Computer Science and Information Systems 11(4):1271–1289 DOI:10.2298/CSIS140815077P

# XML Schema metrics for quality evaluation

Maja Pušnik[1], Marjan Heričko[1], Zoran Budimac[2] and Boštjan Šumak[1]

[1]Faculty of Electrical Engineering and Computer Science, Smetanova ulica 17,
2000 Maribor, Slovenia
{maja.pusnik, marjan.hericko, bostjan.sumak}@um.si
[2] Faculty of Sciences, University of Novi Sad, Trg D. Obradovića 4,
21000 Novi Sad, Serbia
zjb@dmi.uns.ac.rs

**Abstract.** In XML Schema development, the quality of XML Schemas is a crucial issue for further steps in the life cycle of an application, closely correlated with the structure of XML Schemas and different building blocks. Current research focuses on measuring complexity of XML Schemas and mainly do not consider other quality aspects. This paper proposes a novel quality measuring approach, based on existing software engineering metrics, additionally defining the quality aspects of XML Schemas using the following steps: (1) definition of six schema quality aspects, (2) adoption of 25 directly measurable XML Schema variables, (3) proposition of six composite metrics, applying 25 measured variables and (4) composite metrics validation. An experiment was conducted using 250 standard XML Schemas collected from available e-business information systems. The results illustrate the influence of XML Schema's characteristics on its quality and evaluate the applicability of metrics in the measurement process, a useful tool for software developers while building or adopting XML Schemas.

**Keywords:** XML Schema, metrics, quality variables, Quality index, evaluation, validation.

## 1. Introduction

XML Schemas, recommended by W3C (World Wide Web Consortium) [1], are an important part of information system design and their quality affects several architecture levels and steps during the development life cycle. Information about the quality of a developed or adopted XML Schema is useful information and an indicator of the information system's quality. Appropriate metrics are needed in order to evaluate the quality of XML Schemas. Existing software-related metrics are mostly applied regarding software complexity and (less often) during quality measurements. Metrics' results are no longer a strategic advantage for software developers but a necessary indicator, identifying poor quality of well-formed XML Schemas in terms of inappropriate structures and bad practices, which can have expensive and long-term influences on software application [9]. Software quality needs to be evaluated as early as possible, as in those cases, where modifications are needed, additional iterations increase cost, time, and effort. [9].

XML Schemas are broadly used in most e-business companies [18, 17]. Using XML is included at various levels of contemporary multitier information solution

architectures: web service interface definitions, data models, specification of business cooperation protocols between different companies, etc.

Although addressing mainly the structure of XML documents, XML Schemas can also restrict the contents of elements and attributes, thus creating a more controlled environment for data definition [23]. By definition, they are extensible, flexible, and reusable in all kinds of environments, providing a fundamental technology for e-business infrastructure. If done properly, an XML Schema's quality is reflected in time and cost reduction during production, creating lower-cost services, improved user experience, higher flexibility, shorter time to market, and many other features. The mentioned advantages play important roles in enabling high performance of the end product (information system), and operating with XML documents and XML Schemas [15, 2].

The main objective of this paper was to identify and evaluate a metric set, suitable for measuring and evaluating quality of XML Schemas. The metrics must address aspects of structure and content, adopting the legacy of existing software metrics. We aimed to provide answers to following research questions:

- (R1) which XML Schema properties influence its quality or lack of it, and
- (R2) is there a correlation between the XML Schema's complexity and the assessed XML Schema's quality?

The research questions were addressed using several complementary research methods, including a literature review and interviews with a number of XML Schema experts, providing an insight into XML Schema problem areas. In order to measure XML Schema quality and to establish correlation between XML Schema's properties (later translated into measured variables), a set of representative (standard) XML Schemas was provided, gathering all possible (descriptive) data. Analysing the set of selected XML Schemas provided problem area identification and an insight into the efficiencies of existing XML Schema metrics. In addition, an examination of existing metric systems was made and a new theoretical approach for evaluating the quality of an XML Schema was presented. The novel set of metrics, proposed in the study, was evaluated on a set of XML Schemas in the field of e-business and the integration of complex business information systems. For quality measurement purposes, quality aspects were defined by addressing different views of an XML Schema's needs and demands.

The paper is organized in six sections. The first two introduce the research area and related work, exposing the lack of research in the field of measuring the quality of XML Schemas. The third section presents the XML Schema's quality aspects and explains the preliminary research, conducted in order to define the quality aspects of XML Schemas. In the section that follows, the proposed approaches to measure the quality aspects of XML Schemas using metrics are presented. Section five provides interpretations of the XML Schema's quality metrics evaluation results. Conclusions are provided in section six.

## 2. Related Work

A systematic literature review resulted in finding over 200 related scientific and professional papers, addressing quality in XML Schemas and in software applications in

general within the time scope of 15 years. There had been few attempts to evaluate and measure XML Schemas, however the number of evaluated XML Schemas was low and the quality aspect of XML Schemas was mostly one-dimensional, focusing only on complexity. Several metrics were summeded-up by Zhang in [24]. Significantly related work was done by McDowell, Schmidt and Yue in [11] and Narasimhan, Hendradjaya in [14], where attempts to measure XML Schemas and software in general were made. Finding suitable metrics for XML technologies has been addressed in other papers, however their backgrounds being mainly software metrics have been inapplicable for all the needs of an XML Schema's quality measurements, which is the main focus of this paper. Complexity measuring was researched in [20] by addressing larger and more complex applications. The authors focused on analysing the fractal dimensions of software networks, emphasising scalar numbers as characteristics of the whole system, presenting a synthetic metric for the complexity and quality of software systems. The paper attempts to address the technique of separating XML Schemas into fragments and trying to indicate each aspects importance to an XML Schema's quality.

