Redefining Software Quality Metrics to XML Schema Needs

MAJA PUŠNIK, BOŠTJAN ŠUMAK AND MARJAN HERIČKO, University of Maribor
ZORAN BUDIMAC, University of Novi Sad

The structure and content of XML schemas, important and widely used document definitions, has a significant influence on the quality of XML data and XML technologies in general, therefore the quality of XML Schemas and accurate assessment of the quality is a fundamental research challenge in all fields of XML application. A good quality estimation of an XML schema can directly and indirectly lead to a higher efficiency of its usage, simplification of information solutions, efficient maintenance, and higher quality of data and business processes. This paper addresses challenges in measuring the level of XML schema quality by employing general software quality metrics; a set of holistically defined and document-oriented metrics is proposed. Proposed XML Schema quality metrics base on existing software metrics, adapted according to needs of XML schemas, addressing it mostly from a structural perspective.

---

1. INTRODUCTION

The primary role of XML schemas is the definition of XML data and supporting rules regarding the use of XML data, an important part of information technologies. XML schemas and related technologies present an important part of IT solutions in most Slovenian companies [Sušnik 2008], EU and the world [Rishel 2011]. Using XML has spread from the field of e-business and data exchange to data presentation into various levels of contemporary information solution architectures: (1) web service interface definitions, (2) data models, (3) specification of business cooperation protocols between different companies (their many uses are evident from different scientific and technical papers), etc.. Due to the widespread use, the question of XML schema quality is often open, particularly from the aspect of structure (and content) of XML schemas, which indirectly influence the quality of data that XML schema describes. Therefore measuring XML schemas quality is the basic research challenge in our paper. Solution of the problem (the composite of metrics) will directly or indirectly lead to greater efficiency in the use of XML schemes, simplifying IT solutions, facilitating maintenance, improving the quality of data and associated business processes. Ideally the metrics should apply the aspect of structure, content and domain, in which the XML schema is applied, however this paper will focus mostly on structural aspect, trying to take advantage of existing software metrics.

There have been several attempts to evaluate and measure XML schemas. Few of them are summed in [Zhang 2008]. Significantly related work was also done in [McDowell, Schmidt, Yue 2004] and [Narasimhan, Hendradjaya 2007], where attempts to measure XML schemas as well as software in general were made. The subject was addressed in other papers, not included in this overview, however the background are mainly software metrics, which do not necessary always apply needs of XML schema quality (and complexity) measurements.

Based on surveys and interviews, conducted within the University of Maribor and nearby companies, XML Schemas are often built irrationally in a manner, which satisfies the minimum requirements of syntactic correctness and content sufficiency. Existing metrics only partially address the problem basing

---

on existing solutions known in software engineering and not addressing the problem of an objective quality evaluation of an XML Schema. Dynamic creation and adaptation of XML schemas schedules and presents an additional research challenge that requires the use of new approaches and solutions, universal and specific according to a domain.

The aim of this paper is definition of a new theoretical approach for evaluating the quality of XML Schema, basing on the original concept of semantically related analysis of XML schemes and XML documents, by using a new set of metrics. The design correctness of the newly redefined metrics was confirmed on an expanded set of test data of already established XML schemes in the field of e-business and integration of complex business information systems. For quality measurement purposes we gathered quality parameters, addressing different aspects of XML Schema needs and demands.

This paper is organized into four chapters. After the presentation of this papers background and the description of included XML quality parameters, chapter two presents all aspects in metric types. Chapter 3 presents metric application and chapter four includes discussion of our present work and future plans.

1.1 XML schema quality parameters

The results of a systematic review of literature in the field of measuring XML schemas showed that several metrics were applied to XML schema evaluation, extracted mainly from the methods of software engineering measurements, focusing mostly on the complexity of XML Schemas. To include a variety of parameters addressing complexity and quality, we searched different fields on quality measurement. The first group of parameters was related to the structural characteristics of XML schemes (we included a survey, where all currently defined metrics are taken from several authors in [Zhang 2008]):

- *XML schema size,*
- *Number of XML nodes and annotations,*
- *Number of global and local element declarations,*
- *Number of global or local complex types definitions,*
- *Number of derived complex types, number of global and local definitions of simple types,*
- *Number of global or local definitions of models groups (groups),*
- *Number of global or local definitions of groups of attributes,*
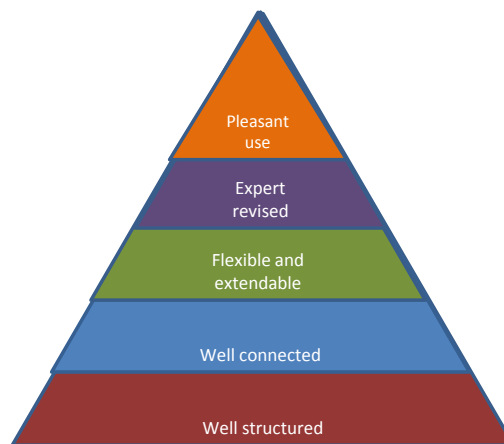- *Branch elements, the average cardinality of elements, etc.*

