Position Paper for the W3C Workshop on Binary Interchange of XML Information Item Sets

Erik Wilde
Computer Engineering and Networks Laboratory
Swiss Federal Institute of Technology, Zürich

August 2003

1 Introduction

The W3C's initiative to start working on a binary representation of XML (or XML Infosets) is a very promising and important decision. Several areas of XML applications, where XML's verboseness has always been a concern, will certainly benefit from such an effort. Application areas of a binary XML representation range form network-level applications that by their very nature are concerned with bandwidth consumption, up to certain industries such as banking, where the sheer volume of transactions that need to be processed makes it necessary to use efficient and compact encodings.

The Swiss Federal Institute of Technology (ETHZ) as a member of the W3C wishes to participate in the process of discussing the requirements for such a binary encoding at the "W3C Workshop on Binary Interchange of XML Information Item Sets", and may also be interested in the following process of working on and defining a binary representation of XML.

2 Motivation

The "W3C Workshop on Binary Interchange of XML Information Item Sets" is a step in a direction that will probably have a severe impact on XML's future. Using the terminology of Pras and Schönwälder [3], the approach to define a *binary* representation for the *XML Infoset* mixes two issues, which are *information* and *data* models[1]. While XML exchange currently is based on the full information model of XML, a binary syntax based on the XML Infoset would change this, making it impossible for users of this encoding to use the full XML information model.

In a recently published technical report [5], we have proposed an information model for hyperlinks in XML. This information model simply is an Infoset-level reformulation of the syntax-oriented XLink recommendation. In two other publications [4, 6], the case for an extensible architecture of XML's information model has been described in more detail. Extensibility adds complexity, which

[1] An *information model* is an abstract description of the entities that are of concern in a given application area, as well as their relations. Specifically, an information model is not concerned with defining representations or interfaces or any other kind of access mechanism for the entities it is modelling. A *data model* can be regarded as the lower-level implementation of a higher-level information model.

is undeniable. However, at least some extensions of XML's information model, such as hypermedia, or XML Schema's *Post Schema Validation Infoset Contributions*, maybe beneficial to a sufficiently large number of people to justify this extensibility.

So far, the W3C position has been to define many things syntactically (i.e., on the data model layer), rather that their information model. There are a few examples where it has become clear that this position should at least be reconsidered:

- *XML Infoset:* The Infoset has been created as a result of the questions that arose in many contexts, what exactly the information of an XML document is, and what may be considered as syntactic sugar. Different groups had different opinions, and as a result, today there is a variety of XML information models, which is irritating for many XML users.
- *XLink:* XML hyperlinking has been defined in terms of XML syntax, rather than an information model. As a result, hyperlinking in XHTML 2.0 has been unable to re-use XLink directly, because of some syntax conflicts between XLink and XHTML 2.0 (the currently proposed solution, HLink [2], is a good motivation for avoiding syntax-centered designs).

In an effort to minimize the chances for similar problems in the future, a possible way to go would be to define information models and data models separately (XML Schema was the first W3C specification doing this). Doing so would admittedly add a layer of abstraction in the architecture of the Web, but it could very well be a very useful layer that enabled specification writers as well as users to separate more clearly the issues of information vs. data model. While the Infoset often has been regarded as being of interest to specification writers only, it should be considered much more important, because many XML users most of the time do not work with XML markup, but instead with XML data models such as DOM or XPath which are based on the Infoset (or something similar).

In recent specifications, the Infoset has become increasingly important, for example SOAP 1.1 had been based on XML syntax, while SOAP 1.2 is based on the Infoset. For some, this is disturbing, because it introduces an additional layer of abstraction. However, this additional layer makes it trivial to use SOAP 1.2 with a future binary encoding of Infosets, while this would be much harder with SOAP 1.1.

Therefore, we propose to reconsider the Infoset as it exists, and in particular clarify some of the open questions that arise when reading the Infoset specification (these questions are described in [6]). A more precisely defined Infoset could also be designed to support extensions, which in the spirit of XML could be identified by namespaces. This way, the Infoset could be extended by anyone wishing to do so, and in particular by the W3C to represent extensions of interest to the public (for example for linking, for PSVI contributions, or even for very simple things such as IDness, which currently also is being tackled on the syntax level [1]).

Defining a binary format for XML or XML Infosets or extensible Infosets should be a second step after carefully reconsidering the information vs. data model issue underlying the syntax problem. The "W3C Workshop on Binary Interchange of XML Information Item Sets" could be a great start to rethink XML's foundations (without changing the syntax or the Infoset!), and then to continue work on a retrofitted XML architecture. Our work on Infoset extensions (for XLink) and general Infoset extensibility explains our general interest in the area of binary XML based on the Infoset.

Apart from the issues touching the architecture of XML as it is today, we are using XML in various scenarios where compactness and efficiency are of concern, such as novel networking architectures. It would be beneficial to have more compact and efficient ways of representing XML. In particular, the soft- and hardware requirements (both in space as well as in time) to process this binary XML should be well-known before settling on any particular solution, so that it is possible to design systems based on these well-known time and space bounds.

3 Conclusions

While a binary representation certainly is a good idea and should be pursued, the approach of "Binary Interchange of XML Information Item Sets" should be considered as a starting point, rather than the only possible way. Restricting the binary representation to a subset of XML has the effect of not creating an alternative encoding for XML, but a profile of XML with a binary encoding. We think that either

- the binary encoding should be able to represent full XML, without any loss of information (such as subset of XML defined by the Infoset), or
- if the binary encoding is based on a different information model than XML (most likely, the Infoset), then this information model should be extensible to enable users to exchange their flavor of XML (which may be full XML, including all the information that the Infoset ignores or pre-processes).

We suggest to tackle the problem in a two-step process, first rethinking the foundations of XML and the Infoset, and in a second step using the outcome of the first step to define a binary representation with well-known algorithmic behavior.

References

[1] Jonathan Marsh. xml:id Requirements. World Wide Web Consortium, Working Draft WD-xml-id-req-20030806, August 2003.

[2] Steven Pemberton and Masayasu Ishikawa. HLink: Link recognition for the XHTML Family. World Wide Web Consortium, Working Draft WD-hlink-20020913, September 2002.

[3] Aiko Pras and Jürgen Schönwälder. On the Difference between Information Models and Data Models. Internet informational RFC 3444, January 2003.

[4] Erik Wilde. Making the Infoset Extensible. In *Proceedings of XML 2002*, Baltimore, Maryland, December 2002.

[5] Erik Wilde. A Proposal for XLink Infoset Contributions. Technical Report TIK-Report No. 148, Computer Engineering and Networks Laboratory, Swiss Federal Institute of Technology, Zürich, Switzerland, August 2002.

[6] Erik Wilde. The Extensible XML Information Set. Technical Report TIK-Report No. 160, Computer Engineering and Networks Laboratory, Swiss Federal Institute of Technology, Zürich, Switzerland, February 2003.