How Much is XML Involved in DB Publishing?

Gyula I. Szabó

XML has been intensive investigated lately, with the sentence, that ”XML is (has been) the standard form for data publishing”, especially in data base area [1].

That is, there are assumptions, that the newly published data take mostly the form of XML documents, especially when databases are involved. That is the reason of the heavy investments applied for researching the handling, quering and comprising XML documents [2],[3].

This study would like to check these assumptions, while investigating the documents accessible over the Internet, possible to go under the surface, into the ”deep WEB”. The investigation focuses on the large scientific databases, but the commercial data stored in the ”deep WEB” will be handled also.

The technique of randomly generated IP addresses will be used to reach publicly accessible sites for testing the whole WEB as suggested by Kevin ChenChuan Chang et al. [4].

The random IP addresses (when accessible) can be used for analyzing the files of the addressed site, we would like to check the amount of xml documents among the entity of files present on the given site.

Another aim of the study is finding a simple attribute to be used for declaring an XML-document being a database (one can assume, if the size of the document is “large enough” it can be accepted as database). This hypothese will be also investigated and an acceptable value of size criterion proposed. By counting the documents that are used for storing databases using the proposed minimal size as criterion, the proportion of databases among the XML documents stored on the WEB can also be estimated. We would like to check a few number of known sites of large scientific databases (first of all biochemical and astronomical databases)

We would also like to investigate the rate of ”masked XML” files, the documents, that declares themselves as XML documents, but in fact they are built up as HTML files.

We don’t want to create a new search engine, and the aims of these investigations cannot be fulfilled by using the known search engines, because they try to find a given text in the documents present on the WEB, while we would like to get statistical data over documents with a given structure (or semi-structure).

These investigations can be repeated in the future in order to get a dynamic picture of the growing rate of the number of the XML documents present on the WEB, and also over the growing rate of the size of the databases stored as XML documents..

Acknowledgements

This work was supported by ELTE IK TAMOP

References

[1] Wenfei Fan and Leonid Libkin: On XML Integrity Constraints in the Presence of DTDs, in Journal of the ACM (JACM), Volume 49 , Issue 3, pp 368 - 406, May 2002.

[2] Ray, I. Muller, M.:Using Schemas to Simplify Access Control for XML Documents, in Lecture Notes in Computer Science, 2004, ISSU 3347, pages 363-368

[3] G. Leighton and D. Barbosa. Optimizing XML Compression. In Proceedings of the Sixth International XML Database Symposium (XSym 2009), LNCS 5679, pp. 91-105, 2009.

[4] K. C.-C. Chang, B. He, C. Li, and Z. Zhang. Structured databases on the web: Observations and implications. Report UIUCDCS-R-2003-2321, Dept. of Computer Science, UIUC, Feb. 2003.