XML Schema Quality Index in the Multimedia Content Publishing Domain

MAJA PUŠNIK, MARJAN HERIČKO AND BOŠTJAN ŠUMAK, University of Maribor
GORDANA RAKIĆ, University of Novi Sad

The structure and content of XML schemas impacts significantly the quality of data respectively documents, defined by XML schemas. Attempts to evaluate the quality of XML schemas have been made, dividing them into six quality aspects: structure, transparency and documentation, optimality, minimalism, reuse and integrability. XML schema quality index was used to combine all the quality aspects and provide a general evaluation of XML schema quality in a specific domain, comparable with the quality of XML schemas from othe r domains. A quality estimation of an XML schema based on the quality index leads to a higher efficiency of its usage, simplification, more efficient maintenance and higher quality of data and processes. This paper addresses challenges in measuring the level of XML schema quality within the publishing domain, which deals with challenges of multimedia content presentation and transformation. Results of several XML schema evaluations from the publishing domain are presented, compared to general XML schema quality results of an experiment, that included 200 schemas from 20 different domains. The conducted experiment is explained and the state of data quality in the publishing domain is presented, providing guidelines for necessary improvements in a domain, dealing with multimedia content.

Categories and Subject Descriptors: **H.0. [Information Systems]:** *General*; **D.2.8 [Software Engineering]:** ***Metrics*** *— Complexity measures; Product metrics*; **D.2.9. [Software Engineering]:** *Management — Software quality assurance (SQA)*

General Terms: *Software quality assurance*

Additional Key Words and Phrases: *software metrics, quality metrics, XML Schema, multimedia, publishing*

INTRODUCTION

This paper is focused on the publishing domain, documents in the publishing domain and the quality level of documents' structure, defined by XML schemas. XML schemas are a widely used technology for structure definition of XML documents and can be on very different quality levels, measured by predefined XML schema metrics, in our case with a quality index, explained in this paper. The activities in this paper include (1) collecting available XML schemas from the publishing field, (2) measuring the characteristics of XML schemas based on quality metrics (the quality index) and (3) critically evaluating their quality and setbacks as well as (4) comparing the results with XML schemas from other (previously evaluated) domains.

The publishing process is performed in both printed and electronic form, however there is an increasing number of eBooks (Shaffer, 2012). In recent times eBooks have started to become more interactive to the extent of providing rich multimedia content; hyperlinks to resources on the web; allowing the reader to highlight text, add notes and bookmark pages, videos, interactive games and other. Newer interactive features include different multimedia content such as embedded audio, video, slide shows and image galleries (Fenwick Jr, J.B., Phillips, R. , Kurtz, B.L., Weidner, A. , Meznar, 2013) and there are many publishing aspects that need to be addressed based on new

Authors's addresses: Maja Pušnik, Marjan Heričko, Boštjan Šumak, Faculty of Electrical Engineering and Computer Science, University of Maribor, Smetanova 17, 2000 Maribor, Slovenia, email: maja.pusnik@uni-mb.si, marjan.hericko@um.si, bostjan.sumak@um.si
Gordana Rakić, University of Novi Sad, Faculty of Sciences, Department of Mathematics and Informatics, Trg Dositeja Obradovica 4, 21000 Novi Sad, Serbia; email:gordana.rakic@dmi.uns.ac.rs

In: Z. Budimac, Z. Horváth, T. Kozsik (eds.): Proceedings of the SQAMIA 2016: 5th Workshop of Software Quality, Analysis, Monitoring, Improvement, and Applications, Budapest, Hungary, 29.-31.08.2016. Also published online by CEUR Workshop Proceedings (CEUR-WS.org, ISSN 1613-0073)

demands. In this paper, we are evaluating the quality of XML schema support, included in different multimedia content types for the publishing purposes, and comparing the results with other domains.

The paper is organized in the following manner: related work and research questions are included in the first introduction section. The quality index with quality aspects is presented in the second section. Section three includes results of applying quality index on XML schemas from the publishing field and a comparison with other domains is discussed. Limitations and threats to validity are explained in section 4 and conclusion and future work is presented in section 5, followed by listed references.

1.1 Related work

Publishing domain, its multimedia content and XML technologies have little history compared to other domains, however several papers were found, addressing its connection and influence on quality issues. Only for the last 5 years, more than 100 search results were identified. 16 papers were extracted, focusing on topics regarding multimedia content including XML documents and the publishing of different document types. Based on existing research, several publishing fields with extensive multimedia content were identified: geographical documentation, medical documentation and general software documentation.

