INTERNATIONAL STANDARD

ISO/IEC 24824-1

First edition
2007-05-01

Information technology — Generic applications of ASN.1: Fast Infoset

Technologies de l'information — Applications génériques de ASN.1: Infoset rapide

Reference number
ISO/IEC 24824-1:2007(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

CONTENTS

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work. In the field of information technology, ISO and IEC have established a joint technical committee, ISO/IEC JTC 1.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of the joint technical committee is to prepare International Standards. Draft International Standards adopted by the joint technical committee are circulated to national bodies for voting. Publication as an International Standard requires approval by at least 75 % of the national bodies casting a vote.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

ISO/IEC 24824-1 was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 6, *Telecommunications and information exchange between systems,* in collaboration with ITU-T. The identical text is published as ITU-T Rec. X.891.

ISO/IEC 24824 consists of the following parts, under the general title *Information technology — Generic applications of ASN.1*:

— *Part 1: Fast infoset*

— *Part 2: Fast Web Services*

The following part is under preparation:

— *Part 3: Fast infoset security*

Introduction

This Recommendation | International Standard specifies a representation of an instance of the W3C XML Information Set using binary encodings (specified using the ASN.1 notation and the ASN.1 Encoding Control Notation). The encoding specified in this edition of this Recommendation | International Standard is identified by the version number 1 (see 12.9).

The technology specified in this Recommendation | International Standard is named Fast Infoset. It provides an alternative to W3C XML syntax as a means of representing instances of the W3C XML Information Set. This representation generally provides smaller encoding sizes and faster processing than a W3C XML representation.

The representation of an instance of the W3C XML Information Set specified in this Recommendation | International Standard is called a fast infoset document. Each fast infoset document is an encoding of an abstract value of an ASN.1 data type (the `Document` type – see 7.2) representing an instance of the W3C XML Information Set.

This Recommendation | International Standard specifies the use of several techniques that minimize the size of a fast infoset document and that maximize the speed of creating and processing such documents.

These techniques are based on the use of vocabulary tables, which allow typically-small integer values (vocabulary table indexes) to be used instead of character strings that form (for example) the names of elements or attributes in an XML 1.0 serialization of an instance of the W3C XML Information Set.

There are a number of vocabulary tables (see clause 8), of which the most basic (the eight character string tables) map typically-small integers to strings of characters. There are, however, also vocabulary tables (the element name table and the attribute name table) that provide a further level of indirection, with a vocabulary table index mapping to a set of three vocabulary table indexes, identifying a prefix, a namespace name, and a local name.

Another important technique is the use of a restricted alphabet vocabulary table. This contains entries that list a subset of ISO/IEC 10646 characters. If a character string needs to be encoded for which there is an entry in this table, then it can be encoded by identifying that this vocabulary table is being used, giving the vocabulary table index, and then encoding each character in the minimum number of bits needed for that particular subset of ISO/IEC 10646 characters. There are a number of built-in restricted alphabets that always form the first few entries of this table, covering such commonly occurring strings as dates and times, and numeric values.

A further important optimization uses the encoding algorithm vocabulary table. This table identifies specialized encodings that can be employed for commonly occurring strings, again with a number of built-in algorithms. For example, if there is a string which looks like the decimal representation of an integer in the range –32768 to 32767, then that string can be encoded by identifying that this vocabulary table is being used, giving the vocabulary table index, and then encoding the integer as a two-octet signed integer. Floating-point numbers and arrays of such numbers are supported in the same way.

In order to ensure fast processing without sacrificing compactness, many components of a fast infoset document (such as character strings and components representing information items of the XML infoset) are octet-aligned, while other components (such as lengths and vocabulary table indexes) are not necessarily octet-aligned but always end on the last bit of an octet. To provide a formal specification of these optimized encodings, the ASN.1 Encoding Control Notation (defined in ITU-T Rec. X.692 | ISO/IEC 8825-3) is used (see A.2), but use of ECN tools for implementation is not necessary and a complete description of the encoding is provided (see Annex C).

The vocabulary tables for a particular fast infoset document can be initialized by information at the head of the document, and are normally added to dynamically, providing flexibility for an encoder. The initial vocabulary tables can be provided by a reference to the set of final vocabulary tables of some other identified fast infoset document (or by other means). This vocabulary reference can then be supplemented by further table additions to provide the initial vocabulary tables for this document. Further dynamic additions are normally made to the tables during the creation or the processing of the document.

