2012 IEEE 36th International Conference on Computer Software and Applications

(Open) Data Quality

Ina Schieferdecker
Competence Center Modeling and Testing of Systems and Service Solutions,
Fraunhofer FOKUS Berlin
ina.schieferdecker@fokus.fraunhofer.de

I. Open Data

Be it environmental data, public utility data, geo data, or some other data the open data approach is a way to open up the most important resource of the 21st century information and further develop our knowledge-based society. Innovation happens when data is open, available online, structured, and readable by man and machine.

Open data is hardly a new concept, its origins being easily traced to open science data, a fairly common practice among scientists. Its translation to governmental data is credited to Edd Dumbill in the 2005 XTech conference, acknowledged by the OECD [2], and supported by Tim Bray and Tim OReilly in 2006 [1]. Actors in the modern society will have to decide to what extent they want to share information with third parties, such as application developers or commercial companies, without losing a competitive advantage, or worse, violating the privacy of its users.

The provision of open data will make processes and software-based systems more understandable and strengthen the trust between industry, governments and civil societies. With the Internet of Things, the amount of data will continue to grow. It is for these reasons that there is a need to manage the data and their quality.

II. Data Quality

Requirements on the quality of data depend on the exact context of a software and has to be set specifically for the different kinds of data in software-based systems. Quality of data characteristics according to ISTQB [3] include aspects of

- Currency: Is it volatile? Is it punctual?
- Relevance: Information demand met?
- Consistency: Is the data consistent?
- Reliability: Transformation and origin of data correct and traceable?
- Correctness: Is the data syntactically and semantically correct? Does the data portray reality?
- Completeness: Does the detailing and the amount of data fit to the purpose?

III. Data Quality Assurance

According to the Global Data Synchronisation Network (GDSN [4]), data quality assurance consists mainly of product inspections, internal data alignment, data quality management system, as well as of education and training. However, data-oriented testing, review, static analysis or dynamic testing to assess and improve the data quality are side considerations only.

Also, when considering data quality in testing one can observe that mainly

- input data is generated to steer functional and non-functional testing,
- data constraints are used to partition systems's test data and to slice system behavior, and
- database testing is used to check the structure and logic of databases and their applications.

Hence, data is mainly a means for testing software-based systems but not itself the target of testing. One can conclude that data quality assurance needs to be improved by extracting, agreeing, developing principles, methods and approaches checking the quality of data. This is particularly needed because of increasing demands on the correctness, suitability, timeliness, and trustworthiness of data as well as their increasing complexity and size.

IV. A Data Quality Vision

Consequently, a data quality framework needs to be developed addressing

- the classification of data into data kinds,
- the definition of data quality per data kinds, and
- the development of concepts and methods to statically and/or dynamically check and improve data quality

These need to be integrated with approaches on

- data management,
- data governance, and
- data maintenance and cleansing.

The main challenge is developing a basic understanding on appropriate references to which the data in software-based systems can be compared to. What are feasible and practical requirements, specifications and/or models for such data. In addition, both data and its metadata need to be quality-assured wrt. its structures as well as its contents.

0730-3157/12 $26.00 © 2012 IEEE
DOI 10.1109/COMPSAC.2012.120

Once these have been elaborated, new concepts, methods and techniques to

- test data and meta-data,
- test the data semantics (e.g. along linked data),
- test real-time data (e.g. streams and flows),
- test geo-data, images, etc., or
- test statistical data, data aggregates, etc.

Along this, measures for the quality characteristics, coverage criteria and according data selection and data comparison methods are to be developed.

Eventually, a data quality framework can evolve that is as strong as the existing frameworks on the design, structure, and behavior of software-based systems.

V. Short CV

Prof. Dr.-Ing. Ina Schieferdecker studied Mathematical Computer Science at Humboldt-University Berlin and did her PhD in 1994 at Technical University Berlin on performance-extended specifications and analysis of QoS characteristics. Since 1997, she is heading the Competence Center for Testing, Interoperability and Performance (TIP) at the Fraunhofer Institute on Open Communication Systems (FOKUS), Berlin and is heading now the Competence Center Modelling and Testing for System and Service Solutions (MOTION). She is Professor on Model-Driven Engineering and Quality Assurance of Software-Based Systems at Free University Berlin since 20011 and was Professor on Engineering and Testing of Telecommunication Systems at Technical University Berlin 2003-2011. Prof. Schieferdecker works since 1994 in the area of design, analysis, testing and evaluation of communication systems using specification-based techniques like UML (Unified Modelling Language), MSC (Message Sequence Charts) and TTCN-3 (Testing and Test Control Notation). Prof. Schieferdecker authored many scientific publications in the area of system development and testing. She is co-founder of the Testing Technologies IST GmbH, Berlin, member of the German Testing Board, and member of the German Academy of Technical Sciences (acatech).

References

[1] T. Bray and T. OReilly, *Open Data*, Proc. of the OSCON Conference, Portland, OR, USA, July 2006 (http://conferences.oreillynet.com/os2006).

[2] OECD, Science, Technology and Innovation for the 21st Century, Meeting of the Committee for Scientific and Technological Policy at Ministerial Level, Paris, France, Jan. 2004 (http://www.oecd.org/document/15/0,3746,en_21571361_34590630_25998799_1_1_1_1,00.html).

[3] International Software Testing Qualifications Board (ISTQB): Certified Software Tester - Foundation and Advanced Level Syllabus, Software Testing Glossary (http://www.istqb.org).

[4] GCI/CapGemini Report: Internal Data Alignment, May 2004.