Metrics for measuring XML Schema quality are addressed in [4] and [16]. The authors in [4] addressed the metrics for evaluating an XML Schema's structure complexity, pointing out entropy. The proposed metric was empirically evaluated based on a case study. While applying the metric, the authors excluded the variety of preferences for XML Schemas based on their purposes or domains. The authors in [16] shared the approach to measuring an XML Schema's quality from the previous paper; however they considered the structural aspect as well as the aspect of content. Their metric enables user modification of quality measurement settings. In our paper we take a step forward and include several other aspects of an XML Schema's quality although without providing user modification possibilities.

Improving an XML Schema's quality is the focus in [19], addressing the problem of changing the existing XML documents (according to XML Schema changes). In order to present the problems, each change has an assigned cost and preference weight. A measurement is proposed for each XML Schema's quality aspect. Assigning properties and rules in XML Schemas is addressed in [13], focusing on information system integration and the role of XML Schemas in the integration process. In order to ensure standard data exchange quality, naming and design rules are used, thus assuring consistency, readability, and reuse of XML Schemas. The afore-mentioned attributes are applied in aspects of an XML Schema's quality in our paper as well. Furthermore, the evaluation of standard quality in [13] is a simplified version of the evaluation proposed in our paper. The structures of XML Schemas are stressed in [22], searching for similarities between XML data, comparing data based on XML data structure. However, the method is more adapted to DTDs (Document Type Definition) than XML Schemas. The DTDs, as a behindhand technology, are excluded from this paper's research focus.

The aspect of complexity of XML Schemas was addressed in [5], where the authors proposed metric for measuring XML Schema complexity based on an XML Schema's structure. The metric was empirically evaluated based on 65 publically-available XML Schemas. The paper is focused mainly on structure and less on its contents and other aspects of quality, as also proposed in our paper. The complexity aspect is also addressed in [3], based on the internal structures of XML Schemas and providing validation variables and a measurement method for calculating complexity. Similar is done in [7]; measuring quality based on data control and ISO standard ISO/IEC 15939.

The above mentioned papers lack a holistic approach to quality measurement, as provided in our paper, however do depict several variables, important for XML Schema quality evaluation and measurement and some are also applied in this paper.

## 3. XML Schema quality aspects

Based on literature review and interviews, XML Schemas are often built irrationally, satisfying minimal requirements of syntactic correctness and content sufficiency. Existing metrics only partially address the problem based on existing solutions known in software engineering, not addressing the problem of an objective XML Schema quality evaluation. The results of the literature review in XML Schemas measuring fields resulted in identification of several metrics, applied to XML Schema evaluation, however extracted mainly from software engineering measurement methods, focusing heavily only on the complexity aspect. In order to include a variety of variables addressing complexity and quality, different fields on quality measurement were identified. In addition, the organisational structuring of XML Schemas was addressed and a classification based on the number of external XML Schemas was included. The following three schema types were defined: (1) independent primary XML Schemas without included or imported external XML Schemas, (2) dependent primary XML Schemas with a root element, some or all data types imported and (3) secondary XML Schemas – without elements and solely data type definition, used by other primary XML Schemas.

As existing metrics, found in the literature review, addressed mostly quality evaluation parameters of general software, a preliminary research was conducted in order to identify building blocks and structure concepts that create difficulties, specifically in the XML Schema life cycle, or jeopardize an XML Schema-related end-product's quality. Interviews as a supporting research method with 30 experts on XML technologies provided results mainly addressing structural problems. Their answers were a foundation for further research, modelled within a theoretical research model and a set of quality metrics was provided.

Structured interviews with experts allowed several open answers, enabling experts to express their own opinions, only indirectly connected to the question. Each group of questions provided insight into possible problems areas regarding the quality of XML Schemas. Experts were able to describe their own thoughts and problems regarding XML Schema usage. In the following subsections the answers are summarised and interpreted.

- *Subset 1: XML Schemas quality influencing related solutions* - Focusing on the relationship of XML Schemas with other technologies and how the quality of an XML Schema affects that correlation, investigating whether a poorly developed XML Schema causes difficulties when XML data is being read or reorganised. An enquiry was made as to what extent developers' use generating tools and skip creating schemas by hand and how such practice affects the qualities in their opinions.
- *Subset 2: XML Schema quality affected by building constructs (elements and attributes)* - Quality is affected by naming problems and conflicts of elements and attributes characteristics. Do experts actually have problems when naming elements,

do they tend to rush and not re-think a consistent naming style, creating inconsistencies and building a schema in a redundant (not minimal nor optimal manner), thus making re-use less available.

- *Subset 3: Quality challenges in existing XML Schemas* - Experts' perceptions of how practical XML Schemas actually are and whether the existing ones need to be modified. Do they have problems with reading schemas, even authorial after a period of time, do they regain a good enough overview of the schema hierarchy when complexity arises, how high the reusability of schemas or their parts are, and does documentation help. The assumptions were made as to whether the complexity decreases the XML Schema's general quality and loss of overview.
- *Subset 4: Quality of XML Schema building blocks* - Expert's perception of how well they take advantage of existing building blocks in XML Schemas, do they feel that using them adds to XML Schema quality, flexibility, simplification and other aspects of alleged quality.

### 3.1. Quality aspects analysis

The qualitative data was organised using the analytical tool QDM Miner and figure (Fig. 1) presents the frequencies of a certain problem area that mostly occurred. The Figure shows that those areas having the most problems included poor quality of generated schemas, definitions of own data-types, severe conflict and inconsistencies risks, need for re-use and other.