Finally, a mechanism is provided for the generator of a fast infoset document to include data (called additional processing data) related to optional additional processing of the fast infoset document, together with a URI that identifies a complete specification of the form and semantics of that additional processing data. The optional additional processing data is ignored by any subsequent processor of the fast infoset document if the URI is not known, or the processing that it specifies is not supported or not required.

NOTE – An example of such additional processing data would be data that provides indexes that enable immediate access to parts of the fast infoset document, so that the whole document need not be processed if the only interest is in those parts of the fast infoset document that correspond to a specific XML tag.

Annex A forms an integral part of this Recommendation | International Standard, and contains an ASN.1 module (see ITU-T Rec. X.680 | ISO/IEC 8824-1) and two ECN modules (EDM and ELM – see ITU-T Rec. X.692 | ISO/IEC 8825-3) which together specify the abstract content and the bit-level encoding of a value of the `Document` type, which conveys the value of an instance of the W3C XML Information Set.

Annex B forms an integral part of this Recommendation | International Standard, and contains the specification of a MIME media type identifying a fast infoset document.

Annex C does not form an integral part of this Recommendation | International Standard, and provides a complete description of the encodings formally specified in clause 12 and A.2.

Annex D does not form an integral part of this Recommendation | International Standard, and provides examples of fast infoset documents generated from some XML documents. Annex D also gives the size of the XML representation and the Fast Infoset representation of these examples.

INTERNATIONAL STANDARD
ITU-T RECOMMENDATION

Information technology – Generic applications of ASN.1: Fast infoset

1 Scope

This Recommendation | International Standard specifies an ASN.1 type (see ITU-T Rec. X.680 | ISO/IEC 8824-1) whose abstract values represent instances of the W3C XML Information Set. It also specifies binary encodings for those values, using ASN.1 Encoding Control Notation (see ITU-T Rec. X.692 | ISO/IEC 8825-3).

NOTE – These encodings are called fast infoset documents.

This Recommendation | International Standard also specifies techniques that:

– minimize the size of fast infoset documents;

– maximize the speed of creating and processing fast infoset documents;

– allow the specification (by the generator of a fast infoset document) of additional processing data.

The first two techniques involve the use of conceptual vocabulary tables. The set of vocabulary tables and the nature of their entries is fully defined in this Recommendation | International Standard, but their representation in computer memory is outside the scope of this Recommendation | International Standard. Provision for transfer or storage of, or a formal notation for displaying or specifying, vocabulary tables to be used as an external vocabulary is also outside the scope of this Recommendation | International Standard.

The third technique involves the provision of additional processing data and a URI that identifies the form and semantics of that data. The specification of specific forms of additional processing data and their use is outside the scope of this Recommendation | International Standard.

URIs can be used to identify final vocabularies that can be used as either part or all of some new initial vocabulary, but the assignment of specific URIs to specific final vocabularies is outside the scope of this Recommendation | International Standard.

This Recommendation | International Standard specifies built-in restricted alphabets, the addition to vocabulary tables of further restricted alphabets by enumeration, and the use of these vocabulary tables for efficient encoding of character strings.

This Recommendation | International Standard further specifies built-in encoding algorithms for the optimum encoding of certain character strings, and the addition to vocabulary tables of further encoding algorithms identified by URIs, but the definition of these further encoding algorithms and their associated URIs is outside the scope of this Recommendation | International Standard.

In addition, this Recommendation | International Standard specifies a Multipurpose Internet Mail Extensions (MIME) media type that identifies a fast infoset document.

2 Normative references

The following Recommendations, International Standards and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation | International Standard. At the time of publication, the editions indicated were valid. All Recommendations, International Standards and other references are subject to revision, and parties to agreements based on this Recommendation | International Standard are encouraged to investigate the possibility of applying the most recent edition of the Recommendations, International Standards and other references listed below. The Telecommunication Standardization Bureau of the ITU maintains a list of currently valid ITU-T Recommendations. Members of IEC and ISO maintain registers of currently valid International Standards. The IETF maintains a list of RFCs, together with those that have been obsoleted by later RFCs. The W3C maintains a list of currently valid W3C Recommendations. The reference to a document within this Recommendation | International Standard does not give it, as a stand-alone document, the status of a Recommendation or International Standard.

