whereby most research looks at individual quality tools outside of their deployment context and without reference to any business context.

3.5.1 Potential Approaches

These integration reference points could bring together disparate semantic quality reporting, metadata and data quality vocabularies into a coherent whole, enabling multi-vendor solutions. This would extend the deployment scope of semantic data quality tools by defining a new approach to quality-centric ontology-based data access (OBDA), where to date only consistency measures have been addressed [Console14]. New mechanisms for data quality management of semantic mappings and semi-structured data could be developed to allow semantic quality approaches to be applied to semi-structured data.

An exemplar architecture for automated, value-driven semantic data quality management is sketched in figure 1 below. Support for dealing with the diversity of real world data infrastructure will be provided by four management capabilities: unified dataset metadata agents, unified ontology-based data access for quality tools, and a unified PROV-based log service. Together these semantic approaches use the power of RDF-based data to span multiple local schemata, provide formal models of semi-structured data mappings (R2RML-F), reuse rich metadata specifications and unified data access for multiple storage technologies. However RDF is only required at the data quality management, metadata and semi-structured data mapping integration reference points – existing access to data silos based on end-user applications or technologies are unaffected. This is important to both: (1) be able to deploy solutions for real-world data sources and (2) to ensure that the integration reference point and data quality criteria specifications are flexible enough to cope with diverse data ecosystems rather than being tailored to an idealised "green-field" deployment of semantic web technology.

The architecture components shown in figure 1 are as follows:

Automated Data Quality System: this will monitor, analyse and enforce data quality within the data quality management system based on the value-driven autonomic data lifecycle approach of Chen [Chen05].

Semantic Data Quality Tools from the state of the art such as TCD's Dacura Quality Service for OWL-based validation (Feeney, 2017) and AKSW's RDFUnit tool for SPARQL and SHACL-based data unit testing (Kontokostas, 2014).

Data Access for Quality Management: Access to RDBMS for metadata agents and semantic data quality tools would be based on the mature and highly performant ontology-based data access platforms such as OnTop (Calvanese et al, 2016).

Dataset Log Agents will convert, create and maintain dataset usage and governance information using the W3C's PROV standard.