Propositions, the experts often pointed out in Fig. 1, were transformed into evaluation fields, addressing XML Schema quality (Fig. 2). The experts address the problematic lack of clarity in the structure of larger XML Schemas. Within XML Schemas experts warn of excessive usage of attributes in favour of more manageable and accessible elements; attributes should be used only in cases of unique or constant values.

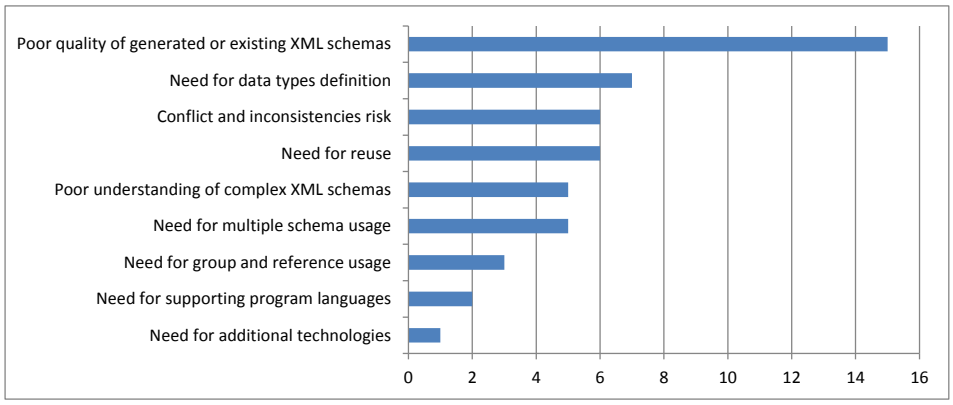


**Fig. 1**. Expert's expressed problem areas

The use of external XML Schemas was reportedly low; however experts admitted the usefulness of the data/types division aspect. The experts' advice was included in metric 1 (structure in Fig. 2). In order to enhance clarity, using global elements and attributes is

advised, enabling flexible XML Schema structure. Using groups, which reasonably combine related and connected building blocks, additionally aids the achieving of clarity regarding XML Schemas. The need of a higher level of using groups was addressed in metric 2 (clarity in Fig. 2). The common opinion was that using groups is better in comparison to using global complex types, composed from other elements, and mostly insufficiently used and exploited. The advice is included within the third metric 3 (optimality in Fig. 2). Additionally, the inevitably necessary building blocks and data types must be used in an economical manner, minimising the oversizing problem of XML Schemas and creating them as less complex. The minimalistic concept was addressed within metric 4 (minimalism in Fig. 2). While application of global elements can be problematic, the more common difficulty is identification of the root element. Within such scenarios the developer must make sure that the global declaration is really necessary due to the re-use or other reasons. However, the number of references on elements must be higher or equal to the number of existing global elements. The re-use aspect is addressed in metric 5 (re-use in Fig. 2). The problem of flexibility that opens up several possibilities for mistakes was also addressed and is included in metric 6 (flexibility in Fig. 2). The Quality index is a term, used for the combined qualities with all aspects equally included, and is a result of several composite metrics.


**Fig. 2.** Quality aspects of XML Schemas

## 4. Approach Overview

The quality aspects of XML Schemas were defined based on preliminary research. Each quality aspect was presented through measurable variables within a composite metric and validated based on several representative XML Schemas. 25 variables were measured on a set of 250 standard XML Schemas within different fields (domains) of use. The schemas were attained through available search portals (Google) filtered by standard schemas in 2013. The variables were included within 6 proposed metrics: (M1) structure, (M2) clarity, (M3) optimality, (M4) minimalism, (M5) reuse, and (M6) flexibility. According to their relevance within each quality aspect, the variables were used in metrics and are presented in table (Table 1). A cross X is present when a variable is included.

Variables were measured for each attained XML Schema, often receiving non-standard values from 0 to over 1000. In order to compare individual variables, standardisation of values was conducted. All variables that were used within metrics and their results were transformed into a scale from 0 to 1, where 0 represented the least desirable value for each variable and 1 the most desirable value. Transformations were based on linear programming rules, assuming that the growth relationship was always linear. Each of the variables underwent a process of standardisation using one of the scaling techniques:

a) The most desirable outcome of the variables is to be as minimal as possible, therefore the minimal value holds the estimation 1 and it linearly goes to 0.
b) The most desirable outcome is to be of average value and holds value 1; the extreme minimal and maximal values hold value 0.
c) The most desirable outcome of the variables is to be as maximal as possible, therefore the maximal value holds the estimation 1 and it linearly drops to 0.

**Table 1.** Variables and quality aspects

| | Structure | Clarity | Optimality | Minimalism | Reuse | Flexibility |
|---|---|---|---|---|---|---|
| $N_{an}$ = number of annotations | X | X | | X | | |
| $N_{ri\_all}$ = number of external XML Schemas | X | | | | X | X |
| $N_E$ = number of elements | X | X | X | X | X | X |
| $N_{E\_g}$ = number of global elements | | X | | | | |
| $N_{E\_l}$ = number of local elements | | | X | | | |
| $N_{E\_s}$ = number of simple elements | | | X | | | |
| $N_{E\_gc}$ = number of global complex elements | | | X | | | |
| $N_{E\_gs}$ = number of global simple elements | | | X | | | |
| $N_{at}$ = number of all attributes | X | X | X | X | X | X |
| $N_{at\_l}$ = number of local attributes | | | X | | | |
| LOC = lines of code | | | | X | | |
| $N_g$ = number of all groups | | X | | | X | X |
| $N_{E\_group}$ = number of element groups | | X | | | | X |
| $N_{A\_group}$ = number of attribute groups | | X | | | | X |
| $N_{re\_all}$ = number of references on elements | | | | | X | X |
| $N_{ra\_all}$ = number of attribute references | | | | | X | X |
| $N_{rg\_all}$ = number of group references | | | | | X | X |
| $N_r$ = number of restrictions | X | | | | | |
| $N_{t\_i}$ = number of derived data types | X | | | | X | |
| $N_t$ = number of all data types | | | X | X | X | X |
| $N_{rt\_all}$= number of all used data types | | | | X | X | X |
| $N_{t\_s}$ = number of simple data types | X | | | | | |
| $N_{t\_c}$ = number of complex data types | X | | | | | |
| $N_{E\_U}$ = number of unbounded elements | X | | X | | | X |

