Software and data quality

Jakost softwaru a dat

J. Vaníček

Czech University of Agriculture, Prague, Czech Republic

Abstract: The paper presents new ideas in the International *SQuaRE* (***S***oftware ***Qua***lity ***R***equirements and ***E***valuation) standardisation research project, which concerns the development of a special branch of international standards for software quality. Data can be considered as an integral part of software. The current international standard and technical report of the ISO/IEC 9126, ISO/IEC 14598 series and ISO/IEC 12119 standard covert the whole software as an indivisible entity. However, such data sets as databases and data stores have a special character and need a different structure of quality characteristic. Therefore it was decided in the SQuaRE project create a special international standard for data quality. The main idea for this standard and the critical discussion of these ideas is presented in this paper. The main part of this contribution was presented on the conference Agricultural Perspectives XIV, aligned by Czech University of Agriculture in Prague, September 20 to 21, 2005.

Key words: software quality; data quality; international standardization; project SQuaRE

Abstrakt: Příspěvek pojednává o novém záměru v mezinárodním normalizačním výzkumném projektu SQuaRE (Požadavky na jakost software a jejich hodnocení), který spočívá v zpracování zvláštní skupiny norem pro jakost softwaru. Za integrální součást software mohou být považována i data. Stávající mezinárodní normy řad ISO/IEC 9126, ISO/IEC 14598 a norma ISO/IEC 12119 pokrývají software jako nedělitelnou entitu. Avšak takové datové množiny jako jsou báze dat a datové sklady mají zvláštní charakter, který vyžaduje odlišnou strukturu charakteristik jakosti. Proto bylo v rámci projektu SQuaRE rozhodnuto vytvořit zvláštní mezinárodní normu pro jakost dat. V tomto příspěvku jsou uvedeny hlavní myšlenky této připravované mezinárodní normy a podrobeny kritické diskusi. Podstatná část tohoto příspěvku byla přednesena na konferenci Agrární perspektivy XIV, uspořádané Českou zemědělskou univerzitou v Praze 20.–21. září 2005.

Klíčová slova: jakost softwaru, jakost dat, mezinárodní normalizace, projekt SQuaRE

This contribution reflects the current state of the work on the SQuaRE project and is based on the valid and prepared international standards and technical reports of the ISO/IEC 9126, ISO/IEC 14598 and ISO/IEC 250xx series; mainly use of the principles of Azuma (2001) or Azuma and Vaníček (2001).

Quality is the key attribute of each product. Product is a result of each process, which is intended to delivery. There are four generic categories of products:
– *Services*
– *Software* (in the more general sense than only computer software)
– *Hardware* (also not only a computer hardware)
– *Processed material.*

Information systems as product are mostly a combination of software, hardware and associated services.

Quality in the general sense is the degree to which a set of inherent characteristics fulfils the requirements. Requirements are needs or expectations that are stated, generally implied, or obligatory. From the user's point of view, the product quality can be considered as a degree in which the product fulfils the user needs. The transformation of needs to exact requirements can be a problem, which is not easy to solve.

Quality of each product is the key characteristic for the success in the market. The quality care is also the main motive power of the industrial and social progress General requirements for the quality management to creating process of each product are given in the ISO 9000 and ISO 9001 standards. These standards are mostly process oriented and are intended previously for developers. It is not easy to

Supported by the Ministry of Education, Youth and Sports of the Czech Republic (Grant No. MSM 6046070904 – Information and knowledge support of strategic management).

establish such general standards for project from the acquirers and users point if view, because products are of the different nature. The main effort for software production consists in developer's process, and in product maintenance. The necessary effort to manifold software is only marginal.

For software products and products, which contain software as their crucial part, such as information systems, the international standards are prepared by the Joint Technical Committee "Information Technology", organized by the cooperation of ISO and IEC. Today the following main series of international standards are available:

– *ISO/IEC 15939 "Information engineering – Software engineering – Software measurement process"*, which describes the measurement framework for general software attributes, including quality attributes.
– *ISO/IEC 9126 series "Information technology – Software product quality"*, which contains the standard *9126-1 "Quality model"*, and three technical report *9126-2, 9126-3* and *9126-4*, describing the *"Internal quality measures", "External quality measures"* and *"Quality in use measures"*, respectively.
– *ISO/IEC 14598 series "Information technology – Software product evaluation"*, with six standards describing the quality evaluation process from various points of view.
– *ISO/IEC 12119 "Information technology – Software packages – Quality requirements and testing"*, which describes the necessary information concerning quality, which is the supplier of the off the shelf software obliged to be published before the contract and the rules of testing these requirements.

