WAP Binary XML Content Format

Document id WAP-192-WBXML-20000515

Version 1.3

Approved Version 15th May 2000

This Document **Document Identifier 192**
Date **15th May 2000**
Subject: **Version 1.3 WBXML**

Wireless Application Protocol
Binary XML Content Format Specification

Disclaimer:
This document is subject to change without notice.

Contents

1. Scope

Wireless Application Protocol (WAP) is a result of continuous work to define an industry-wide specification for developing applications that operate over wireless communication networks. The scope of the WAP Forum is to define a set of specifications to be used by service applications. The wireless market is growing very quickly and reaching new customers and services. To enable operators and manufacturers to meet the challenges in advanced services, differentiation and fast/flexible service creation, WAP defines a set of protocols in transport, session and application layers. For additional information on the WAP architecture, refer to "*Wireless Application Protocol Architecture Specification*" [WAP].

This specification defines a compact binary representation of the Extensible Markup Language [XML]. The binary XML content format is designed to reduce the transmission size of XML documents, allowing more effective use of XML data on narrowband communication channels. Refer to the [WML] specification for one example use of the binary XML content format.

The binary format was designed to allow for compact transmission with no loss of functionality or semantic information. The format is designed to preserve the element structure of XML, allowing a browser to skip unknown elements or attributes. The binary format encodes the parsed physical form of an XML document, i.e., the structure and content of the document entities. Meta-information, including the document type definition and conditional sections, is removed when the document is converted to the binary format.

2. Document Status

This document is available online in the following formats:

- PDF format at http://www.wapforum.org/.

2.1 Copyright Notice

2.2 Errata

Known problems associated with this document are published at http://www.wapforum.org/.

2.3 Comments

Comments regarding this document can be submitted to the WAP Forum in the manner published at http://www.wapforum.org/.

2.4 Document History

Document ID	Date	Version
WAP-104	1998 04 30	1.0
WAP-135	1999 06 16	1.1
WAP-154	1999 11 04	1.2
WAP-192 and WAP-192.100 SCD	2000 02 19 and 2000 05 17	1.3 Draft and SCD
WAP-192-WBXML-20000515	2000 07 18	1.3 Approved

2.5 Changes in this version

Change 1:
The specification currently asserts that there are reserved tokens, which there are not. In addition, the comment on the reserved code page 255 is buried in the wrong section instead.
The following section(s) is known to be impacted:
5.8 - Token Structure: Move reserved code page discussion from 5.8.4.8 (change #1).
5.8.5.8 - Reserved Tokens: Delete section; there are no reserved tokens. (change #2).
Change 2:
The literal tokens LITERAL_A, LITERAL_C, and LITERAL_AC are missing from the BNF and discussion.
The following section(s) is known to be impacted:
5.3 - BNF for Document Structure: Add additional literal tags to BNF (change #1).
5.8.2 - Tag Code Space: Add additional literal tags to discussion (change #2).
5.8.4.5 - Literal Tag or Attribute Name: Add additional literal tags to discussion (change #3).
7.1 - Global Tokens: Expand comments for literal tags (change #4).
Change 3:
It may not be clear from the description what an encoder may or may not do.
The following section(s) is known to be impacted:
6.3 Encoding Default Attribute Values: Addition of clarifying text. (change #1).
Change 4:
Because there is a context sensitivity when using switch_page in front of an
extension, its effect requires clarification.
Sections affected, and additional explanation of details of change

…5.8.4.2 Global Extension Tokens extension = [switchPage] ((EXT_I termstr)|(EXT_T index)|EXT)
Change 5:
How the XML tokenizer processes white space characters depends on the XML application. For example, WML has its own rules for white space handling. Another XML application may have different rules. The current WBXML specification says that "insignificant" white space can be "altered or removed". This is very misleading. The tokenizer is only allowed to remove white space if the XML applications allows for it. And to "alter" or change white space is probably never legal.
Change 6:
Additional clarification on the MAY MUST SHOULD definitions in section 4.1

Change 7:
CRs to add 3 new public identifiers, 2 for WTA and 1 for provisioning:
These were submitted as
CREC-WBXML-ERICSSON-14-Apr-2000.5.doc
CREC-WBXML-ERICSSON-08-MAY-2000.6.doc
CR-Provisioning-WBXML-24-Mar-2000.pdf

Change 8:
Editorial change: Added section for document history to reflect use of document ID's

Change 9:
Document now specification at approved status. Doc name WAP-192-WBXML-20000515

3. References

3.1 Normative References

[IANACharset] IANA MIBEnum Character Set Registry,
URL: ftp://ftp.isi.edu/in-notes/iana/assignments/character-sets

[ISO10646] "Information Technology - Universal Multiple-Octet Coded Character Set (UCS) - Part 1: Architecture and Basic Multilingual Plane", ISO/IEC 10646-1:1993.