The desired outcome was defined based on expert' opinions and the standpoints of other authors as presented in the literature review. In general, the paper favours using elements over attributes, discourages the use of several global elements or definitions of unused elements and emphasises the global definition of simple or complex data types. Standardised variable values were used within defined metrics; addressing the proposed six aspects of an XML Schema's quality as presented in Fig. 2. They are further explained in the following sections.

### 4.1. Quality of structure

Structure measuring of XML schemes for calculating the complexity and quality was done by McDowell and others in [11] as well as Burris in [6]. The authors present a composite of metrics, taken mainly from "quality model" ISO standard, combining them within a single formula. Each variable is further multiplied with a not-interpreted constant, and the values are not standardised. During our research, we analysed and partly used the given formula in our calculations of quality. In this paper we redefined metrics into a composite metric, as presented in equation 1. The variable short names are explained in the table (Table 1).

$$M_1 = N_{ri\_all} + \frac{N_E}{N_{at}} + \frac{N_r}{N_{t\_s}} + \frac{N_{t\_s}}{N_{t\_c}} + \frac{N_{an} + N_{t\_i} + N_{E\_U}}{N_E} \quad \text{(Eq.1)}$$

### 4.2. Quality of clarity

The importance of well-documented and a clear, easy-to-read and understand XML Schema is addressed in the following relationship: the number of annotations ($N_{an}$) depending on the number of items ($N_E$) and attributes ($N_{at}$) illustrates the documentation part of XML Schemas, presuming that more information about the building blocks increases the quality. Using groups of elements or attributes also creates a more organised and clear overview of XML Schema. The metric is presented in the following equation 2.

$$M_2 = \frac{N_{an}}{N_E + N_{at}} + \frac{N_{E_{group}}}{N_E} + \frac{N_{A_{group}}}{N_{at}} + \frac{N_g}{N_E} \quad \text{(Eq.2)}$$

### 4.3. Quality of XML Schema optimality

The term optimum is the most favourable condition or amount of building blocks for obtaining a desired result. The metric addressing quality in the optimal use of XML Schema building blocks searches the more optimal ratio between different building concepts, combining several variables, indicating the optimal structure of XML Schemas. The variable usage discloses whether the in-lining pattern, the least preferable in XML Schema building, has been used. The metric is presented in equation 3, where the following relationships were addressed: ratio between local and all elements, ratio between local attributes and all attributes, ratio between global and complex elements of

all the complex elements, ratio between global and all the simple elements, ratio between data types and all elements and attributes, ratio between number of groups and global complex elements, ratio between number of unbounded elements and all elements.

$$M_3 = \frac{1}{7}\left(\frac{N_{E\_l}}{N_E} + \frac{N_{at\_l}}{N_{at}} + \left(1 - \frac{N_{E\_gc}}{N_E - N_{E\_s}}\right) + \frac{N_{E\_gs}}{N_{E\_s}} + \frac{N_t}{N_E + N_{at}} + \frac{N_g}{N_{E\_gc}} + \left(1 - \frac{N_{E\_U}}{N_E}\right)\right) \quad \text{(Eq.3)}$$

### 4.4. Quality of minimal usage of building blocks

In this metric, a combination of variables is presented, indicating the minimum XML Schemas' building blocks, where the concept of minimalism is defined as the level where we can anticipate that there is no other smaller full descriptive set of building blocks (equation 4).

$$M4 = \frac{N_{an} + N_E + N_{at}}{LOC} + \frac{N_{rt\_all}}{N_t} \quad \text{(Eq.4)}$$

### 4.5. Quality of XML Schema reuse

Equation 5 was inspired by authors Washizaki and Fukazawab in [21], where a summed-up definition of a metrics set for measuring the re-use of software was displayed. This metric includes variables that allow the re-use and are inherently global. We included the following variables:

$$M5 = \frac{N_{re\_all} + N_{ra\_all} + N_{rg\_all} + N_{ri\_all} + N_{rt\_all} + N_{t\_i}}{N_E + N_{at} + N_g + N_t} \quad \text{(Eq.5)}$$

### 4.6. Quality of an XML Schema's flexibility

The definition of the equation was taken from the idea of density of software components [6], where the authors calculated the density of other software segments and the frequencies of interactions between them (operations, classes, modules).We have adjusted and simplified the formula into the following equation 6:

$$M6 = \frac{N_{E_{group}} + N_{A_{group}} + N_g + N_{re_{all}} + N_{ra_{all}} + N_{rg_{all}} + N_{ri_{all}} + N_{An} + N_{rt\_all} - N_{E\_U}}{N_E + N_{at} + N_t + N_g} \quad \text{(Eq.6)}$$

### 4.7. Quality index

The presented metrics address six aspects of an XML Schema's quality (structure, clarity, optimality, minimalism, re-use, flexibility). Metrics include 25 quality and complexity variables, defined based on related research. A composite metric is given in equation 7 equally combining all six metrics and presenting a general Quality index.

$$Q_i = 1/6( Q_1+ Q_2+ Q_3+ Q_4+Q_5+Q_6) \quad \text{(Eq.7)}$$

## 5. Evaluation and analysis

The proposed metrics were applied on a set of 250 XML Schemas, gathered through an internet search. Each XML Schema was validated as to whether it was a well-formed and a standardly used schema. We developed a supporting application, a tool enabling validation and measuring of XML Schemas according to our pre-set variables, included in all six aspects of XML Schema quality. The metric's validity was assured using action ability and appropriate continuity [12] as a composite metric defined for all measured values [10] and would enable a software developer to conduct decision-making based on the calculated Quality index [8]. In case the Quality index is low, six aspects' quality levels indicate the problem areas, needing attention.