In the present state, the set of standards for software product quality is a little bit inconsistent and rather blind. These standards have not a unified terminology and do not fully reflect the current state of art in software engineering. Therefore, the international project *SQuaRE* starts with the aim to develop a consistent series *ISO/IEC 250xx* for the software product quality. The intended structure of the SQuaRE project is the following architecture of the 250xx series, see Figure 1; more details in Azuma, Vaníček (2001) or in the textbook Vaníček (2005).

During the works on SQuaRE project, the requirement to extend the set of 250xx standards to a special standard for a date quality occurs. The reason why the special data quality standard is prepared and why it is not sufficient to evaluate the quality of data as an integral part of software (software is defined as a program with the associated data and documentation) is the following. In the modern data processing, the same data warehouse can be used for various software products and also in different information systems. Therefore, the separate and independent data quality evaluation seems to be useful. The problem how to add the data quality standard into an intended SQuaRE standard structure is not solved till now. This problem shall be probably solved by

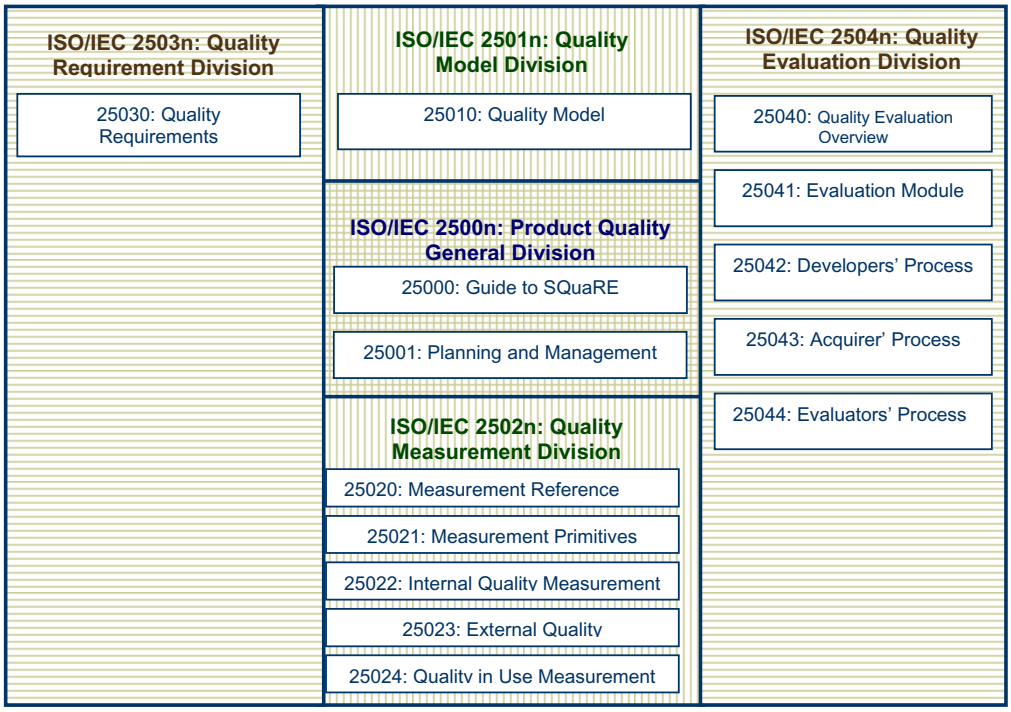

Figure 1. Intended structure of ISO/IEC 250xx set of standards

the multilevel structure simultaneously with some of other ideas how to extend the ISO/IEC 250xx series for another quality standards with the special branches of software. Such directions are the ready made software which quality will be standardized by the standard "*Software Engineering – Software product evaluation – Requirements for quality of Commercial Off-The-Shelf (COTS) software product and instructions for testing*", intended to replace the current ISO/IEC 12119 and other related standards for special software branches.

This structure is outlined in Figure 2.

The intention of this paper is to describe the base terminology, approaches and principles, which will be probably used in the new ISO/IEC standard for the data quality as well as the possible questions and problems, which accurse during the work on this standard.

TERMS INTENDED TO BE USED IN INTENDED DATA QUALITY STANDARD

The following meanings are intended for the main terms in data quality standard:

Attribute is an inherent property or characteristic of an entity that can be distinguished quantitatively or qualitatively by human or automated means.