[RFC822] "Standard for the Format of ARPA Internet Text Messages", STD 11, RFC 822, D. Crocker, August 1982. URL: http://www.ietf.org/rfc/rfc822.txt

[RFC2119] "Key words for use in RFCs to Indicate Requirement Levels", S. Bradner, March 1997.
URL: http://www.ietf.org/rfc/rfc2119.txt

[WAP] "Wireless Application Protocol Architecture Specification", WAP Forum, 30-April-1998.
URL: http://www.wapforum.org/

[XML] "Extensible Markup Language (XML), W3C Proposed Recommendation 10-February-1998, REC-xml-19980210", T. Bray, et al, February 10, 1998. URL: http://www.w3.org/TR/REC-xml

3.2 Informative References

[ISO8879] "Information Processing - Text and Office Systems - Standard Generalised Markup Language (SGML)", ISO 8879:1986.

[UNICODE] "The Unicode Standard: Version 2.0", The Unicode Consortium, Addison-Wesley Developers Press, 1996. URL: http://www.unicode.org/

[WML] "Wireless Markup Language Specification", WAP Forum, 4-November-1999.
URL: http://www.wapforum.org/

4. Definitions and Abbreviations

4.1 Definitions

The following are terms and conventions used throughout this specification.
The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY" and "OPTIONAL" in this document are to be interpreted as described in [RFC2119]. In the absence of any such terms, the specification should be interpreted as "MUST".
Author - an author is a person or program that writes or generates WML, WMLScript or other content.
Content - subject matter (data) stored or generated at an origin server. Content is typically displayed or interpreted by a user agent in response to a user request.
Resource - a network data object or service that can be identified by a URL. Resources may be available in multiple representations (e.g., multiple languages, data formats, size and resolutions) or vary in other ways.
SGML - the Standardised Generalised Markup Language (defined in [ISO8879]) is a general-purpose language for domain-specific markup languages.
User - a user is a person who interacts with a user agent to view, hear or otherwise use a resource.
User Agent - a user agent is any software or device that interprets the binary XML content format or other resources. This may include textual browsers, voice browsers, search engines, etc.
XML - the Extensible Markup Language is a World Wide Web Consortium (W3C) standard for Internet markup languages, of which WML is one such language. XML is a restricted subset of SGML.

4.2 Abbreviations

For the purposes of this specification, the following abbreviations apply.

API	Application Programming Interface
BNF	Backus-Naur Form
LSB	Least Significant Bits
MSB	Most Significant Bits
MSC	Mobile Switch Centre
RFC	Request For Comments
SGML	Standardised Generalised Markup Language [ISO8879]
UCS-4	Universal Character Set - 4 byte [ISO10646]
URL	Universal Resource Locator
UTF-8	UCS Transformation Format 8 [ISO10646]
W3C	World Wide Web Consortium
WAP	Wireless Application Protocol [WAP]
WML	Wireless Markup Language [WML]
XML	Extensible Markup Language [XML]

5. Binary XML Content Structure

The following data types are used in the specification of the XML tokenised format.

Table 1. Data types used in tokenised format

Data Type	*Definition*
bit	1 bit of data
byte	8 bits of opaque data
u_int8	8 bit unsigned integer
mb_u_int32	32 bit unsigned integer, encoded in multi-byte integer format.

Network byte order is "big-endian". In other words, the most significant byte is transmitted on the network first, followed by the less significant bytes. Network bit ordering within a byte is "big-endian". In other words, bit fields described first are placed in the most significant bits of the byte.

Conformance Rules:

WBXML-1. Binary XML Structure M

5.1 Multi-byte Integers

This encoding uses a multi-byte representation for integer values. A multi-byte integer consists of a series of octets, where the most significant bit is the *continuation* flag and the remaining seven bits are a scalar value. The continuation flag indicates that an octet is not the end of the multi-byte sequence. A single integer value is encoded into a sequence of `N` octets. The first `N-1` octets have the continuation flag set to a value of one (`1`). The final octet in the series has a continuation flag value of zero (`0`).

The remaining seven bits in each octet are encoded in a big-endian order, e.g., most significant bit first. The octets are arranged in a big-endian order, e.g., the most significant seven bits are transmitted first. In the situation where the initial octet has less than seven bits of value, all unused bits must be set to zero (`0`).

For example, the integer value `0xA0` would be encoded with the two-byte sequence `0x81 0x20`. The integer value `0x60` would be encoded with the one-byte sequence `0x60`.

5.2 Character Encoding

In the absence of information provided by an external protocol (e.g. WSP, HTTP or MIME), the WBXML document must be presented to the application (e.g. a WML user agent) in the encoding specified in the WBXML charset field (see section 5.6). If the value of the charset field is set to "unknown" (the value is "0") no information about the character encoding is provided.

When a WBXML document is accompanied by external information (e.g. WSP, HTTP or MIME) there may be multiple sources of information available to determine the character encoding. In this case, their relative priority and the preferred method of handling conflict should be specified as part of the higher-level protocol; for example, see the documentation of the "application/vnd.wap.wbxml" MIME media type.