Each metric was measured individually and then combined with all metrics within the Quality index. The poorest XML Schema's achieved Quality index was 23% and the highest 70%. Individual metrics had different value spans: Metric 1 (M1) included values from 8% suitability to 67%, M2 from 0% to 100%, M3 from 2% to 46%, M4, M5 and M6 from 0% to 100%. The first metric (the structure aspect of a quality) is mainly focused on its complexity. The relationship between Quality index and M1 (complexity metric) was extracted from data (Fig. 3). The graph is dispersed; however there is evidence that XML Schemas with the higher qualities have higher levels of complexities, as do the schemas with the lower qualities. Average XML Schemas with average quality evaluation had relatively lower complexity levels.

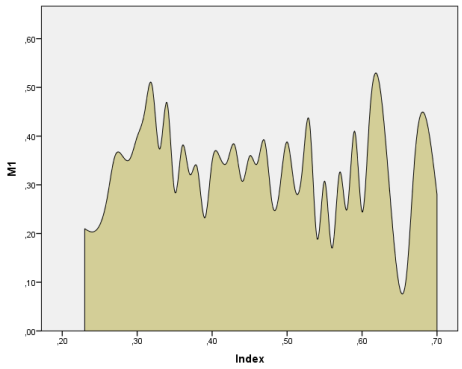


**Fig. 3.** Complexity and quality of XML Schemas

Regression analysis was used to evaluate the results from 250 XML Schemas. 9 out of 25 measured variables had significant impacts on the final quality evaluations of XML Schemas. Two of them had negative influences (lines of code and number of local elements) and seven of them positive influences (XML Schema type, number of derived data types, number of attribute groups, number of included XML Schemas, number of global simple and complex elements, number of annotations). The results are presented in the following sections.

Tables in the following subsections present the influences of specific variables and correlations between the variables and the measured Quality index. They are standard outputs of data interpretation in SPSS Statistics, software for statistical analyses . The 'Model Summary' part of the table presents the R Square value in the third column, indicating the percentage of XML Schemas, where their Quality index can be explained by a specific variable. The "ANOVA" part of each table with the "Sig." column indicates the significance of the variable, which value must be under 0.05 in order for the variable to be significant.

### 5.1. Influence of documentation

Table 2 presents the significant influence of documentation in form of annotation within the measured Quality index. R square value indicates that the model explained 12.4% of the variability regarding the measured XML Schemas. The results illustrated, that the higher amounts of annotations indicates the highest Quality index. Except for a few excessively documented XML Schemas having a higher level of quality, the majority of average XML Schemas had little or no documentation.

**Table 2.** Correlation between variables Annotation number and the Quality index

Model Summary

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | ,352[a] | ,124 | ,107 | ,876 |

a. Predictors: (Constant), Annotation number

ANOVA[a]

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 12,904 | 2 | 6,452 | 7,153 | ,001[b] |
| | Residual | 91,096 | 101 | ,902 | | |
| | Total | 104,000 | 103 | | | |

a. Dependent Variable: Quality index

b. Predictors: (Constant), Annotation number

### 5.2. Influence of building blocks

Table 3 presents the significant influence of variables regarding the number of elements on the measured Quality index. R square value indicates that the model explains 22.4% variability of the measured XML Schemas.

**Table 3.** Correlation between variables regarding the number of elements and the Quality index

Model Summary

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | ,474[a] | ,224 | ,215 | ,09728 |

a. Predictors: (Constant), Number of local elements, Number of global complex elements, Number of global simple elements

ANOVA[a]

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | ,687 | 3 | ,229 | 24,198 | ,000[b] |
| | Residual | 2,375 | 251 | ,009 | | |
| | Total | 3,062 | 254 | | | |

a. Dependent Variable: Quality index

b. Predictors: (Constant), Number of local elements, Number of global complex elements, Number of global simple elements

**Table 4.** Correlation between variables regarding the number of attribute groups and the Quality index

Model Summary

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | ,287[a] | ,082 | ,079 | ,10540 |

a. Predictors: (Constant), Number of attribute groups

ANOVA[a]

| Model | | Sum of Squares | Df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | ,252 | 1 | ,252 | 22,651 | ,000[b] |
| | Residual | 2,811 | 253 | ,011 | | |
| | Total | 3,062 | 254 | | | |

a. Dependent Variable: Quality index

b. Predictors: (Constant), Number of attribute groups

A high number of local elements results in high Quality index; the Quality index grows with number of global simple elements and attains higher results with a larger number of global complex elements. Using several global complex elements causes a

low Quality index, due to the problem of undefined roots. Extremely large (over 1000) and low values (under 10) were excluded. Table 4 presents the significant influences of attribute groups on the measured Quality index. R square value indicates that the model explained 8.2% variability for the measured XML Schemas. The results indicated that the higher the values of using attributes groups in XML Schemas, the higher the Quality index.