Entity is an object that is to be characterised by measuring its attributes.

Data is a reinterpretable representation of information. Data is a representation of the perception of the real world. Any data has a degree of information. Data is considered stored in an electronic format and managed by an information system.

Data value is the atomic content of an attribute.

The structured data are stored on an electronic device and are described through its structure, which is managed by an Information System. Structured data are considered at different levels such as: file (table, segment etc), record (row, topple etc.), specific field (attribute, column etc.) and the atomic content of a field. The data description and the structure in a data base management system are logically stored with the data, while in different legacy data organizations, the data description and the structure can reside separately by the data.

Data format is the description of: type, length or scale/precision of data (e.g. type character, numerical, date, timestamp etc.)

Information is the meaning given to data or the interpretation of data, based on its context. The finished product is the result of the interpretation of such data (English 1999).

Metadata is the data that describes other data.

Technology is the framework for defining, manipulating and managing data. It consists of hardware

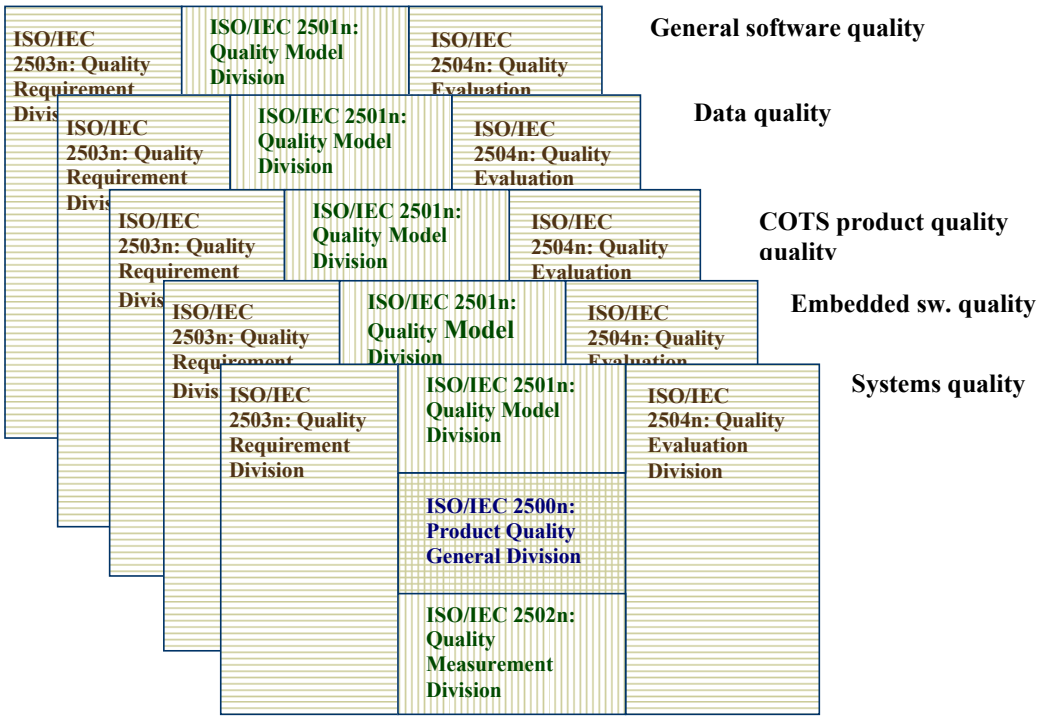

Figure 2. The possible multilevel structure of the set of ISO/IEC 250xx standards

components, software components of operating systems, data base management systems or middleware in general.

INTENDED APPROACH TO DATA QUALITY

The following diagram shows the life cycle for data beginning with its generation from a source through its application by a user. There are 3 stakeholders or contributors to data quality:
– The data *producer*
– The data *evaluator*
– The data *user*.

The data *producer* (or developer) can be an individual, an organization, agency, institution, who has the responsibility for generating data. The data producer shall:
– produce data compliant with the quality characteristics listen on Figure 4
– transmit data in accordance with the procedures defined/agreed by the acquirer/evaluator of the data
– respond to data evaluator and/or data user requests for correction of the faulty or completion of the missing data.

