2007

Novel Approaches to the Delivery of XML and Schemas

S. J. Davis
University of Wollongong, stdavis@uow.edu.au

Novel Approaches to the Delivery of XML and Schemas

A thesis submitted in fulfilment of the
requirements for the award of the degree

Doctor of Philosophy

from

UNIVERSITY OF WOLLONGONG

by

Stephen James Davis
Bachelor of Engineering (Honours Class I)
University of Wollongong, 2001

SCHOOL OF ELECTRICAL, COMPUTER
AND TELECOMMUNICATIONS ENGINEERING
2007

Abstract

Typically XML documents are delivered as whole documents, and the transmission does not consider if all of this data may actually be relevant to the user. This results in inefficiencies in terms of both bandwidth (transferring unnecessary data) and computing resources (extra memory and processing to handle the entire XML document). Through exploitation of XML's tree-like structure, a simple and lightweight protocol is introduced (referred to as RXPP). Designed with mobile devices in mind, RXPP provides users with the ability to navigate and retrieve data from remote documents on a node-by-node or branch-by-branch basis, allowing users to retrieve only fragments of interest. By skipping unwanted XML nodes, this avoids the need to always maintain a full copy of the XML document locally as processing of the document is performed remotely. When only partial views of XML documents are maintained, the processing requirements of mobile devices are less demanding and requires less memory. Furthermore, time and money can be saved when using mobile devices in bandwidth limited environments where data is often charged per kilobyte as only the relevant data is retrieved when the user selects the next node or branch.

Through extension of RXPP, a two-way exchange of XML documents is introduced called RXEP. RXEP allows users to receive XML fragments and also update remote XML documents. In addition to the navigation features of RXPP, RXEP further allows users to construct queries (e.g., using the XPath language), requesting many XML nodes from a remote XML document. In some cases, users can construct well crafted queries to retrieve all the relevant XML fragments using only a single request. RXEP locators are introduced which extend the path features of XPath to the provide precise location of received XML fragments within the clients own local version. RXEP locators provide extra information such as the nodes

absolute location and total number of sibling nodes. RXEP locators thus allow clients to retrieve fragments of XML whilst replicating the exact structure of the original XML document. Through exploitation of RXEP locators and RXEP's two-way exchange, office suites using XML as a document format (such as MS Office and Openoffice), becomes an ideal target for collaborative editing amongst many users. This allows users to download only relevant parts of a document and upload corrections or modifications without the need to upload the entire document.

To further increase the efficiency of RXEP, a binarised (i.e., compressed) version of the protocol is explored. By utilising well established tree-based binarisation techniques significant savings can be achieved through compression of the RXEP structure and requested XML data. A new technique called SDOM is introduced which merges the structural information from XML Schemas with the requested XML document. SDOM allows users to request XML fragments using RXEP techniques where the requested XML data can be compressed on-the-fly using the information contained within SDOM. BinRXEP thus allows users to perform queries or navigation on remote XML documents and receive the results in a compact and compressed form. In many cases, the overhead added by RXEP, is reduced to less than a byte when using binRXEP.

Techniques for the transmission of both XML and XML Schema fragments within a single RXEP packet are proposed. Utilising RXEP, a user can request fragments with a of XML data from a remote document with a further option to request the XML Schema fragment required for validation of that fragment. In this way, the user can avoid retrieving all XML Schemas associated with an XML document, and may only retrieve the relevant XML Schema fragments.

Finally, the collaborative creation of XML Schemas is introduced. Utilising RXEP XML and Schema techniques, users can all contribute to the creation of a schema in realtime, while seeing the progress of other users. This collaborative creation of schemas can lead to quicker creation of XML Schemas. Users may then extend the current set of descriptors or generate new descriptors using ideas from the previous schema updates, thus resulting in a richer set of descriptors.

Statement of Originality

This is to certify that the work described in this thesis is entirely my own, except where due reference is made in the text.

No work in this thesis has been submitted for a degree to any other university or institution.

Signed

Stephen James Davis

18th of February, 2007

Acknowledgments

I would like to thank my supervisor Associate Professor Ian Burnett for his guidance, encouragement and support.

Thanks to my parents, Robert and Evelyn Davis, for their support and sacrifices over the years to get me here. Thanks to Eva Cheng, Chris Davis and Daniel Franklin for proof reading chapters of this thesis.

Finally, I would also like to thank my friends and colleagues of Whisper, SECTE and TITR for their support and enjoyable memories.

Contents

List of Figures

List of Tables

List of Abbreviations

API	Application Programming Interface
ASCII	American Standard Code for Information Interchange
ASN.1	Abstract Syntax Notation number One
AU	Access Unit
BinRXEP	Binary Remote XML Exchange Protocol
BiM	Binary format for MPEG-7
CODEC	enCOder / DECoder
DI	Digital Item
DID	Digital Item Declaration
DOM	Document Object Model
DTD	Document Type Definition
DIDL	Digital Item Declaration Language
FTP	File Transfer Protocol
FRU	Fragment Request Unit
FUU	Fragment Update Unit
GPRS	General Packet Radio Service
HTTP	HyperText Transport Protocol
IEC	International Electrotechnical Commission
IP	Internet Protocol
ISO	International Organization for Standardization
MIME	Multipurpose Internet Mail Extensions
MPEG	Moving Pictures Expert Group
OWL	Web Ontology Language

PSVI	Post Schema Validation Infoset
RAM	Random Access Memory
RXEP	Remote XML Exchange Protocol
RDF	Resource Description Framework
RXPP	Remote XML Pull Protocol
SAX	Simple API for XML
SDOM	Schema Document Object Model
SOAP	Simple Object Access Protocol
SQL	Structured Query Language
TeM	Textual Encoding format for MPEG-7
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
VLC	Variable Length Coding
UTF-8	UCS Transformation Format 8
W3C	World Wide Web Consortium
WAP	Wireless Application Protocol
WSDL	Web Services Description Language
XML	eXtensible Markup Language
XOP	XML-Binary Optimized Packaging