Fig. 1 Quality hierarchy in XML schemas

The typically software metrics parameters were extended with parameters form other quality measurement fields, specifically taken from standards ISO (ISO/IEC 9126 [McDowell, Schmidt, Yue 2004]), decision models theory [Burris 2012] and other papers [Zhang 2008]):

- *XML schemas functionality*
- *XML schemas simplicity*
- *XML schemas scalability*
- *XML schemas comprehensibility*
- *XML schemas re-use,*

- *XML schemas fullness,*
- *XML schemas integrability,*
- *XML schemas Flexibility,*
- *XML schemas Implementation,*
- *XML schemas Maintenance,*
- *Accuracy,*
- *Validity,*
- *Up to date,*
- *Minimalism,*
- *Consistency,*
- *Portability*
- *Security,*
- *Interoperability*
- *Reliability,*
- *Effectiveness,*
- *Visibility*

To determine the quality levels of XML schema usage, we borrowed Maslow's hierarchical nature needs, which can be applied to software and to all supporting technologies, presenting our interpretation in Fig. 1. The gathered parameters were organized into six groups, reflecting six identified XML schema needs respectively XML schema quality demands, meeting the three main XML schema demands: (1) good structure, (2) consistent contents, (3) compliant with domain. All parameters, contributing to XML schema quality and all aspects of quality are combined in Fig. 2.
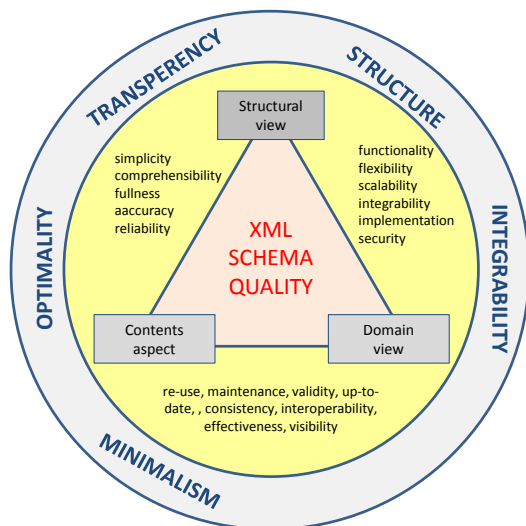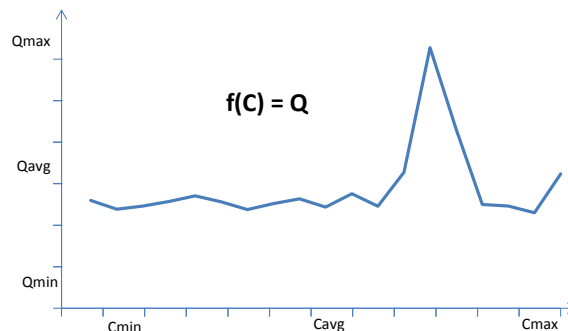

Fig. 2 Quality aspects in XML schemas

Fig. 3 Quality-complexity dependance

2. METRIC TYPES

So that individual metrics could be compared, NORMALIZATION of parameters was conducted. All the parameters that were used within the metrics and their results were transformed to a scale of 0 to 1, where 0 represented the worst value for each parameter and 1 the best value. The transformations based on linear programming, assuming that the growth relationship is linear. The following metrics address all aspects of XML schema quality.

2.1 Structural aspect

Other authors have researched measuring the structure of XML schemes for calculating the complexity and quality by McDowell and others [Burris 2012]. The authors present a number of metrics, taken

mainly from "quality model" ISO standard and link them into a single formula. Each variable is further multiplied, however the factors are not justified, values are not normalized, so the formula cannot be applied, but we have analysed and partly used in our calculation formula of quality.