The data *evaluator* (or acquirer) may be an individual, an organization, agency, institution, who has the responsibility to determine the quality of the data. The data evaluator has the responsibility to determine if the quality of the data is adequate for the intended user. This involves risk analysis of the adequacy of the quality characteristics including the format, correctness completeness, understandability, accuracy, precision and relevance for the intended use. The data quality evaluator uses risk analysis to determine the adequacy of the data for the intended purpose. It may be unwise and costly to require the data that are more precise or accurate than needed.

The data *user* is the stakeholder/s affected by the quality of the data. The user/s may be the producer, the evaluator, or a user that is not the producer or an evaluator. The user may be the processor of the data or the recipient of the processed data. The data user and the evaluator may collaborate to determine risks.

The life cycle for data beginning with its generation by a source through its application by a user is shown on Figure 3.

DATA QUALITY MODEL

The general concept of the intended model, does not directly deal with how databases have been conceptually and logically designed. In this model, "data" refers to "structured data", without making distinctions between them.

There are different views of Data Quality:

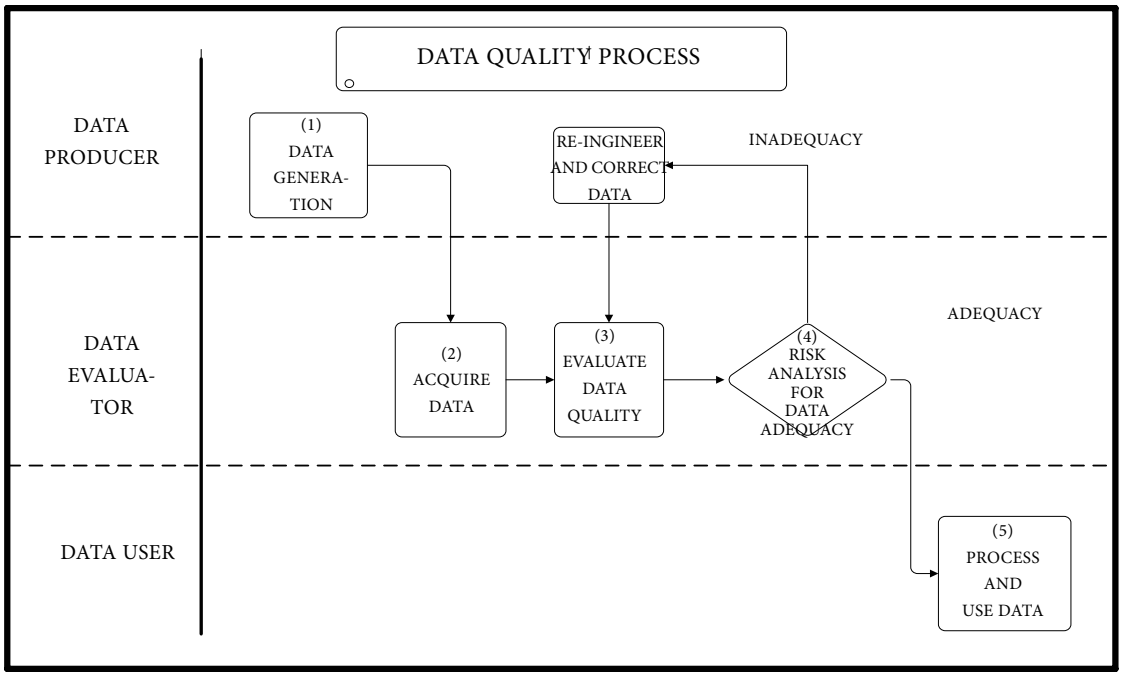

Figure 3. Data quality life cycle

Internal Data Quality: Capability of a set of static attributes of data to satisfy the stated and implied needs when the data are used under specified conditions.

External Data Quality: Capability of data to enable the behaviour of a system to satisfy the stated and implied needs when the system is used under specified conditions

Data Quality in Use: Capability of the data to enable specific users to achieve specific goals with timeless amount of information, relevancy, credibility, understandability in the specific contexts of use.

During the tentative preparation of data quality standard project, the advancement from the first concepts described in Vaníček (2003) in the direction to be more conform to other software quality model occurs. The Figure 4 utlines the Data Quality Model (prepared for the ISO/IEC 25012) after having taken into account the already mentioned considerations. It categorises the internal and external data quality attributes into six characteristics defined in the ISO/IEC 9126-1 and intended to be use also in the ISO/IEC 250xx (functionality, reliability, usability, efficiency, maintainability and portability), which are further subdivided in the mentioned standard into subcharacteristics.