The XML binary representation can support any string encoding, but requires that all strings include an encoding-specific termination character (e.g., a `NULL` terminator) which can be reliably used to detect the end of a string. If a character encoding includes a `NULL` (e.g., Unicode, ASCII, ISO-8859-1, etc.), the `NULL` character must be used as the termination character.

If a tag name or attribute name can not be represented in the target character set, tokenisation terminates with an error.

5.3 BNF for Document Structure

A binary XML document is composed of a sequence of elements. Each element may have zero or more attributes and may contain embedded content. This structure is very general and does not have explicit knowledge of XML element structure or semantics. This generality allows user agents and other consumers of the binary format to skip elements and data that are not understood.

The following is a BNF-like description of the tokenised structure. The description uses the conventions established in [RFC822], except that the "`|`" character is used to designate alternatives and capitalised words indicate single-byte

tokens, which are defined later. Briefly, "(" and ")" are used to group elements, optional elements are enclosed in "[" and "]". Elements may be preceded with <N>* to specify N or more repetitions of the following element (N defaults to zero when unspecified).

```
start        = version publicid charset strtbl body
strtbl       = length *byte
body         = *pi element *pi
element      = ([switchPage] stag) [ 1*attribute END ] [ *content END ]

content      = element | string | extension | entity | pi | opaque

stag         = TAG |  (literalTag index)
literalTag   = LITERAL | LITERAL_A | LITERAL_C | LITERAL_AC
attribute    = attrStart *attrValue
attrStart    = ([switchPage] ATTRSTART) | ( LITERAL index )
attrValue    = ([switchPage] ATTRVALUE) | string | extension | entity | opaque

extension    = [switchPage] (( EXT_I termstr ) | ( EXT_T index ) | EXT)

string       = inline | tableref
switchPage   = SWTICH_PAGE pageindex
inline       = STR_I termstr
tableref     = STR_T index

entity       = ENTITY entcode
entcode      = mb_u_int32    // UCS-4 character code

pi           = PI attrStart *attrValue END

opaque       = OPAQUE length *byte

version      = u_int8 // WBXML version number
publicid     = mb_u_int32 | ( zero index )
charset      = mb_u_int32
termstr      = charset-dependent string with termination
index        = mb_u_int32    // integer index into string table.
length       = mb_u_int32    // integer length.
zero         = u_int8        // containing the value zero (0)
pageindex    = u_int8
```

5.4 Version Number

```
version      = u_int8 // WBXML version number
```

All WBXML documents contain a version number in their initial byte. This version specifies the WBXML version. The version byte contains the major version minus one in the upper four bits and the minor version in the lower four bits. For example, the version number 2.7 would be encoded as 0x17. This document specifies WBXML version 1.2, and will be encoded as 0x02.

5.5 Document Public Identifier

```
publicid     = mb_u_int32 | ( zero index )
zero         = u_int8        // containing the value zero (0)
```

The binary XML format contains a representation of the XML document public identifier. This publicid is used to identify the well-known document type contained within the WBXML entity.

The first form of publicid is a multi-byte positive integer value, greater than zero, representing a well-known XML document type (e.g., -//WAPFORUM//DTD WML 1.0//EN).

mb_u_int32

Public identifiers may also be encoded as strings, in the situation where a pre-defined numeric identifier is not available.

0	index

See section 7.2 for numeric constants related to public identifiers.

5.6 Charset

```
charset      = mb_u_int32
```

The binary XML format contains a representation of the XML document character encoding. This is the WBXML equivalent of the XML encoding declaration. The value of the WBXML charset field is the MIBEnum value assigned by the IANA for the character encoding ((see [IANACharset])). For example, IANA has assigned iso-8859-1 a MIBEnum value of 4, and shift_JIS a MIBEnum value of 17. In the case a character encoding is not registered at IANA, and thus does not have a MIBenum value, the value of the charset field is "0"; that is, "unknown" (the charset field does not allow textual representation of the name of the character encoding). Most character encodings are registered at IANA and do have a MIBenum value.

It is strongly recommended that WBXML tokenisers avoid using the charset value "0". If the XML encoding declaration is not present, the character encoding is either UTF-8 or UTF-16, and the charset field should be set to the MIBenum value defined for the appropriate character encoding. The value "0" should not be used in this case.

5.7 String Table

```
strtbl       = length *byte
```

A binary XML document must include a string table immediately after the charset. Minimally, the string table consists of a `mb_u_int32` encoding the string table length in bytes, not including the length field (e.g., a string table containing a two-byte string is encoded with a length of two). If the length is non-zero, one or more strings follow. The encoding of the strings should be determined by the process specified in 5.6.

Various tokens encode references to the contents of the string table. These references are encoded as scalar byte offsets from the first byte of the first string in the string table. For example, the offset of the first string is zero (`0`).

