An XML Schema to Support Reliable Web-crawling and the Exchange of Web Crawl Graphs

Viv Cothey

viv.cothey@wlv.ac.uk
University of Wolverhampton, School of Computing and IT, Wolverhampton, (UK).

The Requirement

To be reliable the Web crawl graph produced by a Web-crawler needs to be accompanied by a description of the crawl policies that affected the crawler (Cothey, 2004).

Because of the resource consumption involved in Web-crawling including the consumption of Internet resources (network and server), Web crawl graphs are scarce. Therefore their exchange and shared access for the purposes of research should be encouraged. In addition, the reporting of Web crawl graphs should be comprehensive in that as much as possible of the information collected by the Web-crawler should be included so as to maximise research opportunities.

Web crawl graphs are arbitrarily large. They therefore need to be serialised and processed as a sequence of unique nodes and associated arcs.

Existing general languages for describing graphs, (see <url:http://www.graphdrawing.org>) such as GraphML (Brandes et al., 2002) are not optimised to work with Web-crawlers or to report Web crawl graphs.

The Advantages of Using XML (Extensible Markup Language)

XML is an open standard promoted by the WWW Consortium, (see url:http://www.w3.org/TR/REC-xml/) and is supported by all major producers of computer systems. XML is the foundation of the "semantic Web" and is accompanied by a wide range of validation, transformation and processing tools.

XML documents are vendor, platform and hardware independent. They can therefore be exchanged between computer systems without proprietary obstacles or interference.

Any international character (that is not just "American-English") can be included in an XML document.

XML documents can be processed serially. Hence arbitrarily large graphs can be serialised and exchanged.

The content model of an XML document can be defined by a schema and hence the document can be validated.

The Document Type Definition (DTD) schema

A DTD is given here as figure 1. The DTD defines the content model of the blinker (We**b link** crawl**er**) XML crawl graph document that is generated progressively as the output from a Web-crawler. For example, each blinker XML document contains three principal sections all of which must be present and appear in the prescribed order,

header which reports the Web-crawler configuration, that is the crawl policies and the values of the constraining parameters that affect the crawl graph,

crawl which contains a serialisation of the crawl graph as a sequence of nodes and associated arcs, and,

trailer which presents summary and concluding information relating to the crawl graph.

More information is available

The blinker XML schema has been successfully used to report and exchange crawl graphs in respect of both the European and Canadian innovation systems (Cothey, 2005). For further details please see the handout and/or contact <viv.cothey@wlv.ac.uk>.

Acknowledgement

This work was supported by a grant from the Common Basis for Science, Technology and Innovation Indicators part of the Improving Human Research Potential specific programme of the Fifth Framework for Research and Technology Development of the European Commission. It is part of the Web indicators for scientific, technological and innovation research (WISER) project, (Contract HPV2-CT-2002-00015).

References

Brandes, U., Eiglsperger, M., Herman, I., Himsolt, M., & Marshall, M. S. (2002). GraphML progress report: structural layer proposal. *Proceedings 9th Intl. Symp. Graph Drawing*, LNCS 2265 (pp. 501-512). London: Springer.

Cothey, V. (2004). Web-crawling reliability. *Journal of the American Society for Information Science and Technology*, 55(14), 1228-1238.

Cothey, V. (2005). Some preliminary results from a link-crawl of the European Union Research Area Web. *Elsewhere these proceedings*.