# The Challenge of Test Data Quality in Data Processing

CHRISTOPH BECKER, University of Toronto
KRESIMIR DURETEC, Vienna University of Technology
ANDREAS RAUBER, Vienna University of Technology

CCS Concepts: • **CCS →** **Information systems →** **Information retrieval →** **Evaluation of retrieval results →** *Test collections;* **Software and its engineering →** **Software organization and properties →** **Software functional properties →** *Correctness;* **Software and its engineering →** **Software creation and management →** *Software verification and validation*; **Information systems →** **Information systems applications →** *Digital libraries and archives*

Additional Key Words and Phrases: benchmarking, test data, data quality, test oracle, data processing, data formats, ground truth, digital curation, digital preservation, model-based testing, quality model

**ACM Reference Format**:

Christoph Becker, Kresimir Duretec, and Andreas Rauber, 2016. The Challenge of Test Data Quality in Data Processing. *J. Data and Information Quality x, x*, Article zz (month 2016), n pages. DOI:http://dx.doi.org/10.1145/0000000.0000000

## 1. INTRODUCTION

Data processing modules such as format identification and verification tools [Ferro 2016], information object converters [Wing and Ockerbloom 2000] or data acquisition modules [Nardo et al. 2013] are essential components of information systems. Verification of their functional correctness – the "degree to which a product or system provides the correct results with the needed degree of precision" [ISO 2011]– is an inevitable prerequisite for relying on their outputs. High quality test data are an essential element of dynamic verification. For example, the key challenge in developing test suites to ensure correct data acquisition operations in mission-critical environments lies in the test data [Nardo et al. 2013]. In digital preservation, technical interventions such as migration to new formats must respect the original type [Wing and Ockerbloom 2000] so that the result can be considered an authentic reproduction of the original. This must be verified through independent tests that extract features of the original and resulting objects to facilitate comparison [Milic-Frayling 2010]. Dynamic verification in these scenarios relies on mappings between test data and ground truth to address the *test oracle* problem, the question of distinguishing correct from incorrect behavior [Staats et al. 2011].

Data processing components are often homogeneous in function but varied in quality [Becker and Rauber 2010]. This led some domains to develop *benchmarking* initiatives to support research and innovation [Voorhees and Harman 2005]. Each benchmark provides "a procedure, problem, or test that can be used to compare systems or components" [ISO 2010]. A data processing benchmark requires a

Part of this work was supported by the Vienna Science and Technology Fund (WWTF) through the project BenchmarkDP (ICT12-046) and by NSERC through RGPIN-2016-06640.

Author's addresses: C. Becker, Faculty of Information, University of Toronto; C.Becker, K. Duretec and A. Rauber, Institute of Software Technology and Interactive Systems, Vienna University of Technology.



DOI:http://dx.doi.org/10.1145/0000000.0000000

specification of the *task sample* consisting of the specified *function* (e.g. extraction of text from a document [Duretec et al. 2015]), the *data set*, and the *test oracle* or ground truth. This is combined with metrics to compare the components [Sim et al. 2003; Duretec et al. 2015]. The benchmark's quality rests on the quality of its data set and test oracle. However, this *test data quality* is often low or unknown. For example, the recent drive in digital curation towards benchmarking [Duretec et al. 2015; Arocena et al. 2016] is motivated by the need for a robust shared evidence base. Yet, most components remain unverified since there are no data sets with test oracles of sufficient quality [Neumayer et al. 2007; Becker and Duretec 2013].

## 2. CHALLENGES AND RESEARCH DIRECTIONS

This profound lack of high-quality data sets to enable testing and benchmarking of data processing components highlights open research challenges in data quality related to (1) sampled data quality, (2) synthesized data and (3) quality models.

**Sampling** and annotating data from real-world collections has been the common approach in fields such as information retrieval (IR) [Sanderson 2010], digital forensics [Garfinkel 2012] and digital preservation [Neumayer et al. 2007]. In IR, crowd-sourcing has been used successfully to create annotations [Lease and Yilmaz 2013]. Economic resource constraints often limit annotation, and legal constraints prevent many data set owners from sharing [Neumayer et al. 2007; Downie 2003].

The noisy character and the presence of ill-formed data in real collections are essential to testing data processors, but often, no ground truth exists. In data processing, the complex relationships between input formats, computational processes, and results make it highly challenging to formalize, create and validate the test oracle. Some scenarios require mappings between the conceptual elements of digital objects and the symbol structures used for representing them [Becker and Duretec 2013]. The computational complexity involved makes this challenging, although the feasibility of formal mappings between formats and conceptual models has been demonstrated [Hartle et al. 2008]. Since the features are inevitably computed by an algorithm, generating the ground truth for sampled data to provide test oracles is very effort-intensive even for well-understood tasks [Strecker et al. 2009]. Even when the mapping relies less extensively on complex computation, it requires sophisticated models [Nardo et al. 2015]. The problem is circular: The correctness of any algorithm used to compute an annotation on a data set needs to be verified using a test data set [Becker and Duretec 2013; Milic-Frayling 2010].

**Synthesis** is the alternative. The principled solution to the above problem starts from the test oracle and generates artificial test data to match it. The considerable challenges involved in controlling and automating the complex transformations needed for generating fine-granular data sets have only recently been tackled for well-defined domains using model-driven engineering techniques. These allow the formalized representation of multiple linked levels of conceptual models with increased degrees of data format specificity; support transformations from platform-independent models to format-specific implementation constructs; and can generate code that is executed in runtime environments to produce test data. In principle, this approach to model-based testing can simulate realistic data sets with a full trace of ground truth accompanying the generated artifacts [Becker and Duretec 2013; Nardo et al. 2013]. However, it is inevitably limited in generating faulty data [Nardo et al. 2015]. Since the data cannot fully replace real-world collections, a mix of both is needed [Ferro 2016].