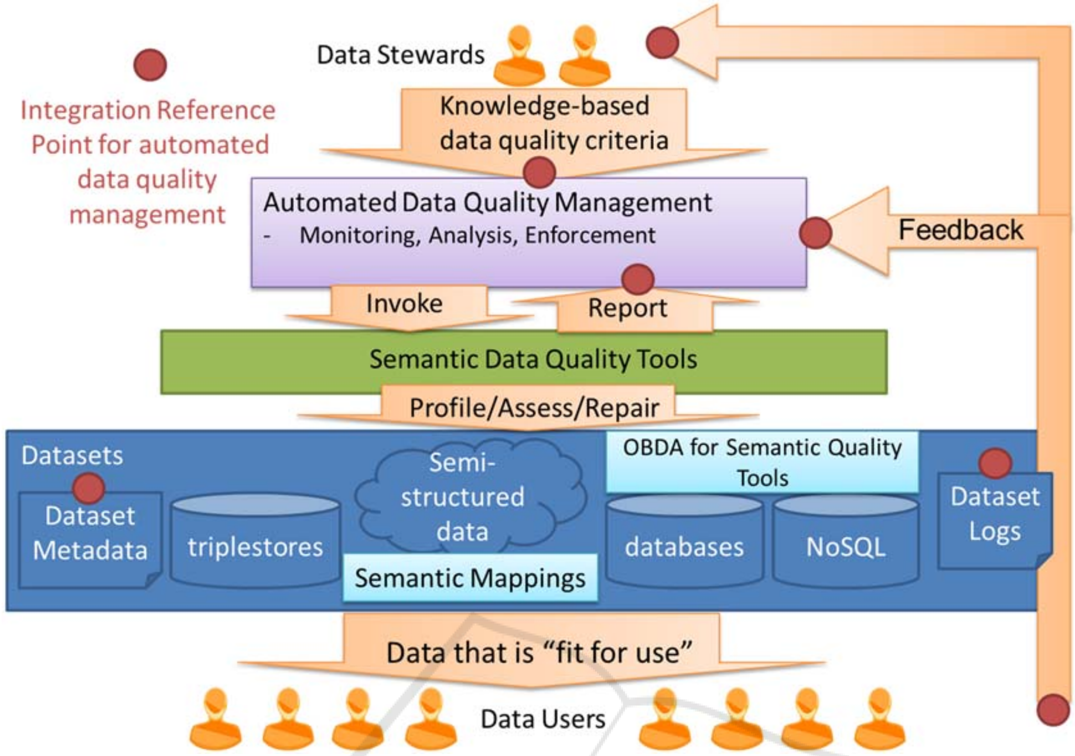

Figure 1: Exemplar Automated Value-Driven Semantic Data Management Architecture.

Dataset Metadata Agents could populate and maintain DataValue extensions of emerging metadata standards like DataID and capture metadata fields relevant to calculating data value such as key entities and dataset provenance.

4 SUMMARY AND OUTLOOK

This paper discussed significant challenges facing data quality researchers and the big data industry as it aims to tackle the dual goals of controlling data quality costs and engineering those systems to be flexible enough to support agile management decision making by data service providers and their customers in the face of the increased scalability demands of Big Data systems.

Most of the development of value-driven systems is currently outside of computer science or informatics academic research and is led by industry specialists such as Doug Laney of Gartner. There is a parallel thread of academic research on knowledge management and organisational impact that emerges from the business schools or economics departments. However these three threads must come together if we are to engineer value-driven systems. This requires bridging the gap between human understanding of business needs and low-level data lifecycle tools. Hence semantics or formal knowledge models are ideally placed to play a significant role in future systems. This compliments the W3C's standardisation role on knowledge-based, data-centric systems.

ACKNOWLEDGEMENTS

This research has received funding from the ADAPT Centre for Digital Content Technology, funded under the SFI Research Centres Programme (Grant 13/RC/2106) and co-funded by the European Regional Development Fund.

The author also wants to thank the reviewers for their many suggestions for improving the final version of this paper.

REFERENCES

Ahituv, N., 1989, Assessing the value of information: problems and approaches. In: DeGross, J.I.,Henderson, J.C., Konsynski, B.R. (eds.) International Conference on Information Systems (ICIS 1989). pp. 315–325. Boston, Massachusetts.

Aiken, P., 2016, EXPERIENCE: Succeeding at Data Management—BigCo Attempts to Leverage Data, Journal of Data and Information Quality (JDIQ), Volume 7 Issue 1.

Bertossi, L. and Bravo, L., 2013, Generic and Declarative Approaches to Data Cleaning: Some Recent Developments, Handbook of Data Quality: Research and Practice, Shazia Sadiq (Ed), Spriner, ISBN 978-3-642-36256-9, 2013.

Brennan, R. , Feeney, K. Mendel-Gleason, G. Bozic, B. Turchin, P. Whitehouse, H. Francois, P. Currie, T. E. Grohmann, S., 2011, Building the Seshat Ontology for a Global History Databank, LNCS , Extended Semantic Web Conference, Heraklion, 29th May - 2nd June, edited by Harald Sack - Eva Blomqvist - Mathieu d'Aquin - Chiara Ghidini - Simone Paolo Ponzetto - Christoph Lange , (9678), Springer, 2016, pp693 – 708.

Calvanese, D., Cogrel, B. , Komla-Ebri, S. Kontchakov, R. Lanti, D. Rezk, M. Rodriguez-Muro, M. Xiao, G., 2016, Ontop: Answering SPARQL Queries over Relational Databases, (Accepted), Semantic Web Journal, Available at: http://www.semantic-web-journal.net/content/ontop-answering-sparql-queries-over-relational-databases-1 [Accessed 09 March 2017]

Chen, Y., 2005, Information valuation for Information Lifecycle Management, Proc IEEE International Conference on Autonomic Computing pp: 135-146, DOI 10.1109/ICAC.2005.35.