### 5.3. Influence of re-use regarding data types

Table 5 presents the significant influence of the number of all types and number of derived data types on the measured Quality index. R square value indicates that the model explained 3.1% variability of the measured XML Schemas.

The increase of Quality index depending on the number of all data types indicated that the larger the number of (derived) data types, the higher the Quality index. Using data types positively affects the Quality index.

**Table 5.** Correlation between variables regarding number of data types and the Quality index

Model Summary

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | ,175[a] | ,031 | ,023 | ,10854 |

a. Predictors: (Constant), Number of all data types, Number of derived data types

ANOVA[a]

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | ,093 | 2 | ,047 | 3,966 | ,020[b] |
| | Residual | 2,969 | 252 | ,012 | | |
| | Total | 3,062 | 254 | | | |

a. Dependent Variable: Quality index

b. Predictors: (Constant), Number of all data types, Number of derived data types

### 5.4. Influence of structure

Table 6 presents the significant influence of the XML Schema type on the measured Quality index. R square value indicates that the model explained 16.9% variability of the measured XML Schemas.

Based on regression analysis there were several more variables having significant impact on the Quality index: Lines of code and number of external XML Schemas. They were not further analysed as they represented the complexity aspect of XML Schemas. The following table (Table 7) presents all those variables with a significant influence on the final Quality index. The Beta value presents a positive or negative value added to the final score for each additional usage of the specific variable,

indicating that each additional external XML Schema raises the quality by 0.164 on the scale of 0 to 1.

**Table 6.** Correlation between XML Schema type and the Quality index

Model Summary

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | ,411[a] | ,169 | ,165 | ,09751 |

a. Predictors: (Constant), XML Schema type

ANOVA[a]

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | ,420 | 1 | ,420 | 44,210 | ,000[b] |
| | Residual | 2,073 | 218 | ,010 | | |
| | Total | 2,493 | 219 | | | |

a. Dependent Variable: Quality index

b. Predictors: (Constant), XML Schema type

**Table 7.** Variables effect on Quality index

| Predictor | | Beta |
|---|---|---|
| Number of external XML Schemas | $N_{ri\_all}$ | ,164 |
| XML Schema type | $N_{ST}$ | -,384 |
| Number of local elements | $N_{E\_l}$ | -,017 |
| Number of global complex elements | $N_{E\_gc}$ | ,308 |
| Number of global simple elements | $N_{E\_gs}$ | ,145 |
| Lines of code | LOC | -,012 |
| Number of attribute groups | $N_{A\_group}$ | ,133 |
| Annotation number | $N_{An}$ | ,130 |
| Number of derived data types | $N_{t\_i}$ | ,047 |

Additionally to correlations between variables, Table 8 describes correlation between metrics M3, M4 and M6, the aspects of optimality, minimal use of building blocks and flexibility.

In order to connect and combine all quality aspects based on significant variables within metrics, a theoretical research model (Fig. 4) was designed, addressing the following hypotheses:

- H1: XML Schema structure impacts significantly the XML Schema Quality index
- H2: XML Schema documentation impacts significantly the XML Schema Quality index
- H3: Optimal and minimal use of building blocks impacts significantly the XML Schema reuse
- H4: Optimal and minimal use of building blocks impacts significantly the XML Schema flexibility

H5: XML Schema reuse impacts significantly the XML Schema Quality index.

**Table 8.** Correlation between metrics M5 and M4

Model Summary

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | ,418[a] | ,175 | ,172 | ,24751 |

a. Predictors: (Constant), M3, M4

ANOVA[a]

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 3,284 | 1 | 3,284 | 53,612 | ,000[b] |
| | Residual | 15,499 | 253 | ,061 | | |
| | Total | 18,784 | 254 | | | |

a. Dependent Variable: M6

b. Predictors: (Constant), M3, M4

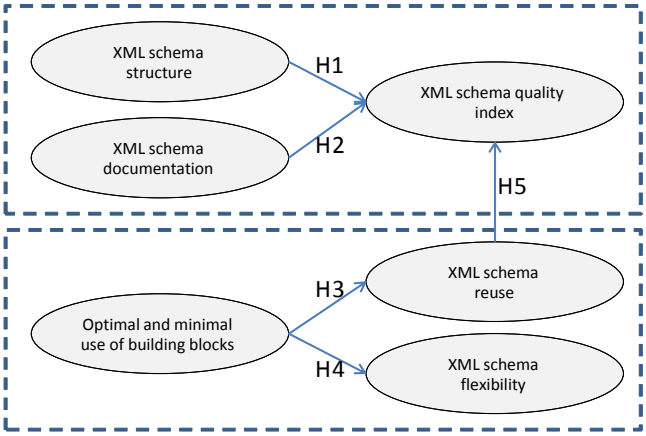


**Fig. 4.** Theoretical research model

Hypotheses were addresses based on research data and statistical analysis of data; however some limitations and threats were present. The results cannot be applied to all XML Schemas due to their differences in preferences within different domains. There is also difference in aspects importance within different domains, emphasising the importance of XML Schema categorisation, which was excluded within this paper and will be a part of our future research. The Quality index, included in this paper, provides a general evaluation of an XML Schema's quality however the separate metric evaluation enables a more specific insight into the aspects of an XML Schema's quality. In several cases, a high value for one metric indicated a low value for another metric; for example, minimalism and clarity were rarely met at the same time. Limitations of the Quality index support the need for further research and reflect the importance of using individual metric's results for guidance while building or adopting an XML Schema.

