Standard format for document exchange and archiving

Discussion paper for comments

Draft: 2009-06-02/Tommi Karttaavi, Martti Poutanen and Juha Vartiainen

Problem: Organisations and individuals should be able to exchange documents regardless of the software the documents have been created with. On the other hand, certain documents need to be archived for long periods time, even hundred years and over. There is also a need for adding semantic information to the documents in a machine readable format. For example, there could be a classified section in an otherwise public document. The confidential part of the document needs to be marked, so that it will not be presented in the published version. In this paper the ability to exchange documents is focused on data exchange between information systems, not office applications.

Background: Document content is basically a flat list of objects of various typese.g.:

- paragraph-level (blocks) objects
- character-level (inline) objects
- tables
- graphics, equations

Word processing applications apply **styles** to these objects and formats them to pages. Pages typically have header/footer-type structures that obtain their content from the actual content of the document or from the metadata.

Styling for example use of headings is used to indicate the hierarchy of the information to the human reader; all text after, say, Heading 1 belongs under it until another text block styled as Heading 1 is encountered.

Working parser applications with defined DTD's for practically any pword processing -documents have been succesfully developed over time. A good example is the Rainbow DTD (SGML) from Electronic Book Technologies (EBT). Rainbow DTD was powerful enough to describe any document as a list of objects of different type with arbitrary formatting properties. This format was used by an application called DynaTag, that was a rule-based WP to SGML converter for Interleaf, Framemaker, WordPerfect and MS Word.

Current document format standards (OOXML, ODF) do not take these considerations into account in a meaningful way:

- they are focused on the presentation of the document on the expense of the content.
- the XML format of the document must also support the full functionality of the editor, which can make the XML structure very complex and prone to changes when the application evolves.
- the complexity may cause information loss over decades of storage

Scope: The standard format aims to provide following functionality:

- low cost to adopt for organizations and easy to use using existing software.
- format with full editing capabilities
- possibility to store sematic context

- amenable to full text search and semantic search
- classification of selected part of the document

The scope excludes the following functionality

- the detailed styling and appearance and paging of the document
- the data, macros and active components that are employed to obtain the output document.

Proposed solution: There should be a XML-based document standard that keeps the document in the simplest possible format without layout information. The document semantics are captured to proper metadata model that stores the document type, author, dates etc. Document type could be used to re-create the semantics (and the styling) of the generic content elements (e.g. sect1/title in a board meeting memo).

The documents are produced using existing word processing software. The flat list of document objects could be transformed into standard compliant structured document with XSLT using a rule-based tranformation.

The standard might be based on some existing specification, such as DocBook (http://www.docbook.org/), that would capture the document structure.
The semantic level would then be added on top of that as a metadata layer derived from document properties, for example. This means that the DocBook schema would need to be enhanced with semantic mark-up elements.
Alternative approaches:

Using custom schemata
The standard could be expressed as an XML-schema that can be used as a custom schema in the most common document editing tools (MS Office, OpenOffice). The downside is that this approach depends on users using the predefined styles and document templates in a diciplined manner.

Using external converters
Using converter applications that transform the document (eg. an OOXML or ODF document) to the specified format is an alternative to using custom schemata within the application that the document is created with. This approach has the same downside as the first one. Furthermore, external converters are liable to make the process more complex and vulnerable for errors.

Using input forms
This approach would be the best way to ensure that documents are compliant with the schema, but not very user friendly.

Conclusions: an international project to define the base XML-schema for document exchange and archiving should be started. The possibility to use DocBook as the starting point should be explored.

Contact:
Juha Vartiainen
Technical Adviser, PhD
IT-standardisation
Finnish Standards Association SFS

PB 116
FI-00241 HELSINKI
FINLAND
email juha.vartiainen@sfs.fi telephone + 358 40 356 7255telefax +358 9 146 4925