Console, M. and Lenzerini, M., 2014, Data Quality in Ontology-Based Data Access:The Case of Consistency, Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence.

Debruyne, C., Walshe, B., O'Sullivan, D., 2015, Towards a Project Centric Metadata Model and Lifecycle for Ontology Mapping Governance, 17th International Conference on Information Integration and Web-based Applications & Services, Brussels, Belgium, 11-13 December , edited by Maria Indrawan-Santiago, Matthias Steinbauer, A Min Tjoa, Ismail Khalil, Gabriele Anderst-Kotsis , ACM, pp356 – 365.

Vander Sande, M., Colpaert, P., Verborgh, R., Mannens, E., and Van De Walle, R., 2014 , RML : A Generic Language for Integrated RDF Mappings of Heterogeneous Data.

Even, A., and Shankaranarayanan, G., 2005, Value-Driven Data Quality Assessment, in Proceedings of the 10th International Conference on Information Quality, Cambridge, MA, USA.

Evan, A., 2010, Evaluating a model for cost-effective data quality management in a real-world CRM setting, Decision Support Systems 50(1):152-163, DOI: 10.1016/j.dss.2010.07.011.

Feeney, K., O'Sullivan, D., Tai, W. and Brennan, R. 2014, Improving curated web-data quality with structured harvesting and assessment, International Journal on Semantic Web and Information Systems 10(2).

Feeney, K. Mendel-Gleason G. and Brennan, R., 2017, Linked data schemata: fixing unsound foundations, Semantic Web Journal, Accepted, Available at: http://www.semantic-web-journal.net/content/linked-data-schemata-fixing-unsound-foundations-0 [Accessed 09 March 2017]

Foley, S., and Fitzgerald, W., 2011, Management of security policy configuration using a Semantic Threat Graph approach, Journal of Computer Security, vol. 19, no. 3, pp. 567-605.

Helfert, M., and Herrmann, C., 2002, Proactive Data Quality Management for Data Warehouse Systems - A Metadata based Data Quality System. In: 4th International Workshop on Design and Management of Data Warehouses (DMDW), Toronto, Canada.

ISO/IEC JTC 1, 2014, Information Technology, Big Data, Preliminary Report 2014. Available at: http://www.iso.org/iso/big_data_report-jtc1.pdf [Accessed 09 March 2017]

Janssen, M. and Vilminko-Heikkinen, R., 2016, Coordinating Decision-Making in Data Management Activities: A Systematic Review of Data Governance Principles, EGOV 2016, LNCS 9820, pp. 115–125.DOI: 10.1007/978-3-319-44421-5_9.

Khatri, V., Brown, C.V., 2010, Designing data governance. Communications of the ACM 53(1), 148–152.

Kontokostas, D. , Westphal, P. , Auer, S., Hellmann, S., Lehmann, J. , Cornelissen, r., Zaveri, a., 2014, Test-driven Evaluation of Linked Data Quality, Proceedings of the 23rd International Conference on World Wide Web, pp747-758.

Logan, D. 2016, Ten Steps to Information Governance, Gartner Report G00296492. Available at: https://www.edq.com/resources/analyst-reports/gartner-10-steps-to-data-governance/ [Accessed 09 March 2017]

Maali, F., Erickson, J., (Eds.), 2014, Data Catalog Vocabulary (DCAT), W3C Recommendation, Available at https://www.w3.org/TR/vocab-dcat/

Meehan, A. Kontokostas, D. Freudenberg, M. Brennan, R. O'Sullivan, D., 2016, Validating Interlinks between Linked Data Datasets with the SUMMR Methodology, ODBASE 2016 - The 15th International Conference on Ontologies, DataBases, and Applications of Semantics, Rhodes, Greece, , Springer Verlag.