Within the complexity calculations we can conclude that the higher the value of the individual, the greater the complexity (the relationship is shown in Fig. 3). According to XML schema needs we redefined metrics into the following composite metric (1) with the following parameters:

- *S1 - relationship between simple and complex data types*
- *S2 - relationship between annotations and the number of elements*
- *S3 - average number of restrictions on the declaration of a simple type*
- *S4 - percentage of the derived type declarations of total number of declarations complex types*
- *S5 - diversification of the elements or 'fanning' which is influenced by the complexity of XML schemas suggesting inconsistencies in XML schemas that unnecessarily increase the complexity*

$$Q1 = \frac{S1 + S2 + S3 + S4 + S5}{5} \tag{1}$$

2.2   Transparency and documentation of the XML Schema

The importance of well documented and easy-to-read/understand XML schema is addressed in the following relationship: number of annotation (N_{An}) depending on the number of items (N_E) and attributes (N_{At}) illustrates the documentation of XML schemas, supposing that more information about the building blocks increases the quality. The parameters in metric 2 regard transparency and documentation.

$$Q2 = \frac{N_{An}}{N_E + N_{At}} \tag{2}$$

2.3   XML schema optimality

In metric 3 we combined several parameters, indicating the optimal structure of an XML Schema. The metric evaluates whether the in-lining pattern has been used, the least preferable one in XML schema building. In doing so, we focus on the following relationships:

- *(O1) The relationship between local and all elements*
- *(O2) The relationship between local attributes and all attributes*
- *(O3) The relationship between global and complex elements of all the complex elements*
- *(O4) The relationship between global and all the simple elements of simple elements.*

Ratio between XML schema building blocks (O1, O2, and O4) should be minimized; meaning minimisation of local elements and attributes and more global simple and complex types; the number of global elements (O3) should be as low as possible, due to the problem of several roots (such flexibility is not always appreciated). This particular parameter differentiates domains into two groups (the flexible ones appropriate to validate multiple different XML schemas, and the strict ones, striving to one root policy for validity or other reasons). In metric 3 we assumed the majority of XML schemas want a certain level of flexibility, therefore the aspect of security was disregarded.

$$Q3 = \frac{O1 + O2 + (1 - O3) + O4}{4} \tag{3}$$

The metrics, described in the following subchapters, use a similar set of parameters:

- *(N_E) Number of elements*
- *(N_{At}) Number of attributes*
- *(N_{An}) Number of annotations*
- *(LOC) Number of lines of code*
- *(N_{re_all}) - number of references to elements (simple and complex)*

- (N_{ra_all}) - *number of references to attributes*
- (N_{rg_all}) - *number of references to groups (elements and attributes)*
- (N_{ri_all}) - *the number of schemes and imported*
- (Ng) - *The number of groups*

2.4   XML schema minimalism

In this metric we combine the parameters that indicate the minimum XML schemas building blocks, where the concept of minimalism is defined as the level, where one can anticipate that there is no other set of less building blocks, however still descriptive full:

$$Q4 = \frac{N_{An} + N_E + N_{At}}{LOC} \tag{4}$$

2.5   XML schema re use

The equation was inspired by author [Washizaki, Fukazawab 2005], where we summed up and defined a set of metrics for measuring the re-use of the software. The metric includes parameters that allow the re-use and are inherently global. We included the following parameters:

$$Q5 = \frac{N_{re_{all}} + N_{ra\_all} + N_{rg\_all} + N_{ri\_all}}{N_E + N_{At} + N_g} \tag{5}$$

2.6   XML schema integrability

Definition of equation was taken from the idea of density of software components [Narasimhan 2007], where the authors calculate the density of the other segments of the software and the density of interactions between them (lines of code, operations, classes, modules ...).We adjusted and simplified the formula into the following equation:

$$Q6 = \frac{N_E + N_{At} + N_g + N_{re_{all}} + N_{ra\_all} + N_{rg\_all} + N_{ri\_all} + N_{re_{all}} + N_{An}}{N_E + N_{At}} \tag{6}$$

3. METRICS APPLICATION

We tested proposed metrics on a set of 200 XML schemas, subtracted from different domains, acknowledging several standards, available on the market in a certain domain. Each XML schema was evaluated manually and automatically with proposed metrics, eliminating possible duplicates due to crossing of different fields. The results of all metrics were combined and nominated to a scale from 1-3, where a level 1 schema is of high quality and level 3 XML schema is of low quality (using identical scale in case of the manual evaluation). Comparing the two types of evaluation, 83% of data received an equal evaluation (Fig. 4).
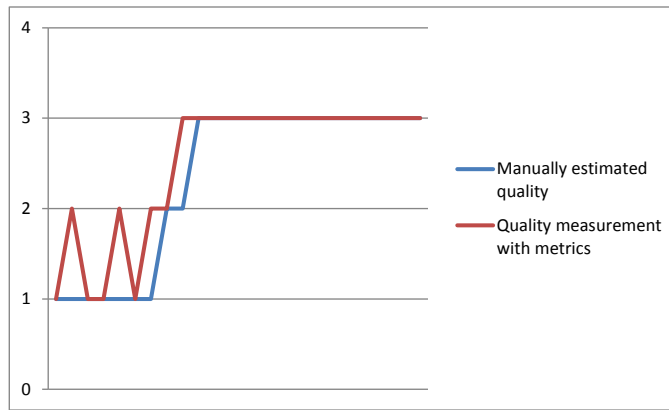
Fig. 4 Manual and metrical measurement of XML schema quality.