INTERNAL AND EXTERNAL DATA QUALITY

The intention is to use for data quality the same set of six characteristics of data considered also from the internal and external point of view. The six characteristics (functionality, reliability, usability, efficiency, maintainability and portability) are mainly related to the values of data, to the format used for storing the data and the technological environment, i.e. hardware, software and middleware. For data quality, it seems to be useful to divide the accuracy subcharacteristic into syntactical and semantical accuracy and to add new subcharacteristic consistency, currency, completeness and precision instead of suitability, or a sub-subcharacteristic of suitability. The following definition for data quality characteristics and subcharacteristics are prepared on the SQuaRE project.

Functionality

Functionality is the capability of the data to provide functions, which meet its functional requirements. Functionality requirements are refined into the requirements for the data to be suitable, accurate, interoperable, secure and compliant with relevant functional standards and regulations. The suitability,

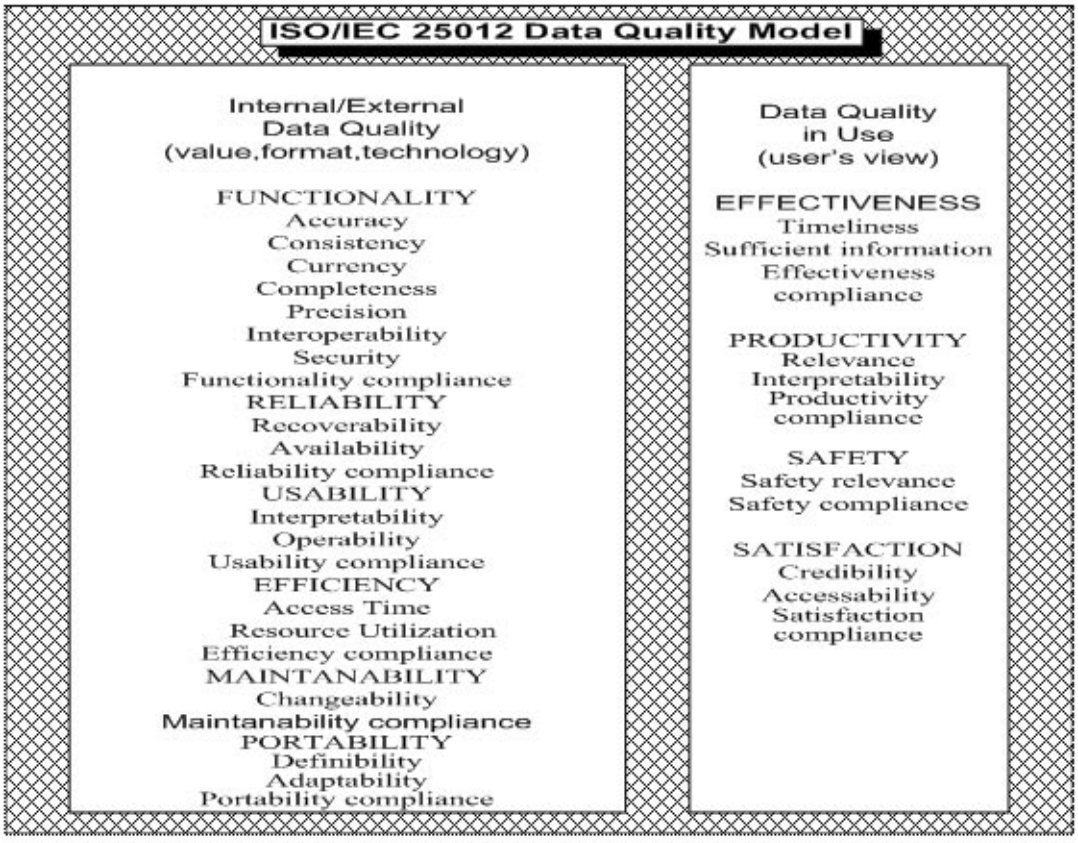

Figure 4. Data quality model

for what concerns data, is expanded into: consistency, currency, completeness and precision.

Consistency

Consistency refers to the absence of apparent contradictions in a database. For example, the data value name and sex related to the same person must be consistent, Mary cannot be a man; the information of "Name" and "Sex" must be compatible. Inconsistency can be verified on the same or different tables.

Currency

Currency is the extent, to which data are up-to-date (Redman 1996). It is critical for volatile data (by "volatility" we mean how frequently the data is updated) (Bovee et al. 2001).

