

Uniform XML-Based Container Element - 1.0

Abstract

This document sets forth a generic format for XML-based container elements so that they can be transmitted, stored and included in XML documents or in other container elements in a uniform way. It also describes the conditions applicable to XML-based markup languages which use them. XML container elements represent data related to a particular topic.

Status

This document is an official technical specification of the Frogans technology.

This technical specification was adopted by the OP3FT on June 28, 2022.

Comments on this document are welcome and may be made on the Frogans technology mailing lists, accessible at the following permanent URL: <https://lists.frogans.org/>.

Location

This document is accessible at the following permanent URL: <https://www.frogans.org/en/resources/uxce/access.html>.

Copyright Statement

This document must be used in compliance with the Frogans Technology User Policy, accessible at the following permanent URL: <https://www.frogans.org/en/resources/ftup/access.html>.

Copyright (C) 2022 OP3FT. All rights reserved.

Trademark Notice

In order to enable all users worldwide to use the Frogans technology in a clearly defined, secure, and perpetual environment, the OP3FT Bylaws provide for the implementation of an intellectual property policy. In this context, the OP3FT is the holder of the "Frogans" trademark and other trademarks that are registered in France, the United States, and other countries around the world.

The right to use these trademarks is granted in the OP3FT Trademark Usage Policy, accessible at the following permanent URL:
<https://www.frogans.org/en/resources/otup/access.html>.

Table of Contents

1. Introduction	2
1.1. Background	2
1.2. Purpose	3
1.3. Intended Audience	4
1.4. Compliance	4
2. Generic Format	4
3. Validation	5
4. Usage	6
5. References	7
5.1. Normative References	7
5.2. Informative References	7

1. Introduction

1.1. Background

Started in 1999, the Frogans project aims to introduce a new medium for publishing content and services on the Internet, called Frogans.

From a technical standpoint, this new medium is designed as a new generic software layer running on top of the original Internet infrastructure, i.e. the TCP and IP protocols and the Domain Name System (DNS), alongside other existing generic software layers such as E-mail or the World Wide Web.

Frogans as a medium is intended for publishing Frogans sites. A Frogans site is made up of free-form pages called Frogans slides which are interconnected.

The technology making up the new medium, i.e. the Frogans technology, involves the use of markup languages in various areas, such as:

- the process for updating Frogans Player, the software used to browse Frogans sites
- the resolution of the addresses of Frogans sites
- the format for publishing Frogans sites

These markup languages are based on the Extensible Markup Language 1.0 (XML) defined by the W3C [XML].

XML was chosen because it presents several advantages: XML offers a foundation for designing markup languages with strict parsing rules; XML supports internationalization as it is based on the Unicode Standard [Unicode]; and XML documents can be easily authored, stored, processed, and transmitted.

After years of development, the Frogans technology is reaching the point where it can be made available to the general public, with a new series of technical specifications to be published by the OP3FT.

Some of these new technical specifications will introduce specific XML-based container elements to be used across different XML-based markup languages, each container element representing data related to a particular topic.

In this context, there is a need for a uniform way to transmit, store and include these XML-based container elements in XML documents or in other container elements, notably as concerns their character set and encoding.

1.2. Purpose

The purpose of this document is to set forth a generic format for XML-based container elements.

This document also describes the conditions applicable to XML-based markup languages which use these container elements.

Given that a generic format for XML-based container elements could be useful in contexts outside of the Frogans technology, the format for these container elements must not depend on components specific to the Frogans technology. As a result, this document must not contain any normative references to the technical specifications of such components.

Furthermore, this generic format must not limit the size of container elements nor constrain either the name of their elements and attributes or their internal structure. Such rules are the responsibility of the technical specifications describing specific container elements.

1.3. Intended Audience

This document is intended for those who design either XML-based container elements or XML-based markup languages which use these container elements.