All metrics were considered as equal, therefore no priority weights are applied to each metric. This limitation was used due to simplification of our early stage metric framework; weights were omitted for the length purposes, since the paper does not include domain/aspect priorities clarification. We treated all aspects of XML schema as equal due to heterogeneous domain, which were not explored in this paper. Definition of weights will be a part of our future work. For the purposes of this paper, we used the following equation:

$$Q = Q_1 + Q_2 + Q_3 + Q_4 + Q_5 + Q_6 \qquad (7)$$

A presentation of metrics application is shown in figure (Fig. 5).A sum of 220 real-life standard or semi-standard XML schemas was used to apply defined metrics. Evaluation software produced a resulting XML document with a summary of all data, some warnings or eventual errors and metric results.
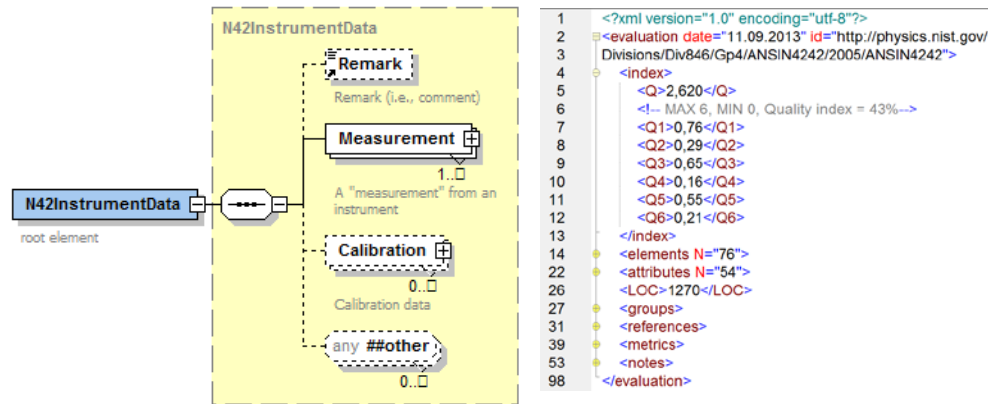

Fig. 5 Metric application example based on an XML schema.

4. DISCUSSION

The focus of the paper was definition of a full set of parameters for assessing the quality of XML schemes, trying to include all aspects and needs of XML schema quality. We defined six metrics, focusing on important aspects of XML schema quality, and repositioned XML schema facts into parameters, measuring the importance of each building block. To assure correctness, we evaluated each XML schema manually based on a simple overview, noting clearness and readability; and compared our results with metrics' results. The overlapping was at 83%.

Correct (and quick) measurement of XML Schema quality provides a strategic decision-making and improvement in data organization, as a standard mechanism (internal or global) for evaluation of XML

schemes quality. Software metrics are a good basis for XML schema quality measuring, however some accommodations are necessary according to their needs and demands. As users operate with different data from multiple domains of XML technologies application, the quality measurements vary depending on the flexibility (or inflexibility) of structures.

In future work we will further explore applicability of defined metrics, their success and validity on practical examples and the need for metrics adaptability according to the domain in which an XML schema is used.

REFERENCES

Zhang, Y. (2008). *Literature Review and Survey: XML Schema Metrics*.
Wes Rishel. (2011). *Does XML Schema Earn its Keep?* The Gartner Blog Network. http://blogs.gartner.com/wes_rishel/2011/12/31/ok-xml-schema-does-earn-its-keep-in-hl7/
Sušnik, M. (2008). *V slogi je e-račun! Monitr Pro*, http://www.monitorpro.si/41040/praksa/v-slogi-je-e-racun/.
Standard ISO/IEC 9126 Software engineering
McDowell, A., Schmidt, C., Yue, K. (2004). *Analysis and Metrics of XML Schema*. Proceedings of the International Conference on Software Engineering Research and Practice, SERP'04, v 2, p 538-544, 2004.
Burris, E. (2012), Hierarchical Nature of Software Quality, Programming in the Large, The Practice of Software Engineering, http://programminglarge.com/hierarchical-nature-of-software-quality/.
Narasimhan, V.L., Hendradjaya, B. (2007). *Some theoretical considerations for a suite of metrics for the integration of software components*. Information Sciences, Volume 177, Issue 3, 1 February 2007, Pages 844-864. http://dx.doi.org/10.1016/j.ins.2006.07.010
Washizaki, H., Fukazawab, Y. (2005). *A technique for automatic component extraction from object-oriented programs by refactoring*. Volume 56, Issues 1–2, April 2005, Pages 99–116. http://dx.doi.org/10.1016/j.scico.2004.11.007