Completeness

Completeness is the fraction (ratio) between the amount of data stored and the amount of data in an ad-hoc reference set. For example: if a database contains information about company employees, all the employees must be recorded.

Precision

The level of detail of the data that measures the ability to distinguish between the nearly equal values. For example, a three-digit numeral to the base 10 discriminates among 1 000 possibilities.

Accuracy

Accuracy is defined as the degree of conformity of an acquired, measured and calculated value to its actual or specified value. Accuracy has two main aspects:

Syntactical accuracy

Syntactical accuracy is defined as the closeness of the data values to a set of values defined in a domain considered syntactically correct. For example: a low degree of syntactical accuracy is when the name Mary is stored as Mari.

Semantic accuracy

Semantic accuracy is defined as the closeness of the data values to a set of values defined in a domain considered semantically correct. For example: a low degree of a semantical accuracy is when the name Mary is stored as John. John is syntactically accurate, because of the domain of reference in which it resides, but it is another name.

Interoperability

Interoperability is the capability of data to be accessed and exchanged among different platforms and systems.

Security

Security is the capability of the data to be accessed only by authorized users.

Functionality compliance

Functionality is the capability of the data to adhere to standards, conventions or regulations in laws and similar prescriptions relating to functionality.

Reliability

Reliability is the capability of the data to maintain a specified level of operations in a specific environment. Reliability requirements are refined into requirements for the data to be recoverable, available and compliant with relevant reliability standards and regulations.

Recoverability

Recoverability is the capability of the data to maintain and preserve its physical and logical integrity, even in the event of failure (for example with features like: commit/synch point, rollback and backup-recovery).

Availability

Availability is the capability of data to be always retrievable.

A particular case of availability is the concurrent access (in reading and writing) by more than one user and/or application.

Reliability compliance

Reliability compliance is the capability of the data to adhere to standards, conventions or regulations in laws and similar prescriptions relating to reliability.

Usability

Usability is the capability of the data to be understood, learned, and used and to catch the user's attention, when used under the specified conditions. Usability requirements are refined into requirements for the data to be untreatable, operable, attractive for the users and capable to adhere to standards, conventions or regulations in laws and similar prescriptions relating to usability.

Understandability

Undesirability is the extent to which data is in the appropriate languages, symbols and units, and to which its definitions are clear (Wang, Strong 1996; Pipino et al. 2002).

Operability

Operability is the capability of the data to be represented appropriately from the functional point of view. For example, the data representing "date" should be represented in "date format" and not like an integer number, to permit operations, such as the difference between dates.

Attractiveness

Attractiveness is the capability of the data to be attractive to the user.

Usability compliance

Usability compliance is the capability of the data to adhere to standards, conventions or regulations in laws and similar prescriptions relating to usability.

Efficiency

Efficiency is the capability of data to be processed by software products which provide the appropriate performance, relative to the amount of resources used, under stated conditions. Efficiency requirements are refined into requirements for the time behaviour, resource compliance with relevant efficiency standards and regulations.

Time behaviour

Time behaviour is the capability of the data to be accessed within the appropriate processing times and within the throughput rates (Wang, Strong 1996).

Resource utilization

Resource utilization is the capability of the data to be stored or accessed using the appropriate amounts and types of resources under the stated conditions.

Efficiency compliance

Efficiency compliance is the capability of the data to adhere to standards, conventions or regulations in laws and similar prescriptions relating to efficiency.

Maintainability

Maintainability is the capability of data to be modified for changes in environments, in requirements or in functional specifications. Maintainability requirements are refined into requirements for the data to be changeable and capable to adhere to standards, conventions or regulations in laws and similar prescriptions relating to maintainability.

Chengeability

Changeability is the capability of the data to be changed in its format.

Portability

Portability is the capability of data to be transferred from one environment to another. Portability requirements are refined into the requirements for the data to be adaptable and capable to adhere to standards, conventions or regulations in laws and similar prescriptions relating to portability.

Adaptability

Adaptability is the capability of the data to be moved from one platform to another.

Maintainability compliance

Maintainability compliance is the capability of the data to adhere to standards, conventions or regulations in laws and similar prescriptions relating to maintainability.

DATA QUALITY IN USE

Data Quality in Use is the capability of data to enable users to reach specific goals in terms of effectiveness, productivity, safety and satisfaction in specified contexts of use.