5.8 Token Structure

Tokens are split into a set of overlapping *code spaces*. The meaning of a particular token is dependent on the context in which it is used. Tokens are organised in the following manner:

- There are two classifications of tokens: global tokens and application tokens.
- Global tokens are assigned a fixed set of codes in all contexts and are unambiguous in all situations. Global codes are used to encode inline data (e.g., strings, entities, opaque data, etc.) and to encode a variety of miscellaneous control functions.
- Application tokens have a context-dependent meaning and are split into two overlapping *code spaces*. These two code spaces are the *tag code space* and the *attribute code space*. A given token value (e.g., `0x99`) will have a different meaning depending on whether it represents a token in the tag or attribute code space.
- The tag code space represents specific tag names. Each tag token is a single-byte code and represents a specific tag name (e.g., `CARD`).
- The attribute code space is split into two numeric ranges representing attribute prefixes and attribute values respectively.

Each code space is further split into a series of 256 *code pages*. Code pages allow for future expansion of the well-known codes. A single token (`SWITCH_PAGE`) switches between the code pages. The code page 255 is reserved for implementation-specific or experimental use. The tokens in this code page will never be used to represent standard XML document constructs.

The definition of tag and attribute codes is document-type-specific. Global codes are divided between a generic set of codes common to all document types and a set reserved for document-type-specific extensions.

5.8.1 Parser State Machine

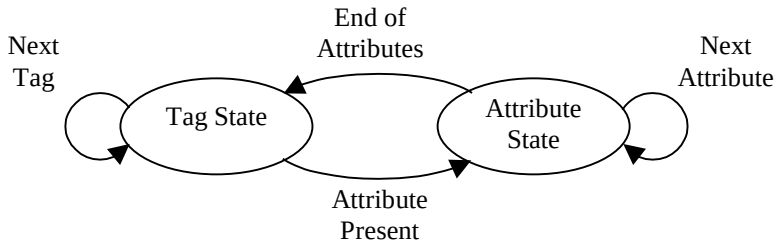
The tokenised format has two states, each of which has an associated code space. The states are traversed according to the syntax described in section 5.3. Code spaces are associated with parser states in the following manner:

Table 2. Parser states

Parser State	***Code Space***
`stag`	Tags
`attribute`	Attributes

Any occurrence of code page switch tokens (`SWITCH_PAGE`) while in a given state changes the current code page for that state. This new code page remains as the current code page until another `SWITCH_PAGE` is encountered in the same state or the document end is reached. Each parser state maintains a separate "current code page". The initial code page for both parser states is zero (`0`).

The following state machine is an alternative representation of the state transitions and is provided as a reference model.



5.8.2 Tag Code Space

Tag tokens are a single `u_int8` and are structured as follows:

Table 3. Tag format

Bit(s)	***Description***
7 (most significant)	Indicates whether attributes follow the tag code. If this bit is zero, the tag contains no attributes. If this bit is one, the tag is followed immediately by one or more attributes. The attribute list is terminated by an END token.
6	Indicates whether this tag begins an element containing content. If this bit is zero, the tag contains no content and no end tag. If this bit is one, the tag is followed by any content it contains and is terminated by an END token.
5 - 0	Indicates the tag identity.

For example:

- Tag value `0xC6`: indicates tag six (`6`), with both attributes and content following the tag, e.g.,
  `<TAG arg="1">foo</TAG>`
- Tag value `0x46`: indicates tag six (`6`), with content following the start tag. This element contains no attributes, e.g.,
  `<TAG>test</TAG>`
- Tag value `0x06`: indicates tag six (`6`). This element contains no content and has no attributes, e.g.,
  `<TAG/>`

The globally unique codes LITERAL, LITERAL_A, LITERAL_C, and LITERAL_AC (see section 5.8.4.5) represent unknown tag names. (Note that the tags LITERAL_A, LITERAL_C, and LITERAL_AC are the LITERAL tag with the appropriate combinations of bits 6 and 7 set.) An XML tokeniser should avoid the use of the `literal` or string representations of a tag when a more compact form is available.
Tags containing both attributes and content always encode the attributes before the content.

5.8.3 Attribute Code Space (ATTRSTART and ATTRVALUE)

Attribute tokens are encoded as a single `u_int8`. The attribute code space is split into two ranges (in addition to the global range present in all code spaces):

- Attribute Start - tokens with a value less than `128` indicate the start of an attribute. The attribute start token fully identifies the attribute name, e.g., `URL=`, and may optionally specify the beginning of the attribute value, e.g., `PUBLIC="TRUE"`. Unknown attribute names are encoded with the globally unique code `LITERAL` (see section 5.8.4.5). `LITERAL` must not be used to encode any portion of an attribute value.
- Attribute Value - tokens with a value of `128` or greater represent a well-known string present in an attribute value. These tokens may only be used to represent attribute values. Unknown attribute values are encoded with string, entity or extension codes (see section 5.8.4).