Mendel-Gleason, G. Feeney K. and Brennan, R., 2015, Ontology Consistency and Instance Checking for Real World Linked Data, 2nd Workshop on Linked Data Quality, Slovenia.

Moody, D. and Walsh, P., 1999, Measuring The Value Of Information: An Asset Valuation Approach, Proc. Seventh European Conference on Information Systems (ECIS'99).

Mosley, M., Brackett, M.,, Earley, S., and Henderson, D., (Eds), 2010, The DMA Guide to the Data Management Body of Knowledge (1st Ed.), Technics Publications LLC, 2010, ISBN 978-9355040-2-3.

Neumaier, S., Umbrich, J., Polleres, A., 2016, Automated Quality Assessment of Metadata across Open Data Portals, Journal of Data and Information Quality (JDIQ), Vol. 8 Issue 1, DOI:10.1145/2964909.

Schork, R., 2009, Integrating Business and Technical Metadata to Support Better Data Quality, MIT Information Quality Industry Symposium. Available at at http://mitiq.mit.edu/IQIS/Documents/CDOIQS_200977/Papers/02_02_1C-2.pdf [Accessed 09 March 2017]

Solanki, M., Bozic, B., Freudenberg, M., Kontokostas, D., Dirschl C., and Brennan,R., 2016, Enabling combined Software and Data engineering at Web-scale: The ALIGNED suite of Ontologies, ISWC 2016: The Semantic Web – ISWC 2016 pp 195-203.

Radulovic, F., Mihindukulasooriya, N., García-Castro, R. and Gómez-Pérez, A., 2016, A comprehensive quality model for Linked Data, Accepted, Semantic Web Journal, Available at: http://www.semantic-web-journal.net/system/files/swj1488.pdf [Accessed 09 March 2017]

Shankaranarayan, G., Ziad, M., and Wang, R. Y., 2003, Managing Data Quality in Dynamic Decision Environments: An Information Product Approach, Journal of Database Management (JDM), 14 (4).

Shvaiko, P., Euzenat, J., 2013, Ontology Matching: State of the Art and Future Challenges, IEEE Trans. Knowledge and Data Engineering 25(1).

Tallon, P. and Scannell, P., 2007, Information Lifecycle Management, Communications of the ACM, Volume 50 Issue 11, pp. 65-69.

Tallon, P., 2013, Corporate Governance of Big Data: Perspectives on Value, Risk, and Cost, IEEE Computer, Vol 46, Issue: 6.

Turchin, P., Brennan, R., Currie, T., Feeney, K., Francois, P., Hoyer, D., Manning, J., Marciniak, A., Mullins, D., Palmisano, A., Peregrine, P., Turner, E.A.L., Whitehouse, H., 2015, Seshat: The Global History Databank, Cliodynamics: The Journal of Quantitative History and Cultural Evolution, 6, (1), 2015, p77 – 107.

Viscusi, G., and Batini, C., 2014, Digital Information Asset Evaluation: Characteristics and Dimensions", in Smart Organizations and Smart Artifacts: Fostering Interaction Between People, Technologies and Processes, pp. 77-86, Springer International Publishing. DOI:10.1007/978-3-319-07040-7_9.

Weber, K., Otto, B. and Osterle, H., 2009, One size does not fit all—a contingency approach to data governance. J.Data Inf. Qual. 1(1), 1–27.

Wijnhoven, F., Amrit, C. and Dietz, P., 2014, Value-based File Retention: File Attributes as File Value and Information Waste Indicators, ACM journal of data and information quality, 4 (4). 15 -. ISSN 1936-1955.

Yousif, M., 2015, The Rise of Data Capital, IEEE Cloud Computing.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., and Auer, S., 2015,Quality Assessment for Linked Data: A Survey, Semantic Web – Interoperability, Usability, Applicability, IOS Press, ISSN: 1570-0844.