2.1 Identical Recommendations | International Standards

- ITU-T Recommendation X.667 (2004) | ISO/IEC 9834-8:2005, *Information technology – Open Systems Interconnection – Procedures for the operation of OSI Registration Authorities: Generation and registration of Universally Unique Identifiers (UUIDs) and their use as ASN.1 Object Identifier components.*
- ITU-T Recommendation X.680 (2002) | ISO/IEC 8824-1:2002, *Information technology – Abstract Syntax Notation One (ASN.1): Specification of basic notation.*
- ITU-T Recommendation X.681 (2002) | ISO/IEC 8824-2:2002, *Information technology – Abstract Syntax Notation One (ASN.1): Information object specification.* †
- ITU-T Recommendation X.682 (2002) | ISO/IEC 8824-3:2002, *Information technology – Abstract Syntax Notation One (ASN.1): Constraint specification.* †
- ITU-T Recommendation X.683 (2002) | ISO/IEC 8824-4:2002, *Information technology – Abstract Syntax Notation One (ASN.1): Parameterization of ASN.1 specifications.* †
- ITU-T Recommendation X.690 (2002) | ISO/IEC 8825-1:2002, *Information technology – ASN.1 encoding rules: Specification of Basic Encoding Rules (BER), Canonical Encoding Rules (CER) and Distinguished Encoding Rules (DER).* †
- ITU-T Recommendation X.691 (2002) | ISO/IEC 8825-2:2002, *Information technology – ASN.1 encoding rules: Specification of Packed Encoding Rules (PER).* †
- ITU-T Recommendation X.692 (2002) | ISO/IEC 8825-3:2002, *Information technology – ASN.1 encoding rules: Specification of Encoding Control Notation (ECN).*
- ITU-T Recommendation X.693 (2001) | ISO/IEC 8825-4:2002, *Information technology – ASN.1 encoding rules: XML Encoding Rules (XER).* †

NOTE – The complete set of ASN.1 Recommendations | International Standards are listed above, as they can all be applicable in particular uses of this Recommendation | International Standard. Where these are not directly referenced in the body of this Recommendation | International Standard, a † symbol is added to the reference.

2.2 Additional references

- ISO 8601:2004, *Data elements and interchange formats – Information interchange – Representation of dates and times.*
- ISO/IEC 10646:2003, *Information technology – Universal Multiple-Octet Coded Character Set (UCS).*
- *The Unicode Standard, Version 4.0,* The Unicode Consortium *(Reading, MA, Addison-Wesley).*

NOTE 1 – The graphics characters (and their encodings) defined by Unicode are identical to those defined by ISO/IEC 10646-1, but Unicode is included as a reference because it also specifies the names of control characters and defines the abbreviation UTF-16BE.

- W3C XML 1.0:2004, *Extensible Markup Language (XML) 1.0 (Third Edition), W3C Recommendation, Copyright © [4 February 2004] World Wide Web Consortium (Massachusetts Institute of Technology, Institut National de Recherche en Informatique et en Automatique, Keio University), http://www.w3.org/TR/2000/REC-xml-20040204/.*
- W3C XML 1.1:2004, *Extensible Markup Language (XML) 1.1, W3C Recommendation, Copyright © [4 February 2004] World Wide Web Consortium (Massachusetts Institute of Technology, Institut National de Recherche en Informatique et en Automatique, Keio University), http://www.w3.org/TR/2000/REC-xml11-20040204/.*

NOTE 2 – References to both W3C XML 1.0 and W3C XML 1.1 are included as neither is a subset of the other. These references are used solely in 3.4.10.

- W3C XML Information Set:2004, *XML Information Set (Second Edition), W3C Recommendation, Copyright © [04 February 2004] World Wide Web Consortium (Massachusetts Institute of Technology, Institut National de Recherche en Informatique et en Automatique, Keio University), http://www.w3.org/TR/2004/REC-xml-infoset-20040204/.*
- W3C XML Namespaces 1.0:1999, *Namespaces in XML, W3C Recommendation, Copyright © [14 January 1999] World Wide Web Consortium (Massachusetts Institute of Technology, Institut National de Recherche en Informatique et en Automatique, Keio University), http://www.w3.org/TR/1999/REC-xml-lnames-19990114/.*
- W3C XML Namespaces 1.1:2004, *Namespaces in XML 1.1, W3C Recommendation, Copyright © [4 February 2004] World Wide Web Consortium (Massachusetts Institute of Technology, Institut National de Recherche en Informatique et en Automatique, Keio University), http://www.w3.org/TR/2004/REC-xml-l-names11-20040204/.*

NOTE 3 – References to both W3C XML Namespaces 1.0 and W3C XML Namespaces 1.1 are included as neither is a subset of the other. These references are used solely in 3.4.10.