**Table 9.** Hypothesis evaluation

| | | |
|---|---|---|
| *H1: XML Schema structure impacts significantly the XML Schema Quality index* | Confirmed | Several of variables, included in quality and structure (M1) had a significant impact on the Quality index: number of external XML Schemas, annotation number and number of derived data types. |
| *H2: XML Schema clarity impacts significantly the XML Schema Quality index* | Confirmed | Clarity was measured in metric M2 and it included annotation number and number of attribute groups, variables with a significant impact factor on the Quality index. |
| *H3: Optimal and minimal use of building blocks impacts significantly the XML Schema reuse* | Confirmed | Optimal use was measured in M3 and it included following variables with a significant impact factor on the Quality index: number of local elements and number of global complex elements.<br>Minimal use was measured in M4, including the following variables with significant impact on the Quality index: annotation number and lines of code. |
| *H4: Optimal and minimal use of building blocks impacts significantly the XML Schema flexibility* | Confirmed | M3 and M4, measuring optimal and minimal use had a significant impact on M6, measuring flexibility for 17.5% of variances. |
| *H5: XML Schema reuse impacts significantly the XML Schema Quality index* | Confirmed | Reuse was measured with M6, including variables with a significant impact on the Quality index: number of external XML Schemas and number of derived data types. |

Following regression analyses of all XML Schemas the results are presented in Table 9. All hypotheses were confirmed, indicating the suitability of composite metrics definition. Although the included XML Schemas were standard schemas within different domains and had high rates of application, several of them were of extremely poor quality and affected the results of the Quality index.

## 6. Conclusion

The focus of this paper was on defining a full set of composite metrics for assessing an XML Schema's quality. Based on the results of a preliminary research, we defined six metrics, focusing on the more important aspects of XML Schema quality, measuring each building block/concept properties. For quality aspects: (1) structure, (2) clarity, (3) optimality, (4) minimalism, (5) re-use and (6) flexibility, separate metric sets were defined, composing an overall XML Schema Quality index. A tool for collecting XML

Schema properties and using the metric sets was developed, providing automation when evaluating the quality of developed or adopted XML Schemas.

Linear regression was utilised for validation of the proposed XML Schema's Quality index and for analysing the impact of used variables on the Quality index. When addressing the research question R1, a set of nine variables was identified as significant (influencing quality positively or negatively); including the use of external XML Schemas, the XML Schema's type, three variables regarding element application, number of attribute groups, lines of code and number of derived data types. All significant variables are in various ways included within separate metrics, addressing all quality aspects. The results of experimental evaluation of 250 XML Schemas provided evidence that there is a correlation between complexity of XML Schemas and their quality, where XML Schemas with the higher quality have a higher level of complexity, addressing the research question R2. We have established that XML Schema structure clarity and possibility of reuse significantly impact the Quality index, where optimal and minimal uses of an XML Schema's building blocks influences reuse and flexibility.

Objective, systematic, consistent and quick measurement of an XML Schema's quality provides a strategic decision-making and improvement in data organisation, as a standard mechanism (internal or global) for evaluating of XML Schemas' quality. Software metrics are a good basis for an XML Schema's quality measuring, however some accommodations are necessary according to their needs and demands (quality aspects). Results of this study however indicate that all aspects of quality cannot be met at the same time, and a high quality for one aspect consequently influences a low quality for another. Different domains of an XML Schema's usage differ in the priorities of quality aspects.

Our primary motivation was to build an extensive set of metrics, applying all aspects of XML Schema quality, including structure and contents. The latter was however difficult to define based on limited documentation of XML Schemas. Additional limitation of this study was also restricted access to full standard XML Schemas with all included files and perceived usefulness from actual users. Given the fact, that XML Schemas were not randomly selected nor can all XML Schemas users' demands and acceptations be predicted, the Quality index based on our research cannot be used as universal for all XML Schemas. The risk of missing literature or overlooked good practices is also present. Additional limitation is the restricted validation by not ensuring the suitability of metrics to all XML Schemas. However, we believe that this paper offers a basis for the further development of metrics in XML technologies and a supporting tool for developers while adopting XML Schemas.

In future work we will continue evaluation of standard XML Schemas, the metrics will be further validated and the set of XML Schemas will be enlarged. The evolution of XML Schemas will also be addressed by comparing the quality of different versions. Growth or decline of quality of XML Schemas' versions will be compared and examined. Applicability of defined metrics will be examined in more detail, focusing on success in practical examples within different fields of use, investigating the need for metrics adaptability according to the domain in which an XML Schema is used. Additionally, we will determine in which domains XML Schemas are widely used, which XML Schema quality aspects prevail in each domain and can the Quality index be designed to evaluate all needs of domain specific XML Schemas.