XML technologies are greatly involved in defining and presenting geographical documentation (Laurini, 2014) and land administration (Lemmen, van Oosterom, & Bennett, 2015) as well as other industries; (Chen, 2016) discusses XML schema benefits in the process of integration, filtering and formatting of graphical information across the globe (Patel, Nordin, & Al-Haiqi, 2014). The presentation of data with XML support is also documented in medical literature (Wagner et al., 2016) and is helpful at evaluation of student's literature understanding (Aparicio, De Buenaga, Rubio, & Hernando, 2012) as well as in correlating fields such as ecology (Chamanara & König-Ries, 2014). Context-awareness and behavior adaptation of different multimedia content based on XML technologies is addressed in (Chihani, Bertin, & Crespi, 2014) and knowledge domain is emphasized in (Ait-Ameur & Méry, 2015). Publishing of learning material is greatly supported by XML; where authors define visual perception improvement through XML annotations (Hiippala, 2012), mobile learning products have even greater need for XML technologies (Sarrab, Elbasir, & Alnaeli, 2016), and proper (educational) literature for dangerous life situations such as earthquakes (Gaeta et al., 2014) also relies on XML based structures. Educational aspect of appropriately designed context can provide support in the system (Chik-Parnas, Dragomiroiu, & Parnas, 2010) and big data within large amounts of documents is addressed as well in (Priyadarshini, Tamilselvan, Khuthbudin, Saravanan, & Satish, 2015). The research of listed papers exposes several aspects of benefits and potential problems in the publishing field in general and creates a basis for the following research questions, which will be addressed in this paper:

(1) Does the publishing domain use XML documents and what standard XML schemas are being used?
(2) What is the quality level of XML schemas in the publishing domain?
(3) How are they compared to XML schemas in other domains such as computer science and other?
(4) How can the level of quality be improved?

Literature review and an experiment based on existing XML schema metrics were methods, used to answer listed research questions in addition to critical comparison to existing results of other domains. A set of metrics for assessing the quality of XML schemas is presented as well, united in a general quality evaluation - XML Schema Quality Index, used to bring all results to a common ground, making them comparable. The quality index is presented in the next section.

QUALITY INDEX

The quality of XML schemas is a general term and includes several aspects (structure, transparency, optimality, minimalism, reuse and integrability), addressed in (Pušnik, Boštjan, Hericko, & Budimac, 2013). The aspects were defined based on a preliminary research of general software metrics, presenting most used latent variables which cannot be always measured objectively. Therefore, the 6 aspects include measurable parameters of XML schema building blocks and their relations, encapsulating the final quality index: (1) structural quality aspect, (2) transparency and documentation of the XML schema quality aspect, (3) XML schema optimality quality aspect, (4) XML schema minimalism quality aspect, (5) XML schema reuse quality aspect and (6) XML schema integrability quality aspect. The exact calculation are presented in (Pusnik, Hericko, Budimac, & Sumak, 2014).

The goal of this paper is to measure the quality of XML schemas, which are an important part of several domains, where data is exchanged in form of XML documents. The quality of XML schemas indirectly impacts the information system quality and further on different companies' business processes (publishing companies as only one example). Companies however often use XML schemas, who meet the minimum criteria of syntactical correctness and content description (Pusnik, Hericko, Budimac, & Sumak, 2014). Quality evaluation was conducted on 200 schemas where results indicated that 30% of identified XML schemas (within 20 different domains) have a very low quality index and are built inappropriately (regarding the structure and documentation), influencing the quality of the information solution. The purpose of this paper was to evaluate and compare the publishing field to the general situation, since the publishing field does include structured data, being transferred on a daily basis. The domains, to which we compare the publishing domain, are presented in the next section.

