Software quality measures validation in the Czech Republic

Validace měr pro jakost softwaru v České republice

J. Vaníček

Czech University of Life Sciences, Prague, Czech Republic

Abstract: The paper concludes the research results performed on the Faculty of Economics and Management, Czech University of Life Sciences in Prague, concerning the utilization and validation of external and internal software quality measures (metrics). The aim of this research was the validation of measures (metrics) recommended in the ISO/IEC 9126-2 and 9126-3 technical reports, with the intention to incorporate selected measures to international standards. The research presents the serious deficiencies and users provisions concerning these measures and the necessity of a deep revision of the set of measures before the decision about incorporating these measures into the ISO/IEC 250xx standards series, developed within the SQuaRE international research project. The main part of this contribution was presented at the conference Agricultural Perspectives XV, organised by the Czech University of Life Sciences in Prague, September 20 to 21, 2006.

Key words: ISO/IEC quality model, quality metrics, external measures, internal measures, quality characteristics, validation, user research

Abstrakt: Příspěvek shrnuje výsledky průzkumu provedeného na Provozně ekonomické fakultě České zemědělské univerzity v Praze o užití a validaci vnitřních a vnějších měr (metrik) jakosti softwaru. Průzkum byl snahou validovat míry (metriky) navržené v technických zprávách ISO/IEC 9126-2 a 9126-3 tak, aby vybrané míry bylo možné včlenit do mezinárodních norem. Průzkum ukázal vážné nedostatky a výhrady uživatelů k těmto mírám a na nutnost hluboké revize množiny měr před rozhodnutím které míry převzít do soustavy norem ISO/IEC 250xx, vyvíjených v rámci mezinárodního výzkumného projektu SQuaRE. Podstatná část tohoto příspěvku byla přednesena na konferenci Agrární perspektivy XV, uspořádané Českou zemědělskou univerzitou v Praze 20.–21. září 2006.

Klíčová slova: model jakosti ISO/IEC, metriky pro jakost, vnitřní míry, jakosti vnější míry jakosti, charakteristiky jakosti, validace, průzkum využívání

Software accomplishing the high level of quality is an essential tool for supporting all processes in economics, administration and environment management. The quality evaluation of software product on the market is so far mainly the subjective process. Therefore, the rules for the objective and unified rating of software quality are extremely eligible and were in the centre of interest for international stan-dardization. Two main international standardization organizations joined its effort for developing standards for software product quality within the Joint Technical Committee ISO/IEC JTC1 "Information Technology".

The history of the main international standardization organization for the software quality standardization is in short the following: In 1991, the ISO has published the first international standard ISO/IEC 9126 "Software Product Evaluation – Quality Characteristics and Guidelines for their use" (1991). In 2001, ISO made revision and enhancement on the ISO/IEC 9126 standard and published a new series of standards. The ISO/IEC 9126:1991 has been replaced by two related multipart standards: ISO/IEC 9126 (Software product quality) and ISO/IEC 14598 (Software product evaluation). All these standards were adapted by the CEN/CENELEC as the European

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

standards and by the Czech Standardization Institute (ČNI) as the Czech technical standards.

The first IS, ISO/IEC 9126, consists of one standard and three technical reports under the common title Information technology – Software product quality.

– ISO/IEC 9126-1: Quality model
– ISO/IEC TR 9126-2: External metrics
– ISO/IEC TR 9126-3: Internal metrics
– ISO/IEC TR 9126-4: Quality in use metrics.

The second multipart IS consists of six standards under the common title the Information Technology – Software product evaluation].

– ISO/IEC 14598-1: General overview
– ISO/IEC 14598-2: Planning and management
– ISO/IEC 14598-3: Process for developers
– ISO/IEC 14598-4: Process for acquirers
– ISO/IEC 14598-5: Process for evaluator
– ISO/IEC 14598-6: Documentation of evaluation modules.

From the authors' point of view, the main weakness of these sets of standards is the lame selection of the selected quality attributes and their *metrics*, in the new terminology accordance with the ISO/IEC 15939 Information technology – Software measurement process *measures* in ISO/IEC 9126-2, -3 and -4. The sets of attributes and measures nominated in these technical reports are too extensive, not consistent and contain also the measures, which are badly (defected, inconsistent or incomplete) defined.