– IETF RFC 2045 (1996), *Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies.*

– IETF RFC 2396 (1998), *Uniform Resource Identifiers (URI): Generic Syntax.*

– IEEE 754-1985, *IEEE Standard for Binary Floating-Point Arithmetic.*

3 Definitions

For the purposes of this Recommendation | International Standard, the following definitions apply.

3.1 ASN.1 terms

This Recommendation | International Standard uses the following terms defined in ITU-T Rec. X.680 | ISO/IEC 8824-1:

a) choice type;

b) sequence type;

c) sequence-of type.

3.2 ECN terms

This Recommendation | International Standard uses the following terms defined in ITU-T Rec. X.692 | ISO/IEC 8825-3:

a) Encoding Definition Modules (EDM);

b) Encoding Link Module (ELM).

3.3 ISO/IEC 10646 terms

This Recommendation | International Standard uses the following term defined in ISO/IEC 10646:

a) Basic Multilingual Plane.

3.4 Additional definitions

3.4.1 Base64: An encoding mechanism that represents an octet string value as a character string using a restricted alphabet of 65 characters (see 10.3 and IETF RFC 2045).

3.4.2 character string: A string of ISO/IEC 10646 abstract characters, without any implication on the way they are encoded.

3.4.3 encoding algorithm: A precise specification of how to efficiently encode a character string with specified characteristics into octets.

NOTE – An example is the encoding of a string such as "-32176" into a two's complement binary integer in two octets. The two-octet encoding would be accompanied by a vocabulary table index identifying this encoding algorithm.

3.4.4 external vocabulary: A set of vocabulary tables referenced by a URI (see 7.2.14).

3.4.5 fast infoset document: An XML infoset represented as specified in this Recommendation | International Standard.

3.4.6 final vocabulary: The content of the vocabulary tables at the end of the creation or of the processing of a fast infoset document.

3.4.7 information item: Each of the kinds of items that constitute an XML infoset.

3.4.8 initial vocabulary: The set of vocabulary tables established by information at the head of a fast infoset document that optionally references an external vocabulary and optionally provides additional table entries.

3.4.9 name surrogate: A set of three vocabulary table indexes (the first two optional) that are used to represent a qualified name (see 3.4.11).

3.4.10 namespace-well-formed XML document: Either a W3C XML 1.0 document that is well-formed according to W3C XML Namespaces 1.0, or a W3C XML 1.1 document that is well-formed according to W3C XML Namespaces 1.1.

3.4.11 qualified name: The set consisting of the **[prefix]**, **[namespace name]**, and **[local name]** properties of an **element** information item or **attribute** information item.

3.4.12 restricted alphabet: An ordered set of distinct ISO/IEC 10646 characters, which permits a compact encoding of any character string that consists entirely of characters from that set.

3.4.13 vocabulary table index: A positive integer value identifying an entry in a vocabulary table.

3.4.14 vocabulary tables: A set of conceptual tables (typically, but not necessarily, dynamically constructed) associated with a fast infoset document, which contain character strings or other information, and support the use of typically-small positive integer values (vocabulary table indexes) identifying their entries.

NOTE – Examples of vocabulary tables are those containing character strings that are the **[local name]** property of **attribute** or **element** information items, or character strings corresponding to sequences of **character** information items that are members of the **[children]** property of **element** information items.

3.4.15 XML declaration: The UTF-8 encoding of a specified character string (see also 12.3) that may be included at the beginning of a fast infoset document to identify the encoding as a fast infoset document and to distinguish it from a W3C XML 1.0 or W3C XML 1.1 document.

3.4.16 XML infoset: An abstract data set describing the information in a namespace-well-formed XML document, as specified in W3C XML Information Set.

3.4.17 XML whitespace: One or more of the characters HORIZONTAL TABULATION (9), LINE FEED (10), CARRIAGE RETURN (13), or SPACE (32) of Unicode.