## References


1. Algergawy, A., Schallehn, E., Saake, G.: Improving XML Schema matching performance using Prüfer sequences. Data & Knowledge Engineering, Vol. 68, 728-747. (2009)
2. Atay, M., Chebotko, A., Liu, D., Lu, S., Fotouhi, F.: Efficient schema-based XML-to-Relational data mapping. Information Systems, Vol. 32, No. 3, 458-476. (2007)
3. Basci, D., Misra, S.: Complexity Metric for XML Schema Documents. Journal of Information Science and Engineering, Vol. 25, No. 5, 1405-1425. (2009)
4. Basci, D., Misra, S.: Entropy as a Measure of Quality of XML Schema Document. The International Arab Journal of Information Technology, Vol. 8, No. 1. (2011)
5. Basci, D., Misra, S.: Measuring and Evaluating a Design Complexity Metric for XML Schema Documents. Journal Of Information Science And Engineering, Vol. 25, 1405-1425. (2009)
6. Burris, E.: Hierarchical Nature of Software Quality, Programming in the Large. The Practice of Software Engineering (2012). [Online]. Available: http://programminglarge.com/hierarchical-nature-of-software-quality/ (current October 2014)
7. Caballero, I., Verbo, E., Calero, C., Piattini, M.: A data quality measurement information model based on ISO/IEC 15939. Proc. of the 12th International Conference on Information Quality, MIT, Cambridge, MA. (2007)
8. Fenton, N. E., Neil, M.: Software metrics: roadmap. Future of Software Engineering, Limerick, Ireland. (2000)
9. Khan, R. A., Mustafa, K., Ahson, S. I.: An Empirical Validation of Object Oriented Design Quality Metrics. J. King Saudi University, Vol. 19, 1-16. (2006)
10. Kitchenham, B., Pfleeger S. L., Fenton, N.: Towards a Framework for Software Measurement Validation. IEEE Transactions on Software Engineering, Vol. 21, No. 12, 929-944. (1995)
11. McDowell, A., Schmidt, C., Yue, K.: Analysis and Metrics of XML Schema. Proceedings of the International Conference on Software Engineering Research and Practice, SERP'04, Vol. 2, 538-544. (2004)
12. Meneely, A., Smith, B., Williams, L.: Validating Software Metrics: A Spectrum of Philosophies. ACM Transactions on Software Engineering and Methodology, Vol. 21, No. 24. (2012)
13. Morris, K. C.: A framework for XML Schema naming and design rules development tools. Computer Standards & Interfaces, Vol. 32, No. 4, 179-184. (2010)
14. Narasimhan, V. L., Hendradjaya, B.: Some theoretical considerations for a suite of metrics for the integration of software components. Information Sciences, Vol. 177, No. 3, 844-864. (2007)
15. Nayak, R., Iryadi, W.: XML Schema clustering with semantic and hierarchical similarity measures. Knowledge-Based Systems, Vol. 20, 336-349. (2007)
16. Pardede, E., Gaur, T.: On the Development of User-Defined Quality Measurement Tool for XML Documents. Information Systems development, Vol. 3, 213-221. (2011)
17. Rishel, W.: Does XML Schema Earn its Keep? The Gartner Blog Network (2011). [Online]. Available: http://blogs.gartner.com/wes_rishel/2011/12/31/ok-xml-schema-does-earn-its-keep-in-hl7 (current October 2014)
18. Sušnik, M.: V slogi je e-račun! Monitr Pro (2008). [Online]. Available: http://www.monitorpro.si/41040/praksa/v-slogi-je-e-racun (current October 2014)
19. Tan, Z., Zhang, L.: Improving XML Data Quality with Functional Dependencies. Lecture Notes in Computer Science, Vol. 6587, 450-465. (2011)
20. Turn, I., Concas, G., Marchesu, M., Tonelli, R.: The fractal dimension of software networks as a global quality metric. Information Science, Vol. 245, 290-303. (2013)
21. Washizaki, H., Fukazawab, Y.: A technique for automatic component extraction from object-oriented programs by refactoring. New Software Composition Concepts, Vol. 56, No, 1–2, 99–116. (2005)

22. Wojnar, A., Mlynkova, I., Dokulil, J.: Structural and semantic aspects of similarity of Document Type Definitions and XML Schemas. Information sciences Vol. 180, 1817-1836. (2010)
23. The World Wide Web Consortium (W3C): Standards, XML Technology, Schema, W3C Recommendations, W3C XML Schema Definition Language (XSD). XML Schema Working Group. [Online]. Available: http://www.w3.org/TR/2012/REC-xmlschema11-1-20120405 (current October 2014)
24. Zhang, Y.: Literature Review and Survey: XML Schema Metrics. University of Windsor. (2008)

**Maja Pušnik** is the corresponding author of this paper. She received her Ph.D. degree in computer science in 2014 from the University of Maribor, Slovenia. She is currently a teaching assistant at the Institute of Informatics, FERI, at the University of Maribor. Her main research interests include XML and related technologies, web application development, information system integration, orchestration and optimization of business processes, decision theories and operational research. She has appeared as an author and co-author in several peer-reviewed scientific journals. She has also presented her work at a number of international conferences.

**Marjan Heričko** is a full professor at the Institute of Informatics. He is the head of the Information systems laboratory and Deputy Head of the Institute of informatics. He received his PhD in Computer Science from University of Maribor in 1998. His main research interests include all aspects of information systems development, software and service engineering, agile methods, process frameworks, software metrics, functional size measurement, SOA, component-based development, object-orientation, software reuse and software patterns. Dr. Heričko has been a project or work co-ordinator in several applied projects, project or work co-ordinator in several international research projects and committee member and chair of several international conferences.

**Zoran Budimac** holds position of a full professor at Faculty of Sciences, University of Novi Sad, Serbia, since 2004. Currently, he is head of Computing laboratory and member of Faculty council. His fields of research interests involve: Programming Languages, Static software analyzers, Educational Technologies, and Software agents. He was principal investigator of more then 20 projects. He is author of 13 textbooks and more then 250 research papers most of which are published in international journals and international conferences. He is/was a member of Organizing and/or Program Committees of more then 70 international Conferences and is member of Editorial Board of "Computer Science and Information Systems Journal".

**Boštjan Šumak** received his Ph.D. degree in computer science in 2011 from the University of Maribor, Slovenia. He is currently an assistant professor at the Institute of Informatics, FERI, at the University of Maribor. His main research interests include contemporary applications architectures, service oriented architectures, e-services, object-oriented analysis and design, business process integration, XML and related technologies, web applications development, Web 2.0 & Web 3.0 and user experience. He has appeared as an author and co-author in several peer-reviewed scientific journals. He has also presented his work at a number of international conferences. In addition, he has participated in many national and international research projects.

*Received: August 15, 2014; Accepted: October 08, 2014*