All tokenised attributes must begin with a single attribute start token and may be followed by zero or more attribute value, string, entity, opaque, or extension tokens. An attribute start token, a `LITERAL` token or the `END` token indicates the end of an attribute value. This allows a compact encoding of strings containing well-known sub-strings and entities.
For example, if the attribute start token `TOKEN_URL` represents the attribute name "`URL`", the attribute value token `TOKEN_COM` represents the string "`.com`" and the attribute value token `TOKEN_HTTP` represents the string "`http://`", the attribute `URL="http://foo.com/x"` might be encoded with the following sequence:

```
TOKEN_URL TOKEN_HTTP STR_I "foo" TOKEN_COM STR_I "/x"
```

In another example, if the attribute start token `TOKEN_PUBLIC_TRUE` represents the attribute name "`PUBLIC`" and the value prefix "`TRUE`", the attribute `PUBLIC="TRUE"` might be encoded with the following sequence:

```
TOKEN_PUBLIC_TRUE
```

An XML tokeniser should avoid the use of the `LITERAL` or string representations of an attribute name when a more compact form is available. An XML tokeniser should avoid the use of string representations of a value when a more compact form is available.

5.8.4 Global Tokens

Global tokens have the same meaning and structure in all code spaces and in all code pages. The classes of global tokens are:

- Strings - inline and table string references
- Extension - document-type-specific extension tokens
- Opaque - inline opaque data
- Entity - character entities
- Processing Instruction - XML PIs
- Literal - unknown tag or attribute name
- Control codes - miscellaneous global control tokens

5.8.4.1 Strings

```
string      = inline | tableref
inline      = STR_I termstr
tableref    = STR_T index
```

Strings encode inline character data or references into a string table. The string table is a concatenation of individual strings. String termination is dependent on the character document encoding and should not be presumed to include NULL termination. References to each string include an offset into the table, indicating the string being referenced. Inline string references have the following format:

STR_I	… char data …

String table references have the following format:

STR_T	mb_u_int32

The string table offset is from the first byte of the first string in the table (i.e., not a character offset). An empty string ("") must be explicitly encoded as string termination byte sequence for the string's character data. For example, an inlined null-terminated UTF8 empty string would be encoded as the byte sequence "03 00". Empty string encoding rules only apply to attribute values where there is no encoding defined by the application.

5.8.4.2 Global Extension Tokens

```
extension    = [switchPage] (( EXT_I termstr ) | ( EXT_T index ) | EXT)
```

The global extension tokens are available for document-specific use. The semantics of the tokens are defined only within the context of a particular document type, but the format is well defined across all documents. There are three classes of global extension tokens: single-byte extension tokens, inline string extension tokens and inline integer extension tokens.

Inline string extension tokens (`EXT_I*`) have the following format:

EXT_I*	… char data …

Inline integer extension tokens (`EXT_T*`) have the following format:

EXT_T*	mb_u_int32

Single-byte extension tokens (`EXT*`) have the following format:

EXT*

The effect of a switchPage preceding an extension will depend upon where the extension appears. If switchPage appears in content, it will change the tag code page. If switchPage appears in an attribute list, it will change the attribute code page.

5.8.4.3 Character Entity

```
entity       = ENTITY entcode
entcode      = mb_u_int32    // UCS-4 character code
```

The character entity token (`ENTITY`) encodes a numeric character entity. This has the same semantics as an XML numeric character entity (e.g., `&#32;`). The `mb_u_int32` refers to a character in the UCS-4 character encoding. All entities in the source XML document must be represented using either a string token (e.g., `STR_I`) or the `ENTITY` token.

The format of the character entity is:

ENTITY	mb_u_int32

5.8.4.4 Processing Instruction

```
pi           = PI attrStart *attrValue END
```

The processing instruction (PI) token encodes an XML processing instruction. The encoded PI has identical semantics to an XML PI. The `attrStart` encodes the `PITarget` and the `attrValue` encodes the PI's optional value. For more details on processing instructions, see [XML].

The format of the PI tag is:

PI	attrStart	attrValue	END

PIs without a value are encoded as:

PI	attrStart	END

5.8.4.5 Literal Tag or Attribute Name

The LITERAL token encodes a tag or attribute name that does not have a well-known token code. The actual meaning of the token (i.e., tag versus attribute name) is determined by the token parsing state. The tokens LITERAL_A, LITERAL_C, and LITERAL_AC are used when the tag posesses respectively attributes, content, or both. All literal tokens indicate a reference into the string table, which contains the actual name.

The format of the literal tags are:

LITERAL*	mb_u_int32

5.8.4.6 Opaque Data

```
opaque      = OPAQUE length *byte
```

The opaque token (OPAQUE) encodes application-specific data. A length field and zero or more bytes of data follow the token. The length field encodes the number of bytes of data, excluding the OPAQUE token and the length field.