The ISO/IEC JTC1 Committee for Information Technology has recognized a need for further enhancement of standards for software product quality primarily a result of advances in the fields of information technologies and changes in environment. Therefore, the team of standardization experts from 18 countries, including the Czech Republic, is now working on the next generation of software product quality standards (see Vaníček 2006a and Vaníček 2006b), which will be referred to as the *S*oftware *Qua*lity *R*equirements and *E*valuation (*SQuaRE* – ISO/IEC 25000). This series of standards will replace the current ISO/IEC 9126 and ISO/IEC 14598 series of standards. The SQuaRE series will consist of five divisions:

– Quality management division (ISO/IEC 2500n)
– Quality model division (ISO/IEC 2501n)
– Quality measurement division (ISO/IEC 2502n)
– Quality requirements division (ISO/IEC 2503n)
– Quality evaluation division (ISO/IEC 2504n).

This work is being carried out by the working group 6 (WG6) of the Software and System Engineering Subcommittee (SC7) of ISO/IEC JTC1. One of the main objectives of (and difference between) the SQuaRE series of standards and the current ISO/IEC 9126 series of standards is the coordination and harmonization of its contents with ISO/IEC 15939 and the care for consistency between the individual standards of the new 250xx series.

The author of this contribution expects that in the preparation of the SQuaRE standards, one problem is still underrated and not solved by the responsive effort. This problem is the selection of an agreed, not unduly extensive set of attributes and measures that covers all quality characteristics and subcharacteristics. This problem is of course arduous due to different experiences and practices in various countries and the different concernments of various software producers and acquires. Nevertheless, the author is convinced that the selection of an appropriate set of attributes and measures are a necessary condition for the success of the SQuaRE project.

OBJECTIVES AND METHODOLOGY

In this situation, within the research project MSM 6046070904 Information and Knowledge Support of Strategic Decision Process, the research with the target to map, which attributes and measures are used by software developers, acquires and evaluators in the practices and to map the experiences with the ISO/IEC 9126 usage have been done. On the Department of Information Engineering, Faculty of Economics and Management, the research was realized in the period 2002–2006. One professor, the member of the international SQuaRE research team (the author of this paper), 6 doctoral students and 12 master degree students take a part in the large scale inquire research, which covers 10 software developers, 12 system integrators and 48 information systems users in the Czech Republic. Developers, system integrators and users companies were of a very different size, from the branches of the main worldwide software companies to the small software houses with the staff from 10 to 20. The methodology of investigation was different for various respondents, from formal questioners fulfilling to informal talk with the staff.

RESULTS

The general perception from the research is that about 80% of software developers and system integrators declare the systematic care for software and system quality. Though they are more concentrated

at the process control than to the product control. The ISO 9000 series standard are known and used. Only about 50% of these companies are familiar with the ISO/IEC 9126 and ISO/IEC 14598 standards and use some measures recommended in these technical reports. Nobody follows consistently the process according to the ISO/IEC 14598 and evaluate all quality characteristics defined in the ISO/IEC 14598 by the internal and external measures recommended in related technical reports. But about 60% of these companies follow some quality prediction methods and quality testing which are similar to the ISO/IEC 9126 quality model and use some measures, which can be found in the respective technical reports or are some modification of the measures listed in these reports.

The contacted acquires and end users prefer for their decision of the information system or software product provision or in the competitive tendering the references about the vendors reputation before the proper systematic quality audit of the offered product. Still each user has declared some quality evaluation. This evaluation is in about 50% of cases concentrated to functionality evaluation only. Only about 20% of the acquires evaluate in fact all six or at least 5 selected five quality characteristics from the ISO/IEC 9126. Mostly they have no information about the ISO/IEC 9126 quality model, but follow some own methodology whitch is similar.

Concerning the measures used in practices, it is the following. Plenty of measures recommended in ISO/IEC 9126-2, -3 and -4 have the form

A/B

where B is the number of some properties of the evaluated entity, which are satisfied and B is the number of properties required. The range of such a measure is always a close interval <0, 1> and the rating "the closer to 1 is the better". Such a "normalization" of measures allows for comparing the measures for different attributes. However, it is only seldom indigent in the practices. More needful is the comparison of the requirements of the different required measures of various users and various stakeholders for the same attribute of the product in question. Second need is often to compute the measure value for some attribute of a complex entity by the reconciliation of measures acquired for its elements. The "measures" of the A/B type do not allow doing it. In fact, such fractions are not quality measures for the given entity, but the result of comparison of the actual measures with measure indicators derived from the individual quality requirements for concrete user or stakeholder. This comparison is not a part of the measure process. This step is important, but shall be realized later, after the measurement. It is the part of the measurement evaluation stage (see the ISO/IEC 15939) and can lead to different results for different stakeholders. This fact and the opportunity to concentrate the measures obtained for instalments of a complex product is the reason, why the fraction measures are not used and evaluators prefer to enrol the nominator A only as a proper measure of the entity.