1.2 XML schemas in the publishing domain

Analysis of 200 schemas from 20 domains was conducted in (Pušnik, 2014). The criteria which domain was included into the analysis was the number of used XML schemas: only top 20 domains with most XML document transactions and XML schema definitions were included. Domains, which provided a general state of used XML schemas, are presented in table (Table *I*)

Table I. Set of domains (Pušnik, 2014)

D1 - Mathematics and Physics	D11 –Decision Science
D2 - Materials Science	D12 –Medicine
D3 - Telecommunications	D13 - Economics and finance
D4 - Manufacturing	D14 - Law
D5 - Energy and Electronics	D15 - Social science
D6 - Engineering	D16 –Health and sport
D7 - IT architecture and design	D17 –Construction
D8 - Traffic	D18 - Librarianship (Library)
D9 - Communications	D19 - Landscape and geography
D10 –Computer Science	D20 –Media, journalism, newspapers

The publishing domain was not specifically investigated within the primary set of domains, however was included in domain D20 - Media, journalism, newspapers. However, due to the expanding use of XML and related technologies in the publishing field, we conducted a similar research of XML schema quality in this specific field and compared it to the average values, received when analysing most often used XML schema. The 10 identified XML schemas, that were valid, publically available and supported by all included and imported XML schemas were evaluated based on the six quality aspects, presented in the next section.

1.3 The six quality aspects

The aspects of XML schemas are evaluated and presented in (Pusnik et al., 2014) and include equations, combining the listed parameters (Table *II*). They are presented in more detail in the following sections.

1.3.1 *Structural quality aspect (QA1).* The structure aspect evaluates the number and relationship among building blocks of XML schemas. It includes several measured parameters and focuses on the level of complexity. Metrics include relationship*s* between simple and complex data types, relationship between annotations and the number of elements, average number of restrictions on the declaration of a simple type, percentage of the derived type declarations of total number of declaration complex types and diversification of the elements or 'fanning', which is influenced by the complexity of XML schemas, suggesting inconsistencies that unnecessarily increase the complexity.

1.3.2 *Transparency and documentation of the XML Schema (QA2).* The importance of well documented and easy-to-read as well as understandable XML schema is derived from the following relationship: number of annotation per number of elements and attributes, illustrating the documentation of XML schemas, supposing that more information about the building blocks increases the quality.

1.3.3 *XML schema optimality quality aspect (QA3).* 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: the relationship between local and all elements, the relationship between local attributes and all attributes, the relationship between global and complex elements of all complex elements, the relationship between global and simple elements of all simple elements. Ratio between XML schema buil*ding blocks should be minimized,* indicating minimization of local elements and attributes and maximization of global simple and complex types. The number of global elements however should be as low as possible, due to the problem of several roots (such flexibility is not always appreciated).

1.3.4 XML schema minimalism quality aspect (QA4). Metric of minimalism is defined as the level, when there is no other full set of less building blocks. Number of annotations, elements and attributes should be according to the size of XML schema (LOC respectively).

1.3.5 *XML schema reuse quality aspect (QA5).* Metric is focused on reuse of the existing software and includes parameters that allow the reuse and are inherently global. References are mostly calculated and number of references to elements (per defined elements) is measured as well as the number of references to attribute (per defined attributes), number of references to groups (per defined groups) and the number of imported or included XML schemas.

1.3.6 *XML schema integrability quality aspect (QA6).* Metrics measure capability of XML schema components to be integrated, including number of elements and references on elements (per defined elements), number of attributes and references on attributes (per defined attributes), number of groups and references on groups as well as number of imported XML schemas and annotations.

1.3.7 XML schema Quality Index (QI). Equally combines all six metrics and provides the average value. The values have been scaled between the 0 and 1 (Pušnik, 2014) for the results to be comparable.

The measurement and evaluation process based on XML schema parameters and metrics is described in (Pušnik et al., 2013) and summarized in Table *II*, addressing basic characteristics of XML schemas. Their values are gathered into the six metrics, composing the holistic quality index in the following section.