OPAQUE	mb_u_int32	… bytes …

5.8.4.7 Miscellaneous Control Codes

5.8.4.7.1 END Token

The END token is used to terminate attribute lists and elements. END is a single-byte token.

5.8.4.7.2 Code Page Switch Token

```
switchPage  = SWTICH_PAGE pageindex
pageindex   = u_int8
```

The code-page switch token (SWITCH_PAGE) indicates a switch in the current code page for the current token state. The code-page switch is encoded as a two-byte sequence:

SWITCH	u_int8

6. Encoding Semantics

6.1 Document Tokenisation

The process of tokenising an XML document must convert all markup and XML syntax (i.e., entities, tags, attributes, etc.) into their corresponding tokenised format. All comments must be removed. Processing directives intended for the tokeniser may be removed. Other meta-information, such as the document type definition and unnecessary conditional sections must be removed. All text and character entities must be converted to string (e.g., `STR_I`) or entity (`ENTITY`) tokens. Character entities in the textual markup (e.g., `&`) must be converted to string form when tokenised, if the target character encoding can represent the entity. Characters present in the textual form may be encoded using the `ENTITY` token when they can not be represented in the target character encoding. XML parsed entities (both internal and external) must be resolved before tokenisation. XML notations and unparsed entities are resolved on an application basis (e.g., using inline opaque data). Attribute names must be converted to an attribute start token or must be represented by a single `LITERAL` token. Attribute values may not be represented by a `LITERAL` token.
It is illegal to encode markup constructs as strings. The user agent must treat all text tokens (e.g., `STR_I` and `ENTITY`) as CDATA, i.e., text with no embedded markup.

To process white space characters correctly, the tokenizer must be aware of the white space rules for the particular XML application being tokenized. If the tokenizer does not recognize the application, all white space must be preserved. In elements with the "xml:space" attribute set to "preserve", white space must be left alone. To determine the value of the "xml:space" attribute, the tokenizer must read the DTD.
For example, white space characters may be removed from a WML document, if removal is done according to the white space processing rules in the WML specification.

Conformance Rules:		
WBXML-2.	Conversion of all XML mark-up, excluding unparsed entities, into tokens	M
WBXML-3.	Removal of Processor Instructions	O
WBXML-4.	Removal of all information not covered in SC16, and SC17	M
WBXML-5.	Conversion of all text into String or Entity tokens	M
WBXML-6.	Conversion of all XML parsed entities into string or entity tokens	M
WBXML-7.	Conversion of all XML unparsed entities into string or entity tokens	O

6.2 Document Structure Conformance

The tokenised XML document must accurately represent the logical structure, as defined by [XML], and semantics of the textual source document. This implies that the source document must be well-formed, as defined in [XML]. Document tokenisation may validate the document as specified in [XML], but this is not required. If the semantics of a particular DOCTYPE are well known, additional semantic checks may be applied during the tokenisation process.

Conformance Rules:		
WBXML-8.	Checking that document is well-formed	M
WBXML-9.	Document validation	O

6.3 Encoding Default Attribute Values

The tokenised representation of an XML document may omit any attributes whose values match the default or implied value. For example,

given the following DTD

```
<!ELEMENT alpha EMPTY>
<!ATTLIST alpha
```

```
init (true|false) "false"
xml:space CDATA #FIXED "preserve"
xml:lang NMTOKEN #IMPLIED
>
```

A tokenized representation of a document containing the element <alpha init="false" xml:space="preserve" xml:lang="en"/> may omit the init and xml:space attributes. The tokenized representation may omit the xml:lang attribute if the implied value is "en".

This implies that a user agent implementation must be aware of the attribute defaults of a given version of the DTD. This information can be inferred from the version number in the tokenised data format.

Conformance Rules:		
WBXML-10.	Encoding default attribute values	O

6.4 Associating XML Documents with WBXML Token Values

An external typing system must be used to associate XML documents with WBXML token values.
If the document is transported by WSP, HTTP, or SMTP, the MIME media type must be used. Since the token values are associated with the media type, and not a particular version of the document type definition, the tokeniser is independent of the document type version; and can tokenise any version of the document type. To ensure compatibility between different versions of user-agents and tokenisers, the user-agent must support both the binary token value and the literal value for all tags, attribute names, and attribute values.

Conformance Rules:		
WBXML-11.	Support both the binary token value and the literal value for all tags, attribute names, and attribute values.	M

7. Numeric Constants

7.1 Global Tokens

The following token codes are common across all document types and are present in all code spaces and all code pages. All numbers are in hexadecimal.