The critical appreciation of users for the concrete software quality measures in the ISO/IEC 9126-2 and -3 come next. The measures for quality in use were not reported in the query. The thing is that these measures are not a product measures but the process measures for the quality of information pressing in the users company. Such information is customarily judged as confidential and companies are not obliged to communicate it.

The ISO/IEC 9126 series judges the software and system quality into six quality characteristics named:
- Functionality
- Reliability
- Usability
- Efficiency
- Maintainability
- Portability.

Each of those characteristics can be divided into some subcharacteristic. In the following text, the perceptions for each characteristic (denoted by ***bold italic*** font) and subcharacteristics (denoted in *italic*) will be listed.

Functionality is on the top of the respondent's concern. All respondents that realized some quality evaluation have the functionality evaluation as a part of the evaluation process.

Suitability subcharacteristic is for functionality judged as underlying by all respondents. It is often measured by the variant of "functional implementation completeness" and/or "functional implementation coverage" attribute and corresponding external or internal measure. The difference between these two attributes and measures is a little bit fuzzy. The first counts all functions, which are "present". The second counts only functions, which are "complete". The source of the problem consists in the absence of the correct and lucid definition of the term "number of functions", which is important quality measure element. Functions form not a linear structure, but a multilevel hierarchy. One complex function can be considered as a collection of some elementary or less complex functions. One upper level function which is realized, but its realization is not complete, can be

considered on the lower level in this hierarchy as a set of functions in which some functions are realized and some are not. The second problem is that not all functionality requirements of users and stakeholders have the same importunity. Some functions are absolutely necessary, some have a great priority, some can be cosy, but not necessary, and some have only a minimal importance. Therefore, the suitability is often measured by the weighted functional coverage or using the multilevel functional coverage.

In the weighted functional coverage, the natural number w_j, so called weight, from some scale is assigned to all elementary function (function on the lowest level of the hierarchy of functions). If we have N function in this level and for each $j = 1, \ldots, N$ we put $s_j = 1$ if the function is realized and $s_j = 0$ if not, then the coverage can be computed as

$$\sum_{j=1}^{N} w_j \cdot s_j$$

The disadvantage of this measure is that if there exists a great number of offered functions with the lower importance, there is the danger that many small benefits can outbalance the serious disability. The world and life is, however, not always Archimedean and therefore this measure can be sometimes also problematic.

In the multilevel coverage, we shall split all functions to M (usually from two to five) levels of the importance. On each level we compute the weighted or not weighted functional coverage. The result is the generalized measure, which is a M-part vector $(c_1, c_2, \ldots, c_M)$. For each component of this vector, the required value of the quality indicator has to be adjusted separately. Ordinarily for c_1 the maximum possible value is required. If we construct the weak order between several products according to this vector attribute, we compare first the first components of the related vector measures, if it is equal, then second, if also the second coordinates are equal, the third and so on. Let us mention that according the well known Birkhoff and Milgram theorem (see for example Krantz et al. 1971), this vector measure can be replaced by the traditional one number measure of the ordinal scale type.

Respondents addressed in the research have good experiences with the multilevel or weighted coverage as a suitability attribute and measure. The same multilevel or weighted principle is often used also for the modification of other measures recommended in the respective ISO/IEC 9126-2, -3 technical reports.

Accuracy was not the topic of evaluation of the respondents, except of one case, the programs for meteorological forecasts. In this case, the stability of matrices inversion algorithm for large matrixes are the subject of interest. For this special problem, the much more sophisticated measures, than the measures recommended in ISO/IEC 9126-2, were used.

Interoperability is the subject of interest of many users. The situation is similar as in the case of the suitability measures. The data format based consideration is performed more frequently than the user success attempt consideration. For evaluation, the weighted or multilevel evaluation is used.

Security is the important issue for many users. However, nobody uses for the evaluation of the security the ISO/IEC quality model and the measures recommended in the technical reports ISO/IEC 9126-2 and -3. Users prefer to use the special security standards; in the first place the Common Criteria Standard.

Compliance in functionality is out of interest of users in question. The same situation occurs in the compliance subcharacteristics of all other quality characteristics.

Reliability is after functionality the second most interesting quality characteristic for the respondents.

Maturity is evaluated by about 50% of respondents. The multilevel variants of the attributes "failure density", "failure resolution" and "mean time between failures" are mostly used. The failures are classified usually into three levels.