**Models of test data quality** are needed for a systematic evaluation of the fitness for purpose of individual test data sets, identify concrete shortcomings, and effectively combine data from different sources. The metrics must at least address a data sets' test coverage in relation to identified tasks; the presence and reliability of the test oracle; and the degree to which the data set approximates real-world collections. Test data adequacy is a long-recognized concern in software engineering, but no comprehensive quality models address the concerns of data processing.

## 3. CONCLUSIONS

The need for robust test data sets for data processing presents challenging research questions in data and information quality. Adequate ground truth must accompany test data to provide the test oracle. Novel approaches to model-based testing use model-driven engineering technologies to synthesize test data and oracles. These seeds of the emergent area of model-driven test data generation for complex data processing tasks present a promising alternative to the prevailing approach of sampling and annotation. Robust quality models for test data sets are needed to evaluate emerging approaches and allow the systematic development of heuristics to combine sampled annotated data with synthetic generated data.

## REFERENCES


Patricia C. Arocena, Boris Glavic, Giansalvatore Mecca, Renee J. Miller, Paolo Papotti, and Donatello Santoro. 2016. Benchmarking Data Curation Systems. *IEEE Data Eng. Bull.* 39 (2016).

Christoph Becker and Kresimir Duretec. 2013. Free Benchmark Corpora for Preservation Experiments: Using Model-driven Engineering to Generate Data Sets. In *ACM/IEEE JCDL*. New York, NY, USA: ACM, 349–358.

Christoph Becker and Andreas Rauber. 2010. Improving component selection and monitoring with controlled experimentation and automated measurements. *Inf. Softw. Technol.* 52, 6 (June 2010), 641–655.

J. Stephen Downie. 2003. Music information retrieval. *Annu. Rev. Inf. Sci. Technol.* 37, 1 (January 2003), 295–340.

Kresimir Duretec, Artur Kulmukhametov, Andreas Rauber, and Christoph Becker. 2015. Benchmarks for Digital Preservation tools. In *International Conference on Digital Preservation (IPRES 2015)*. Chapel Hill, NC, USA.

Nicola Ferro. 2016. Proposal for an Evaluation Framework for Compliance Checkers for Long-term Digital Preservation. In *12th Italian Research Conf. on Digital Libraries (IRCDL)*.

Simson Garfinkel. 2012. Lessons learned writing digital forensics tools and managing a 30TB digital evidence corpus. *Digit. Investig.* 9 (2012), S80–S89.

Michael Hartle, Friedrich-Daniel Möller, Slaven Travar, Benno Kröger, and Max Mühlhäuser. 2008. Using Bitstream Segment Graphs for Complete Description of Data Format Instances. In *ICSOFT (ISDM/ABF)*. 198–205.

ISO. 2011. *IEC 25010: 2011: Systems and software engineering–Systems and software Quality Requirements and Evaluation (SQuaRE)–System and software quality models*, Int. Standards Org.

ISO. 2010. *ISO/IEC/IEEE 24765:2010 - Systems and software engineering -- Vocabulary*, Int. Standards Org.

Matthew Lease and Emine Yilmaz. 2013. Crowdsourcing for information retrieval: introduction to the special issue. *Inf. Retr.* 16, 2 (March 2013), 91–100. DOI:http://doi.org/10.1007/s10791-013-9222-7

Natasa Milic-Frayling. 2010. Digital Object Characterization: Document Conversion and Qualiity Assurance. *Dagstuhl Semin. Proc.* (2010).

D. Di Nardo, N. Alshahwan, L.C. Briand, E. Fourneret, T. Nakić-Alfirević, and V. Masquelier. 2013. Model based test validation and oracles for data acquisition systems. In *IEEE/ACM ASE*. 540–550.

D. Di Nardo, F. Pastore, and L. Briand. 2015. Generating Complex and Faulty Test Data through Model-Based Mutation Analysis. In *IEEE ICST*. 1–10. DOI:http://doi.org/10.1109/ICST.2015.7102589

Robert Neumayer et al. 2007. On the need for benchmark corpora in digital preservation. In *Proceedings of the 2nd DELOS Conference on Digital Libraries*.

Mark Sanderson. 2010. Test collection based evaluation of information retrieval systems. *Found. Trends® Inf. Retr.* 4, 4 (2010), 247–375.

S.E. Sim, S. Easterbrook, and R.C. Holt. 2003. Using Benchmarking to Advance Research: A Challenge to Software Engineering. In *Int. Conf. Softw. Eng. (ICSE)*. Washington, DC, USA: IEEE Computer Society, 74–83.

M. Staats, M.W. Whalen, and M.P.E. Heimdahl. 2011. Programs, tests, and oracles: the foundations of testing revisited. In *Int. Conf. Softw. Eng. (ICSE)*. 391–400.

Thomas Strecker, Joost Van Beusekom, Sahin Albayrak, and Thomas M. Breuel. 2009. Automated ground truth data generation for newspaper document images. In *ICDAR 2009*. IEEE, 1275–1279.

Ellen M. Voorhees and Donna K. Harman. 2005. *TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing)*, The MIT Press.

Jeannette M. Wing and John Ockerbloom. 2000. Respectful Type Converters. *IEEE Trans Softw Eng* 26, 7 (July 2000), 579–593. DOI:http://doi.org/10.1109/32.859529