Table 4. Global tokens

Token Name	***Token***	***Description***
SWITCH_PAGE	0	Change the code page for the current token state. Followed by a single u_int8 indicating the new code page number.
END	1	Indicates the end of an attribute list or the end of an element.
ENTITY	2	A character entity. Followed by a `mb_u_int32` encoding the character entity number.
STR_I	3	Inline string. Followed by a `termstr`.
LITERAL	4	An unknown attribute name, or unknown tag posessing no attributes or content.Followed by a `mb_u_int32` that encodes an offset into the string table.
EXT_I_0	40	Inline string document-type-specific extension token. Token is followed by a `termstr`.
EXT_I_1	41	Inline string document-type-specific extension token. Token is followed by a `termstr`.
EXT_I_2	42	Inline string document-type-specific extension token. Token is followed by a `termstr`.
PI	43	Processing instruction.
LITERAL_C	44	An unknown tag posessing content but no attributes.
EXT_T_0	80	Inline integer document-type-specific extension token. Token is followed by a `mb_u_int32`.
EXT_T_1	81	Inline integer document-type-specific extension token. Token is followed by a `mb_u_int32`.
EXT_T_2	82	Inline integer document-type-specific extension token. Token is followed by a `mb_u_int32`.
STR_T	83	String table reference. Followed by a `mb_u_int32` encoding a byte offset from the beginning of the string table.
LITERAL_A	84	An unknown tag posessing attributes but no content.
EXT_0	C0	Single-byte document-type-specific extension token.
EXT_1	C1	Single-byte document-type-specific extension token.
EXT_2	C2	Single-byte document-type-specific extension token.
OPAQUE	C3	Opaque document-type-specific data.
LITERAL_AC	C4	An unknown tag posessing both attributes and content.

7.2 Public Identifiers

The following values represent well-known document type public identifiers. The first 128 values are reserved for use in future WAP specifications. All numbers are in hexadecimal.

Table 5. Public Identifiers

Value	***Public Identifier***
0	String table index follows; public identifier is encoded as a literal in the string table.
1	Unknown or missing public identifier.
2	"-//WAPFORUM//DTD WML 1.0//EN" (WML 1.0)
3 **DEPRECATED**	"-//WAPFORUM//DTD WTA 1.0//EN" (WTA Event 1.0)
4	"-//WAPFORUM//DTD WML 1.1//EN" (WML 1.1)

Value	***Public Identifier***
5	"-//WAPFORUM//DTD SI 1.0//EN" (Service Indication 1.0)
6	"-//WAPFORUM//DTD SL 1.0//EN" (Service Loading 1.0)
7	"-//WAPFORUM//DTD CO 1.0//EN" (Cache Operation 1.0)
8	"-//WAPFORUM//DTD CHANNEL 1.1//EN" (Channel 1.1)
9	"-//WAPFORUM//DTD WML 1.2//EN" (WML 1.2)
A	"-//WAPFORUM//DTD WML 1.3//EN" (WML 1.3)
B	"-//WAPFORUM//DTD PROV 1.0//EN" (Provisioning 1.0)
C	"-//WAPFORUM//DTD WTA-WML 1.2//EN" (WTA-WML 1.2)
D	"-//WAPFORUM//DTD CHANNEL 1.2//EN" (Channel 1.2)
E- 7F	Reserved

8. Encoding Examples

The following example encodings are for demonstration purposes only, and do not necessarily represent an optimal WBXML encoding.

8.1 A Simple XML Document

The following is an example of a simple tokenised XML document. It demonstrates basic element, string and entity encoding. Source document:

```
<?xml version="1.0"?>
<!DOCTYPE XYZ [
<!ELEMENT XYZ (CARD)+>
<!ELEMENT CARD (#PCDATA | BR)*>
<!ELEMENT BR EMPTY>
<!ENTITY nbsp " ">
]>
<XYZ>
  <CARD>
    X &amp; Y<BR/>
    X = 1
  </CARD>
</XYZ>
```

The following tokens are defined for the tag code space:

Tag Name	***Token***
BR	5
CARD	6
XYZ	7

Tokenised form (numbers in hexadecimal) follows. This example uses only inline strings and assumes that the character encoding uses a NULL terminated string format. It also assumes that the transport character encoding is US-ASCII. This encoding is incapable of supporting some of the characters in the deck (e.g., ` `), forcing the use of the `ENTITY` token.

```
03  01  03  00  47  46  03  ' ' 'X' ' ' '&' ' ' 'Y' 00  05  03  ' '
'X' 00  02  81  20  03  '=' 00  02  81  20  03  '1' ' ' 00  01  01
```

In an expanded and annotated form:

Table 6. Example tokenised deck

Token Stream	***Description***
03	Version number - WBXML version 1.3.
01	Unknown public identifier
03	charset=US-ASCII (MIBEnum is 3)
00	String table length
47	XYZ, with content
46	CARD, with content
03	Inline string follows
' ', 'X', ' ', '&', ' ', 'Y', 00	String
05	BR
03	Inline string follows
' ', 'X', 00	String
02	ENTITY
81 20	Entity value (160)
03	Inline string follows
'=', 00	String
02	ENTITY
81 20	Entity value (160)
03	Inline string follows

Token Stream	***Description***
'1', ' ', 00	String
01	END (of CARD element)
01	END (of XYZ element)