Effectiveness

Effectiveness is the capability of data to enable users to achieve the specified goals with accuracy and completeness in a specified context of use. Requirements for the effectiveness can be refined into the requirements for the data to be timeliness, to contain sufficient information and adhere to standards, conventions or regulations in laws and similar prescriptions relating to effectiveness.

Timeliness

Timeliness is the extent to which data is sufficiently up-to-date for the task at hand (Wang, Strong 1996).

Sufficient information

Sufficient information is the completeness of data from the user's point of view and its stated needs. For example, data have the right "amount of information" if there are all fields that user needs and nullable fields have a "valid value" where the user is expecting one.

Effectiveness compliance

Effectiveness compliance is the capability of the data to contain sufficient information and adhere to standards, conventions or regulations in laws and similar prescriptions relating to effectiveness.

Productivity

Productivity is the capability of data to enable users to expend the appropriate amount of resources in relation to effectiveness achieved in a specific context of use. Requirements for the productivity can be refined into requirements for the data to be relevant, interoperable and to adhere to standards, conventions or regulations in laws and similar prescriptions relating to productivity.

Relevancy

Relevancy is the extent, to which data is applicable and helpful for the task at hand (Wang, Strong 1996; Pipino et al. 2002).

Interpretability

Interpretability is the extent to which data is easy to comprehend. Metadata can help the comprehension of data.

Productivity compliance

Productivity compliance is the capability of the data to adhere to standards, conventions or regula-

tions in laws and similar prescriptions relating to productivity.

Safety

Safety is the capability of data to achieve the acceptable level of risk of harm to people, business in a specified context of use. Requirements for safety can be refined into requirements for the data to be relevant and to adhere to standards, conventions or regulations in laws and similar prescriptions relating to productivity.

Safety relevance

Safety relevance is the capability similar prescriptions relating to safety.

Safety compliance

Safety compliance is the capability of the data to adhere to standards, conventions or regulations in laws and similar prescriptions relating to safety.

Satisfaction

Satisfaction is the capability of data to satisfy users in a specified context of use. Requirements for the satisfaction can be that the data are credible, accessible and adhere to standards, conventions or regulations in laws and similar prescriptions relating to satisfaction.

Credibility

Credibility is the extent to which data is regarded as true and credible (Wang, Strong 1996).

Accessibility

Accessibility is the capability of the data to be accessed, particularly by people who need the supporting technology or special configuration because of some disability.

Satisfaction compliance

Satisfaction compliance is the capability of the data to adhere to standards, conventions or regulations in laws and similar prescriptions relating to satisfaction.

CRITICAL DISCUSSION

The main problem concerning the development of new SQuaRE series of standard and also concerning the data quality standard is the enormous volume of standardisation documents. The quality aspects measurement is not a cheap procedure. If we extend the number and span of standards, nobody will use them. For the personal point of view and experience of the author, we are in this moment in the situation when "less can be more".

The second serious problem is that the SQuaRE development and consideration are permanent only on the general level. Extensive researches about the standard set structure, quality model and live cycle model and meta-models are available, but the concrete attributes and measures for quality requirements and quality evaluation are absent. It seems that the commitment what is tangible is essential for the software and data quality and how to measure it is not outside the door. Without concrete contents, the quality evaluation standards cannot be utilizable in practice.

The last but not least problem from the author's personal point of view consists in the non-fully consistent fragmentation of the quality into internal, external and quality in use. According to the quality definition, the product quality is only the external quality in the ISO/IEC standards terminology. The so-called "internal quality" does not fulfil any users or shack holder's requirements. It can be interpreted only as quality predictors or indicators. The so-called "quality in use" is not an attribute of the product. It is in fact the attribute of the process of product utilization. It depends not only on the product itself, but also on the organization, which uses the product, and several other factors. This aspect shall be evaluated using the ISO 9000 series standard, not according to the SQuaRE ISO/IEC 250xx standards. The author's felling is that the elucidation of this crisscross is absolutely necessary for the SQuaRE project success.

REFERENCES

Azuma M. (2001): Square, the Next Generation of the ISO/IEC 9126 and 14598 International Standard Series on Software Product Quality. Proceedings of the 12th European Software Control and Metrics Conference, London.

Azuma M., Vaníček J. (2001): SQuaRE: Next Generation of ISO/IEC 9126 & 14598. In: EurOpen CZ. XVIII konference Dolní Malá Úpa: 1–16.

Ballou D.P., Pazer H.L. (1985): Modeling data and process quality in multi-input, multi-output information systems. Management Science, *31* (2): 150–162.