NOTE – These characters are those that match the production "S" in both W3C XML 1.0 and W3C XML 1.1 (see W3C XML 1.0, 2.3 and W3C XML 1.1, 2.3). The characters NEXT LINE (133) and LINE SEPARATOR (8232), which may occur in a namespace-well-formed W3C XML 1.1 document (see W3C XML 1.1, 2.11), are converted to LINE FEED characters by end-of-line handling (see W3C XML 1.1, 2.11). If those characters occur in an XML infoset generated from a namespace-well-formed W3C XML 1.1 document, they are not XML whitespace.

4 Abbreviations

For the purposes of this Recommendation | International Standard, the following abbreviations apply:

ASN.1	Abstract Syntax Notation One
BMP	Basic Multilingual Plane
ECN	Encoding Control Notation
MIME	Multipurpose Internet Mail Extensions
UBL	Universal Business Language
URI	Uniform Resource Identifier
UTF-8	Universal Transformation Function 8-bit (see ISO/IEC 10646, Annex D)
UTF-16BE	Universal Transformation Function 16-bit Big Endian (see Unicode, 2.6)
UUID	Universally Unique Identifier
XML	eXtensible Markup Language

5 Notation

5.1 This Recommendation | International Standard uses the ASN.1 notation defined by ITU-T Rec. X.680 | ISO/IEC 8824-1 for the formal definition of data types whose encodings are fast infoset documents.

NOTE – Clause 12 specifies the application of ITU-T Rec. X.692 | ISO/IEC 8825-3 to the ASN.1 type definitions, providing the bit-level encoding of a fast infoset document.

5.2 In this Recommendation | International Standard, **`bold Courier`** is used for ASN.1 notation and **bold Arial** is used for W3C XML syntax and for the names of information items of the XML Information Set.

5.3 The names of information items' properties are in **bold Arial** and enclosed between square brackets (for example, **[children]**).

5.4 The names of categories of character strings (see 8.4.2) and the names of categories of qualified names (see 8.5.4) are in UPPERCASE.

5.5 In this Recommendation | International Standard, bit positions within an octet are specified using the terminology first bit, second bit, etc., to eighth bit, where the first bit is the most significant bit of the octet, and the eighth bit is least significant bit of the octet.

6 Principles of vocabulary table construction and use

6.1 Vocabulary tables are conceptual tables mapping a vocabulary table index into a vocabulary table entry.

NOTE – The representation of vocabulary tables in computer memory is not defined, nor is the means by which an implementation maps a vocabulary table index into a vocabulary table entry for that table.

6.2 The creator of a fast infoset document from an XML infoset determines the contents of the vocabulary tables.

6.3 In the most general case, the head of a fast infoset document can reference a set of vocabulary tables (an external vocabulary), followed by the specification of additions to those vocabulary tables to form the initial vocabulary for this fast infoset document. Further additions to the vocabulary tables occur during the creation and during the processing of a fast infoset document, so that they incrementally grow to form the final vocabulary tables for that document.

6.4 Some vocabulary tables incrementally grow from an initial vocabulary to a final vocabulary during the creation and during the processing of a fast infoset document, and therefore have the word "dynamic" in the name of the vocabulary table. There are no mechanisms for entries to be removed from any table.

6.5 Vocabulary table indexes are implicitly assigned. The first entry to any vocabulary table has a vocabulary table index of one, and each subsequent entry to that table has the next higher integer value for the vocabulary table index. Where this Recommendation | International Standard specifies that something is to be added to a vocabulary table, this implies that the next available vocabulary table index shall be assigned.

NOTE – Vocabulary table indexes start at one and not zero because the value zero (when permitted) has the special meaning of "empty character string" in a field that might otherwise hold a vocabulary table index.

6.6 In order to support this implicit assignment of vocabulary table indexes, the conceptual order of processing the components (at any depth) of a fast infoset document is fully-defined (see 8.1).

NOTE – This order is the same as the order of the encodings of the components in a fast infoset document. It does not necessarily imply that the semantics carried by the document is processed in this order. The order is defined solely for the purposes of ensuring that the same vocabulary table index is assigned for any given vocabulary table entry by both the creator and the processor of a fast infoset document.