Table II Names and abbreviations of all used parameters and metrics

Simple metrics (parameters)	Composite metrics
File size [KB] (P1) *Number of imports (P2)* *Element related parameters: number of all elements (P3), number of* global *elements (P3.1), number of local elements (P3.2), number of simple elements (P3.3), number of complex elements(P3.4), number of global complex elements (P3.1.1), number of global simple elements (P3.1.2).* *Attribute related parameters: number of all attributes (P4), number of local attributes (P4.1), number of global attributes (P4.2)* *Lines of code (P5)* *Group related parameters: number of element groups (P6.1), number of attribute groups (P6.2)* *Reference related parameters: number of element references (P7.1), number of references on simple elements (P7.1.1), number of references on complex elements (P7.1.2), number of references on attributes (P7.2), number of references on element groups (P7.3), number of references on attribute groups (P7.4)* *Number of annotations (P8)* *Number of restrictions (P9)* *Number of derived (extended) types (P10)*	*XML schema type (M1)* *Ratio of simple to complex types of elements(M2)* *Percentage of annotations over total number of elements and attributes (M3)* *Average number of restrictions per number of elements and attributes (M4)* *Number of all data types (M5)* *Percentage of derived data types over all complex types (M6)* *Average use of minimal and maximal occurs per defined elements (M7)* *Average number of attributes per complex types (M8)* *Number of unbounded elements (M9)* *Element fanning (M10)* *Quality index (QI)*

RESULTS

XML schemas, defined within different companies and organizations for the needs of publishing process were analyzed. Based on the analysis of 200 XML schemas from 20 domains, the 10 XML schemas from the publishing domain were compared. The quality aspects of the publishing domain to average results is presented in Fig. 1.

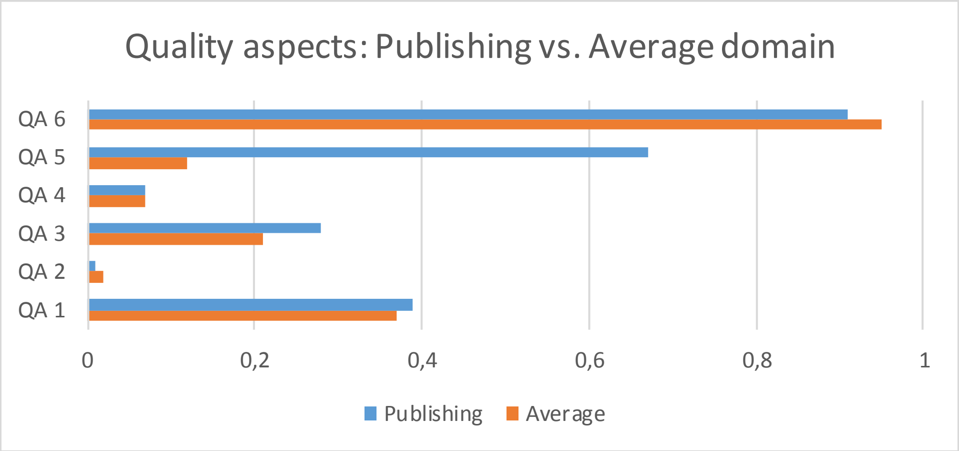

Fig. 1. XML schema quality aspects in the publishing domain compared to average domains

Based on the calculated index quality, a composite of all six quality aspects (equally distributed), the research questions were addressed:

(1) RQ1 –Several XML schemas were found, connected to the publishing field (respectively publishing process) through active research. No standard forms were found, 10 were extracted.
(2) RQ2 – The quality level was measured through metrics, defined in (Pusnik et al., 2014). The average quality of XML schema in the publishing field is 39%.
(3) RQ3 - The quality index of 39% is significantly higher than by the average quality index (of all 20 domains, where XML schemas are most common) which is 29% based on an experiment from 2014.
(4) RQ4 – Comparing to average XML schemas, the publishing field had lower results only at transparency and documentation quality aspect, all other quality aspects were above average.

The results provided better quality index for the publishing domain for most publishing XML schemas compared to average ones (Fig. 1) from the set of 200 XML schemas. The T-test for finding significant difference among the groups was used. The p-value resulted in 0,21 which makes the difference non-significant among parameters. The results comparison is presented in Table III. The metric M1 is left empty, since average value cannot be determined.

Table III. Comparing publishing domain with average results of all domains

PARAMETERS		
	Publishing domain - average	All domains - average
P1	604,000	133,932
P2	1,400	0,795
P3	86,600	77,727
P3.1	48,000	26,755
P3.2	38,600	50,973
P3.3	31,800	27,691
P3.4	54,800	50,036
P3.1.1	35,700	19,250
P3.1.2	12,700	7,514
P4	26,600	47,655
P4.1	26,600	47,091
P4.2	0,000	0,564
P5	14969,200	3188,618
P6.1	1,000	4,364
P6.2	1,300	1,377
P7.1	65,400	69,118
P7.1.1	9,000	3,177
P7.1.2	56,400	65,941
P7.2	0,000	1,927
P7.3	0,200	7,114
P7.4	3,700	11,664
P8	0,900	0,977
P9	30,100	75,118
P10	41,800	35,309