8.2 An Expanded Example

The following is another example of a tokenised XML document. It demonstrates attribute encoding and the use of the string table. Source document:

```
<?xml version="1.0"?>
<!DOCTYPE XYZ [
<!ELEMENT XYZ ( CARD )+ >
<!ELEMENT CARD (#PCDATA | INPUT | DO)*>
<!ATTLIST CARD NAME NMTOKEN #IMPLIED>
<!ATTLIST CARD STYLE (LIST|SET) 'LIST'>
<!ELEMENT DO EMPTY>
<!ATTLIST DO TYPE CDATA #REQUIRED>
<!ATTLIST DO URL CDATA #IMPLIED>
<!ELEMENT INPUT EMPTY>
<!ATTLIST INPUT TYPE (TEXT|PASSWORD)'TEXT'>
<!ATTLIST INPUT KEY NMTOKEN #IMPLIED>
<!ENTITY nbsp " ">
]>
<!-- This is a comment -->
<XYZ>
  <CARD NAME="abc" STYLE="LIST">
    <DO TYPE="ACCEPT" URL="http://xyz.org/s"/>
    Enter name: <INPUT TYPE="TEXT" KEY="N"/>
  </CARD>
</XYZ>
```

The following tokens are defined for the tag code space:

Tag Name	***Token***
CARD	5
INPUT	6
XYZ	7
DO	8

The following attribute start tokens are defined:

Attribute Name	***Attribute Value Prefix***	***Token***
STYLE	LIST	5
TYPE		6
TYPE	TEXT	7
URL	http://	8
NAME		9
KEY		A

The following attribute value tokens are defined:

Attribute Value	***Token***
.org	85
ACCEPT	86

Tokenised form (numbers in hexadecimal) follows. This example assumes an UTF-8 character encoding and NULL terminated strings:

```
03  01  6A  12  'a' 'b' 'c' 00 ' ' 'E' 'n' 't' 'e' 'r' ' ' 'n'
'a' 'm' 'e' ':' ' ' 00  47  C5  09  03  00  05  01  88  06
86  08  03  'x' 'y' 'z' 00  85  03  '/' 's' 00  01  83  04
86  07  0A  03  'N' 00  01  01  01
```

In an expanded and annotated form:

Table 7. Example tokenised deck

Token Stream	***Description***
`03`	Version number - WBXML version 1.3
`01`	Unknown public identifier
`6A`	charset=UTF-8 (MIBEnum is 106)
`12`	String table length
`'a', 'b', 'c', 00, ' ', 'E', 'n', 't', 'e', 'r', ' ', 'n', 'a', 'm', 'e', ':', ' ', 00`	String table
`47`	`XYZ`, with content
`C5`	`CARD`, with content and attributes
`09`	`NAME=`
`83`	String table reference follows
`00`	String table index
`05`	`STYLE="LIST"`
`01`	`END` (of `CARD` attribute list)
`88`	`DO`, with attributes
`06`	`TYPE=`
`86`	`ACCEPT`
`08`	URL="http://"
`03`	Inline string follows
`'x', 'y', 'z', 00`	string
`85`	`".org"`
`03`	Inline string follows
`'/', 's', 00`	string
`01`	`END` (of `DO` attribute list)
`83`	String table reference follows
`04`	String table index
`86`	`INPUT`, with attributes

Token Stream	***Description***
07	TYPE="TEXT"
0A	KEY=
03	Inline string follows
'N', 00	String
01	END (of INPUT attribute list)
01	END (of CARD element)
01	END (of XYZ element)

9. Static Conformance Requirements

This section defines static conformance requirements for WBXML documents, encoder, and decoder. The encoder is producing WBXML documents. The decoder is reading WBXML documents.

9.1 WBXML Document

Identifier	Structure	Reference	Mandatory/ Optional
WBXML-1.	Binary XML Structure	5	M

9.2 WBXML Encoder

If a WBXML encoder does not support an optional feature, the token stream produced may not be optimal, but any WBXML decoder will be able to interpret the tokens without errors.

Identifier	Encoding semantics	Reference	Mandatory/ Optional
WBXML-2.	Conversion of all XML mark-up, excluding unparsed entities, into tokens	6.1	M
WBXML-3.	Removal of Processor Instructions	6.1	O
WBXML-4.	Removal of all information not covered in SC16, and SC17	6.1	M
WBXML-5.	Conversion of all text into String or Entity tokens	6.1	M
WBXML-6.	Conversion of all XML parsed entities into string or entity tokens	6.1	M
WBXML-7.	Conversion of all XML unparsed entities into string or entity tokens	6.1	O
WBXML-8.	Checking that document is well-formed	6.2	M
WBXML-9.	Document validation	6.2	O
WBXML-10.	Encoding default attribute values	6.3	O

9.3 WBXML Decoder

Identifier	Decoding semantics	Reference	Mandatory/ Optional
WBXML-11.	Support both the binary token value and the literal value for all tags, attribute names, and attribute values.	6.4	M