6.7 Vocabulary tables are used for many purposes (see clause 8), but their primary function is to enable the use of a vocabulary table index instead of a vocabulary table entry, where such indexes are smaller (and may be faster to process) than the table entry. A number of built-in entries for some vocabulary tables are specified in clause 9. These entries are always implicitly present in these vocabulary tables, with the vocabulary table indexes specified in clause 9.

6.8 For some categories of character string, the creator of a fast infoset document has the option of adding or not adding a string to a vocabulary table, depending on the expected (or known) number of occurrences of that character string in the XML infoset.

6.9 The precise form and meaning of vocabulary table entries is specified in clause 8, but they are in most cases variable length character strings, often short, but potentially as large as 2^{32} octets.

6.10 A conforming creator of a fast infoset document is required to do all the additions to the vocabulary tables as specified in 7.13.7, 7.14.6, 7.14.7, and 7.16.7. This ensures that the number of vocabulary table entries in each vocabulary table never exceeds 2^{20}.

NOTE – A vocabulary table entry may equal one or more other vocabulary table entries. This is in order to allow efficient creation of fast infoset documents. However, duplicate entries will decrease the efficiency of transfer. A processor is not affected by duplicate entries.

6.11 A conforming processor of a fast infoset document is required to do all the additions to the vocabulary tables as specified in 7.13.8, 7.14.11, and 7.16.8. This ensures that the restriction of 6.10 a has not been violated.

7 ASN.1 type definitions

7.1 General

7.1.1 This Recommendation | International Standard specifies a set of ASN.1 types supporting a representation of the XML Information Set. The root type of this set of types is the **Document** type.

7.1.2 Some restrictions are imposed on the content of the XML infosets and some simplifications are made in the representation (see clause 11) in order to improve the usability of the specification and the efficiency of the encodings produced with it.

NOTE – An XML infoset that does not meet those restrictions cannot be represented as a fast infoset document, nor can it normally be represented as a namespace-well-formed XML document.

7.1.3 For each kind of information item specified in W3C XML Information Set, a corresponding ASN.1 type definition is provided in this Recommendation | International Standard. This type definition is always a sequence type, with components corresponding to the properties of the information item.

7.1.4 Certain properties of information items are not included in the ASN.1 type definitions (see 11.4).

7.1.5 In some cases, the value of a property that is not included in the ASN.1 type definitions can be determined from the value of other properties of the same or other information items that are included. In these cases, the omission of that property simplifies the representation with no loss of information. There are, however, a few cases in which the value of a property that is not included cannot be determined from other properties. In all such cases, the omission of that property is a simplification that does not limit the utility of the specification for most practical use cases.

7.1.6 Clause 12 specifies the encoding of the **Document** type.

7.2 The Document type

7.2.1 The **Document** type is:


```
Document ::= SEQUENCE {
      additional-data          SEQUENCE (SIZE(1..one-meg)) OF
            additional-datum SEQUENCE {
                  id                        URI,
                  data                      NonEmptyOctetString } OPTIONAL,
      initial-vocabulary       SEQUENCE {
            external-vocabulary             URI OPTIONAL,
            restricted-alphabets            SEQUENCE (SIZE(1..256)) OF
                  NonEmptyOctetString OPTIONAL,
            encoding-algorithms             SEQUENCE (SIZE(1..256)) OF
                  NonEmptyOctetString OPTIONAL,
            prefixes                        SEQUENCE (SIZE(1..one-meg)) OF
                  NonEmptyOctetString OPTIONAL,
            namespace-names                 SEQUENCE (SIZE(1..one-meg)) OF
                  NonEmptyOctetString OPTIONAL,
            local-names                     SEQUENCE (SIZE(1..one-meg)) OF
                  NonEmptyOctetString OPTIONAL,
            other-ncnames                   SEQUENCE (SIZE(1..one-meg)) OF
                  NonEmptyOctetString OPTIONAL,
            other-uris                      SEQUENCE (SIZE(1..one-meg)) OF
                  NonEmptyOctetString OPTIONAL,
            attribute-values                SEQUENCE (SIZE(1..one-meg)) OF
                  EncodedCharacterString OPTIONAL,
            content-character-chunks        SEQUENCE (SIZE(1..one-meg)) OF
                  EncodedCharacterString OPTIONAL,
            other-strings                   SEQUENCE (SIZE(1..one-meg)) OF
                  EncodedCharacterString OPTIONAL,
            element-name-surrogates         SEQUENCE (SIZE(1..one-meg)) OF
                  NameSurrogate OPTIONAL,
            attribute-name-surrogates       SEQUENCE (SIZE(1..one-meg)) OF
                  NameSurrogate OPTIONAL }
                  (CONSTRAINED BY {
                        -- If the initial-vocabulary component is present, at least
                        -- one of its components shall be present -- }) OPTIONAL,
      notations                     SEQUENCE (SIZE(1..MAX)) OF
                                          Notation OPTIONAL,
      unparsed-entities             SEQUENCE (SIZE(1..MAX)) OF
                                          UnparsedEntity OPTIONAL,
```

```
          character-encoding-scheme  NonEmptyOctetString OPTIONAL,
          standalone                 BOOLEAN OPTIONAL,
          version                    NonIdentifyingStringOrIndex OPTIONAL
                                       -- OTHER STRING category --,
          children                   SEQUENCE (SIZE(0..MAX)) OF
                CHOICE {
                      element                     Element,
                      processing-instruction      ProcessingInstruction,
                      comment                     Comment,
                      document-type-declaration   DocumentTypeDeclaration }}
```