METRICS		
	Publishing domain - average	All domains - average
M1		
M2	5,155	2,266
M3	0,340	0,513
M4	0,923	1,323
M5	31,900	34,755
M6	2,116	2,897
M7	38,458	0,768
M8	0,655	1,048
M9	73,000	59,473
M10	0,358	0,285

QUALITY APSPECTS		
	Publishing domain - average	All domains - average
QA 1	0,387	0,374
QA 2	0,005	0,022
QA 3	0,280	0,214
QA 4	0,071	0,072
QA 5	0,667	0,117
QA 6	0,908	0,950

LIMITATIONS AND THREATS TO VALIDITY

This research has limitations that have to be identified and discussed. Only one database was used (science direct) and only 10 XML schemas were evaluated due to comparability with other domains (for each domain 10 XML schemas were used). There is also a possibility of a human error and shortcomings of the used measurement tool are possible (Pušnik et al., 2013). The limitations are also connected to XML schema patterns respectively (un)reachability of all included or imported XML schemas. To confirm or disregard validity of results, the research should be repeated, based on other empirical research methods.

This paper does not include DocBook domain, although adopted by many publishing companies and common in the publishing field, as an extended use of DocBook is proposed in (Şah & Wade, 2012). The OASIS DocBook Schema for Publishers is an official variant of DocBook v5.x, specifically designed to support the publishing industry (Walsh., 2011) and is subjected to investigation of how well it supports interactive and multimedia content (Project, C-level, Spring, Supervisor, & Examiner, 2015). Authors pointed out the challenge of no standard nor best practices for writing documents on any subject and the publishers take different approaches with very different solutions. Therefore, in this research an investigation of the existing publishing field was launched, trying to include random and average XML schemas in publishing and DocBook was not included.

CONCLUSION AND FUTURE WORK

The paper addresses publishing domain issues through XML schemas, its characteristics, influence and contribution to organization, focused on assessing the quality of used XML schemas in the publishing field. The literature review reviled the importance of XML schemas within the publishing field and an experiment provided data to compare quality with other XML schemas and identify shortcomings that need to be addressed.

We have reused existing metrics for quality evaluation by comparing the 10 publishing XML schemas to 200 XML schemas of different origin. 6 aspects of quality were defined, combined into one quality metric, the quality index. The quality aspects include (1) structure, (2) transparency, (3) optimality, (4) minimalism, (5) reuse and (6) integrability. We have discovered that XML schemas from the publishing field are above average, providing an answer to the research question: the publishing domain does use XML schemas, the quality of them is above average however they still need to be improved mostly in the quality aspect of transparency and documentation. More detailed impact of XML schema quality is yet to be empirically confirmed and was included as an assumption in the paper.

Future work will extend the domains, where XML schemas will be evaluated. 20 domains have already been investigated, the publishing domain being the 21st. Additional domains will be further explored and compared as well additional XML schemas will be included in the experiment set for specific domains. Versions and the quality movement when changing an existing XML schema will also be explored.

ACKNOWLEDGMENTS

This joint work is enabled by bilateral project "Multidimensional quality control for e-business applications" between Serbia and Slovenia (2016-2017). Furthermore, the last author was partially supported by the Ministry of Education, Science, and Technological development, Republic of Serbia, through project no. OI 174023.

REFERENCES

Ait-Ameur, Y., & Méry, D. (2015). Making explicit domain knowledge in formal system development. *Science of Computer Programming*, *121*, 100–127. http://doi.org/10.1016/j.scico.2015.12.004

Aparicio, F., De Buenaga, M., Rubio, M., & Hernando, A. (2012). An intelligent information access system assisting a case based learning methodology evaluated in higher education with medical students. *Computers & Education*, *58*(4), 1282–1295. http://doi.org/10.1016/j.compedu.2011.12.021

Chamanara, J., & König-Ries, B. (2014). A conceptual model for data management in the field of ecology. *Ecological Informatics*, *24*, 261–272. http://doi.org/10.1016/j.ecoinf.2013.12.003

Chen, Y. (2016). Industrial Information Integration-A Literature Review 2006-2015. *Journal of Industrial Information Integration*. http://doi.org/10.1016/j.jii.2016.04.004