(1) The crash – when the failure causes a serious damage, for example a loss of important data.
(2) The breakdown – when the functionality is completely refused and the user task is suspended, but without serious additional damages.
(3) The cutback – when the functionality is preserved, but on the limited restricted level of performance. The measures based on the potential number of failures, predicted using a reliability growth estimation model, such as "estimated latent failure density" or "estimated latent fault density" not used by respondents.

Failure or fault density measures, which divide the number of failures or faults by the product size, are used by about 50% of software developers, but mostly not for the evaluation of product quality, but yet for the valuation of developers process and staff rating. Of course the problem is in the definition of the term "program size". The assessment of the size by number of lines of source code (so called measure LOC) seems to be applicable only with great problems, because the complexity of the problem, programming languages or CASE tool used and developer's environment can be for the size more important than the physical length of the code.

Majority of the software developers use the measures "test coverage" and "test overcome" in various stages of the product life cycle. In the case, of the test coverage measure, the problem is to distinguish the detached scenarios for which the individual test is required.

Fault tolerance usually is not evaluated by any special attributes and its measures, such as "failure avoidance" and "incorrect operation avoidance", is investigated. The reaction of the system to incorrect data input or/and invalid operation during the program execution is assessed as a part of the system functionality. Operation with error input is considered as a part of functionality requirements. Different risk levels that denounce from the software fault evoked failure are diversified by the classification of failures to appropriate levels (usually as crash, break-down and cut-back).

Recoverability is sometimes evaluated by the attribute "availability", defined as a ratio

$$T_o/(T_o+T_r)$$

where T_o is an operation time of the system and T_r the time consumed to repair it. It is recorded by many users and considered as a major reliability indicator. For the "down time" attribute, the maximum down time variant is considered as more interesting as the mean down time variant. The same is true for the restart time. The pessimistic estimation is considered as a more important attribute.

Usability is rarely assessed and evaluated using the objective measurement process.

Understandability of the product and its attributes like "completeness of description" and "demonstration availability" and "function understandability" is sometimes estimated indirectly during the functional implementation coverage measurement.

Learnability is measured only using the time required to learn the function and operation to perform the task. If the system is able to perform an extensive set of functions, obviously two values are interesting for users. The time necessary to perform the base set of all-important functions and the time necessary to perform all special functions and take the advantage from the utilizing all advanced systems features. Some users also estimate the help system, but in majority only using a subjective estimation in the ordinal scale measure scale type estimation.

Operability is evaluated in the similar way as the learnability above. The discrepant measures proposed on ISO/IEC 9126-2 and -3 are not used. The features like automatic error input correction and/or default value availability are mostly considered as problematic and dangerous with regard to the reliability of the system. The user operation time is measured by the actual time needed to perform the task. For reiterative tasks, the pessimistic estimation of the maximum is considered as a more interesting attribute than the mean or average time.

For the *attractiveness,* respondents sometimes use only the subjective ordinal scale type estimation.

Efficiency is mostly out of the interest of developers and users inquired.

Time behaviour was interesting for the Internet web products with potentially many parallel users, working at the same time. For such product providers, the "response time" is considered as a key parameter of the system. In the case when the anticipation of loading for such a system is disproportionate, the pessimistic worst case "maximal response time" is considered as a more interesting measure than the mean response time. However, the objection of measure users is that the response time problem cannot be adequately described using one number only. More adequate is to measure this time by a function in which the free variable is the number of on line users and the value the response time.

The turnaround time was used as efficiency attribute only by two users, working with the special software realizing the sophisticated mathematical algorithms to solve extensive systems of difference equations.

None of our respondent applied *resource utilization* measures recommended in the ISO/IEC 9126-2 and -3.

Maintainability is considered as an important quality characteristic by developers and also by users. However, only few developers use the internal measures recommended inthe ISO/IEC 9126-3. The reason probably is that for the developer's process softwarehouses use various integral methodologies that contain their own inbuilt tools for the maintainability control. These tools are concentrated on the documentation of the project and the document flow during the stages of the project life cycle. Therefore, we have to concentrate on the external measures applied mostly by software users and system integrators.

Analyzability is measured only by few respondents. The attribute "failure analysis time" is used, still not as a fraction, but as a time at which the causes of failure are found out, only. The same variant of the proposed measure is between whiles used for the "diagnostic support" attribute.

Changeability is measured by some respondents, using the time spent to implement changes in the new version or modification of the software. This measure is a variant of some measures recommended in the ISO/IEC 9126-2.