where the value **one-meg** is:

```
one-meg INTEGER ::= 1048576 -- Two to the power 20
```

The **NonEmptyOctetString** type is:

```
NonEmptyOctetString ::= OCTET STRING (SIZE(1..four-gig))
```

where the value **four-gig** is:

```
four-gig INTEGER ::= 4294967296 -- Two to the power 32
```

The **URI** type is:

```
URI ::= NonEmptyOctetString
```

7.2.2 The **EncodedCharacterString**, **NameSurrogate**, **Notation**, **UnparsedEntity**, **NonIdentifyingStringOrIndex**, **Element**, **ProcessingInstruction**, **Comment**, and **DocumentTypeDeclaration** types are defined in 7.17, 7.15, 7.11, 7.10, 7.14, 7.3, 7.5, 7.8, and 7.9 respectively.

7.2.3 The **URI** type shall be a URI as specified in IETF RFC 2396.

7.2.4 The component **restricted-alphabets** of **initial-vocabulary** (if present) shall carry one or more character strings, each holding the characters of a restricted alphabet. Each character string shall contain at least two characters, and all characters in the character string shall be distinct.

NOTE – The use of a restricted alphabet to optimize encodings of character strings is specified in 7.17.6.

7.2.5 The component **encoding-algorithms** of **initial-vocabulary** (if present) shall carry one or more URIs each identifying an encoding algorithm.

NOTE – There are built-in encoding algorithms defined in this Recommendation | International Standard (see clause 10), with specified vocabulary table indexes, but it is out of the scope of this Recommendation | International Standard to define further encoding algorithms and their associated URIs, nor is the means of defining such algorithms determined here. The information needed to define an encoding algorithm is specified in 8.3.3.

7.2.6 The **Document** type represents the **document** information item of an XML infoset. Since all other information items in an XML infoset are either properties of this information item or properties of an item that is a child or descendant of this item (at any depth), each **Document** represents a complete XML infoset.

NOTE – Each **Document** without a reference to an external vocabulary (see 7.2.13) also defines a final vocabulary that can be used as the external vocabulary of some other fast infoset document.

7.2.7 The **additional-data** component (if present) shall carry one or more **additional-datum** components to permit additional mechanisms for the processing of a fast infoset document.

NOTE 1 – An example would be data that enables a processor to access parts of a fast infoset document without requiring the processing of the whole document. The form of such data is not standardized.

NOTE 2 – The number of **additional-datum** components is restricted to 2^{20} components (see 7.2.1).

7.2.8 Each **additional-datum** shall consist of:

a) the **id** component (a value of the **URI** type); the URI shall reference a specification that defines the form and semantics of the **data** component; and

NOTE – The form of the **additional-datum** may be specified as an abstract type in conjunction with an encoding rule, or by any other suitable means.

b) the **data** component, which is an octet string that holds the additional processing data.

7.2.9 The use of an **additional-data** component is subject to the following:

a) an **additional-datum** component can be ignored by a processor unless the URI is recognized and the additional processing is considered relevant for the activity of that processor;

b) a processor that ignores all **additional-datum** components is nonetheless capable of generating an XML infoset that is equivalent to the XML infoset used to generate the fast infoset document.