Batini C., Lenzerini M., Navathe S.B. (1984): A comparative analysis of methodologies for database schema integration. ACM Computing Survey, *15* (4).

Bovee M., Srivastava R.P., Mak B.R. (2001): A Conceptual Framework and Belief-Function Approach to Assessing Overall Information Quality. Proceedings of the 6th International Conference on Information Quality, Boston, MA.

Calvanese D., De Giacomo G., Lenzerini M., Nardi D., Rosati R. (1998): Information Integration: conceptual Modeling and Reasoning Support. Proceedings of the 6th International Conference on Cooperative Information Systems (CoopIS'98), New York City, NY, USA.

Crosby P.B. (1984): Quality without Tears. McGraw-Hill, New York.

English L.P. (1999): Improving Data Warehousing and Business Information Quality, Wiley, New York.

Gertz M. (1998): Managing Data Quality and Integrity in Federated Databases. Second Annual IFIP TC-11 WG 11.5 Working Conference on Integrity and Internal Control in Information Systems, Airlie Center, Warrenton, Virginia.

Huang K.T., Lee Y., Wang R. (1999): Quality Information and Knowledge. Prentice Hall, Upper Saddle River.

ISO 8402: Quality Management and Quality Assurance-Vocabulary.

ISO/IEC 9126: Information Technology – Software Product Evaluation – Quality Characteristics and Guidelines for their Use.

ISO/IEC 9126-1: 2001. Software engineering – Product quality – Part 1: Quality model.

ISO/IEC TR 9126-2: 2003. Software engineering – Product quality – Part 2: External metrics.

ISO/IEC TR 9126-3: 2003. Software engineering – Product quality – Part 3: Internal metrics.

ISO/IEC TR 9126-4: 2004. Software engineering — Product quality – Part 4: Quality in use metrics

ISO/IEC WD circulation and CD Registration Ballot, WD 25012 , Software Engineering – Software Quality Requirements and Evaluation (SQuaRE) – Data Quality Model.

Kriebel C.H., Mikhail O. (1980): Dynamic pricing of resources in computer networks. Logistics.

Madnick S. (1999): Metadata Jones and the Tower of Babel: The Challenge of Large – Scale Semantic Heterogeneity. Proceeding of the 3rd IEEE Meta-Data Conference (Meta-Data '99), Bethesda, MA, USA.

Natale D. (1995): Qualità e Quantità nei Sistemi Software. Franco Angeli.

Natale D., Scannapieco M., Angeletti P., Catarci T., Raiss G. (2001): Qualità dei dati e standard ISO/IEC 9126: Analisi critica ed esperienze nella Pubblica Amministrazione Italiana", Workshop AIPA "Sistemi in rete nella Pubblica Amministrazione" with the participation of Sogei-Società Generale d'Informatica S.p.A. and the University of Rome "La Sapienza", (in conjunction with The VLDB Very Large Data Base Conference 2001) Rome, September. Esempi di misurazioni.

Olson J.E. (2003): Data Quality: the accuracy dimension. Ed. Morgan Kaufmann Publishers. Wang, R., Storey V.C., Firth, C.F. (1995): A framework for analysis of Data Quality Research. IEEE Transactions on Knowledge and Data Engineering, *7* (4): 623–640.

Pipino L., Lee Y., Wang R. (2002): Data Quality Assessment. Communications of the ACM, *45* (4).

Redman T.C. (1996): Data Quality for the Information Age. Artech House, Boston, London.

Scannapieco M., Catarci T. (2004): Data quality under the Computer Science perspective, University of Rome.

Ulmann J.D. (1997): Information Integration using Logical Views. Proceedings of the International Conference on Database Theory (ICDT '97), Greece.

Vaníček J. (2003): Data Quality model. In: Sborník příspěvků ze semináře k výzkumnému záměru VZ 411100010 "Zpracování dat a matematické modelování", ČZU, Praha: 21–26.

Vaníček J. (2005): Measurement and rating of information systems quality, Part 2 – Quality model. ČZU, Prague.

Wang R.Y., Strong D.M. (1996): Beyond accuracy: what data quality means to data consumers. Journal of Management Information Systems, *12* (4): 5–34.

Arrived on 1st February 2006

Contact address:

Jiří Vaníček, Czech University of Agriculture in Prague, Kamýcká 129, 165 21 Prague 6-Suchdol, Czech Republic
Tel.: +420 224 382 362, e-mail: vanicek@pef.czu.cz