Some respondents evaluate *stability* by recording the failure frequency in consequent versions or modifications of software and computing the ratio of these frequencies. It is conform to the attribute "less encountering failures after change", defined in the ISO/IEC 9126-2. Some users prefer for the stability evaluation the attribute of the main time between two consecutive version unfreezing and deliveries. For such users, there exists some optimal time, which is considered to be the best. Very frequent version changes are considered as annoyance. The large interval between versions is considered as an incompetence to repair faults in software.

Some respondents evaluate *testability*, still not according to the measures recommending in the ISO/IEC 9126-2, but only recording simply the time spending by the test run after the implementation of the new version of software.

Portability is evaluated by developers only in two projects. This evaluation was limited to the problem of application of the software design that works on the Microsoft Windows and Unix-like operation systems environment.

Adaptability is not evaluated by any respondent. No users try to adapt software to the different environment itself.

Installability is considered by respondents usually as a part of functionality or efficiency. The measures recommended in the ISO/IEC 9126-2 are not used.

Replaceability some users evaluate using the time necessary for spending for data migration to the new environment and the user effort spending to the adaptation of the software environment. The measures recommended in the ISO/IEC 9126-2 were not used.

Co-existence is also considered by respondents only as a part of the functionality subcharacteristic interoperability, or as a part of the reliability characteristic. The measure "concurrent multiple software use with less constrains", recommended in the ISO/IEC 9126-2, is based in fact on the number of failures.

DISCUSSION

The research demonstrates that the level of software product quality use in the Czech Republic is relatively low, comparable with advanced countries. The information technology market is more the market of supplier than the market of purchaser. Therefore the standards for developing process are more popular compared to the standards for product evaluation from the users point of view.

There are not enough experiences with product quality standards. Quality is often regarded only as functionality. Acquires are only seldom able to formulate the exact requirements which qualify their real needs. Acquires often make their decision concerning the product choice using different criteria than the product quality.

The quality attributes and measures recommended in the ISO technical report represent an extensive and blind set and it is not easy to realize a feasible choice of attributes in the concrete situation.

CONCLUSION

The described research shall be completed by the research in other countries, with different experience in information technology market and different vendor – buyer culture and the results shall be summarized. As the author knows, the similar research is realized in Korea. After that, the recommendation for a relative small and lucid set of the recommended software product quality attributes shall be given. The selected attributes and measures, probably in two levels as based and optional, shall be integrated into the now building up SQuaRE ISO/IEC 250xx series of standards. Without the responsible choice of attributes and measures, the SQuaRE project cannot be resultful and other branches of the ISO/IEC 250xx series can be only scholastic unhelpful and void ideas.

REFERENCES

IS ISO/IEC 9126 (1991): Software product evaluation – Quality characteristics and guidelines for their use. Former ISO/IEC standard.

IS ISO/IEC 9126-1 (2000): Information Technology – Software product quality – Quality model. ISO/IEC standard.

TR ISO/IEC 9126-2 (2000): Information Technology – Software product quality – External metrics. ISO/IEC technical report.

TR ISO/IEC 9126-3 (2000): Information Technology – Software product quality – Internal metrics. ISO/IEC technical report.

TR ISO/IEC 9126-4 (2000): Information Technology – Software product quality – Quality in use metrics. ISO/IEC technical report.

IS ISO/IEC 14598-1 (1998): Information Technology – Software product evaluation – General Overview. ISO/IEC standard.

IS ISO/IEC 14598-2 (1998): Information Technology – Software product evaluation – Planning and management. ISO/IEC standard.

IS ISO/IEC 14598-3 (1998): Information Technology – Software product evaluation – Process for developers. ISO/IEC standard.
IS ISO/IEC 14598-4 (1998): Information Technology – Software product evaluation – Process for acquirers. ISO/IEC standard.
IS ISO/IEC 14598-5 (1998): Information Technology – Software product evaluation – Process for evaluators. ISO/IEC standard.
IS ISO/IEC 14598-6 (1999): Information Technology – Software product evaluation – Documentation of evaluation modules. ISO standard.
IS ISO/IEC 15939 (2001): Information Technology – Software measurement process. ISO/IEC standard.
Vaníček J. (2006a): Software and data quality. Agric. Econ. – Czech, *52* (3): 138–146.
Vaníček J. (2006b): Software quality requirements. Agric. Econ. – Czech, *52* (4): 177–185.
Krantz D.H., Luce R.D., Suppers P., Tversky A. (1971): Foundation of Measurement. Vol. 1: Additive and Polynomial Expressions, Academic Press Inc., San Diego.

Arrived on 15th December 2006

Contact address:

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