Chihani, B., Bertin, E., & Crespi, N. (2014). Programmable context awareness framework. *Journal of Systems and Software*, *92*(1), 59–70. http://doi.org/10.1016/j.jss.2013.07.046

Chik-Parnas, L., Dragomiroiu, M., & Parnas, D. L. (2010). A family of computer systems for delivering individualized advice. *Knowledge-Based Systems*, *23*(7), 645–666. http://doi.org/10.1016/j.knosys.2010.02.007

Fenwick Jr, J.B., Phillips, R. , Kurtz, B.L., Weidner, A. , Meznar, P. (2013). Developing a Highly Interactive eBook for CS Instruction. *SIGCSE '13: Proceeding of the 44th ACM Technical Symposium on Computer Science Education, Denver, CO, USA*, 135–140.

Gaeta, M., Loia, V., Mangione, G. R., Orciuoli, F., Ritrovato, P., & Salerno, S. (2014). A methodology and an authoring tool for creating Complex Learning Objects to support interactive storytelling. *Computers in Human Behavior*, *31*(1), 620–637. http://doi.org/10.1016/j.chb.2013.07.011

Hiippala, T. (2012). Reading paths and visual perception in multimodal research, psychology and brain sciences. *Journal of Pragmatics*, *44*(3), 315–327. http://doi.org/10.1016/j.pragma.2011.12.008

Laurini, R. (2014). A conceptual framework for geographic knowledge engineering. *Journal of Visual Languages and Computing*, *25*(1), 2–19. http://doi.org/10.1016/j.jvlc.2013.10.004

Lemmen, C., van Oosterom, P., & Bennett, R. (2015). The Land Administration Domain Model. *Land Use Policy*, *49*, 535–545. http://doi.org/10.1016/j.landusepol.2015.01.014

Patel, A., Nordin, R., & Al-Haiqi, A. (2014). Beyond ubiquitous computing: The Malaysian HoneyBee project for Innovative Digital Economy. *Computer Standards and Interfaces*, *36*(5), 844–854. http://doi.org/10.1016/j.csi.2014.01.003

Priyadarshini, R., Tamilselvan, L., Khuthbudin, T., Saravanan, S., & Satish, S. (2015). Semantic Retrieval of Relevant Sources for Large Scale Virtual Documents. *Procedia Computer Science*, *54*, 371–379. http://doi.org/10.1016/j.procs.2015.06.043

Project, B. D., C-level, C. S., Spring, E., Supervisor, R., & Examiner, M. B. (2015). Data models for interactive web based Textbooks.

Pusnik, M., Hericko, M., Budimac, Z., & Sumak, B. (2014). XML schema metrics for quality evaluation. *Computer Science and Information Systems*, *11*(4), 1271–1289. http://doi.org/10.2298/CSIS140815077P

Pušnik, M. (2014). Quality evaluation of domain specific XML schemas. *Doctoral Thesis*, 180.

Pušnik, M., Boštjan, Š., Hericko, M., & Budimac, Z. (2013). Redefining software quality metrics to XML schema needs. *CEUR Workshop Proceedings*, *1053*, 87–93.

Şah, M., & Wade, V. (2012). Automatic metadata mining from multilingual enterprise content. *Journal of Web Semantics*, *11*, 41–62. http://doi.org/10.1016/j.websem.2011.11.001

Sarrab, M., Elbasir, M., & Alnaeli, S. (2016). Towards a quality model of technical aspects for mobile learning services: An empirical investigation. *Computers in Human Behavior*, *55*, 100–112. http://doi.org/10.1016/j.chb.2015.09.003

Shaffer, C. A. (2012). Active eTextbooks for CS: what should they be? *SIGCSE '12 Proceedings of the 43rd ACM Technical Symposium on Computer Science Education*, 680.

Wagner, M., Vicinus, B., Muthra, S. T., Richards, T. A., Linder, R., Frick, V. O., … Weichert, F. (2016). Text mining, A race against time? an attempt to quantify possible Variations in text corpora of medical publications throughout the years. *Computers in Biology and Medicine*. http://doi.org/10.1016/j.compbiomed.2016.03.016

Walsh., N. (2011). Getting Started with DocBook Publishers. Retrieved from http://www.docbook.org/tdg5/publishers/5.1b3/en/html/ch01.html