1.4. Compliance

The rules applicable to XML-based container elements in this specification are defined in succession. The definition of each rule assumes compliance with all preceding rules.

A conforming implementation of this specification is an implementation which is compliant with all descriptions appearing in this document.

Hence, unlike in Request for Comments drawn up by the Internet Engineering Task Force (IETF), requirement levels in this specification are not indicated using key words such as "MUST", "MUST NOT", "SHOULD", and "SHOULD NOT" defined in RFC 2119 [RFC2119] and RFC 8174 [RFC8174]. This applies to all specifications drawn up by the OP3FT.

In this document, normative and informative references detailed in the References section appear between square brackets [].

2. Generic Format

An XML-based container element is a case-sensitive string of characters in the Unicode character set [Unicode] which represents an XML element as defined in section 3.1 of the Extensible Markup Language (XML) 1.0 [XML].

As a result:

- * Each character of the string is either in one of the three following ranges: from U+0020 to U+D7FF (inclusive), from U+E000 to U+FFFD (inclusive), or from U+10000 to U+10FFFF (inclusive); or is one of the following characters: the U+0009 CHARACTER TABULATION character, the U+000A LINE FEED (LF) character, or the U+000D CARRIAGE RETURN (CR) character.
- * The string may contain character references as well as the following entity references: & < > " and '
- * An XML-based container element may have attributes. It may contain other XML elements, texts, comments, processing

instructions (PIs) and CDATA sections as defined in [XML], or may be empty.

Although the XML technical specification [XML] is based on version 5.0.0 of the Unicode Standard, an XML-based container element can contain characters that are unassigned in that version of the Unicode Standard.

When XML-based container elements are transmitted, stored or included in XML documents, they are encoded using UTF-8 [RFC3629] without a byte-order mark (BOM).

3. Validation

The validation of an XML-based container element requires carrying out the following three steps in succession:

Step 1: Normalization

The XML technical specification defines whitespace as a sequence of one or more of the following characters: U+0020 SPACE, U+0009 CHARACTER TABULATION, U+000D CARRIAGE RETURN (CR), or U+000A LINE FEED (LF).

The following white space normalization processes are performed:

- * Ends of lines are normalized as follows:
 - Each U+000D CARRIAGE RETURN (CR) character that is followed by a U+000A LINE FEED (LF) character is removed.
 - Each U+000D CARRIAGE RETURN (CR) character that is not followed by a U+000A LINE FEED (LF) character is replaced by a U+000A LINE FEED (LF) character.
- * Attribute values are normalized as follows:
 - Each of the following characters is replaced by a U+0020 SPACE character: U+0009 CHARACTER TABULATION, U+000D CARRIAGE RETURN (CR), and U+000A LINE FEED (LF).
 - If the type of attribute is not a string, then any leading and trailing U+0020 SPACE characters are removed from the attribute value, and any sequence of more than one U+0020 SPACE character is replaced by a single U+0020 SPACE character.

Characters removed or replaced during these white space normalization processes may be present nevertheless in attribute values or text characters found in elements of the XML-based container element, given that these characters can be introduced via character references.

No additional normalization is performed on the XML-based container element. In particular, the XML-based container element is not normalized using Unicode normalization forms [UAX15].

Step 2: Compliance with generic format

The XML-based container element is checked against the rules defined in Section 2.

Step 3: Compliance with specific format

The XML-based container element is checked against the rules defined in the technical specification describing the specific container element.

The technical specification describing the specific container element may include rules concerning additional normalization processes.

4. Usage

When an XML-based markup language uses an XML-based container element:

- * The character set of the XML documents defined by that markup language is the Unicode character set [Unicode], i.e. that of the XML-based container element.
- * The encoding of these XML documents is UTF-8 [RFC3629], i.e. that of the XML-based container element.

Alternatively, if the character set of the XML documents defined by the markup language is not the Unicode character set, or if their encoding is not UTF-8, then XML-based container elements are included in the XML documents as follows:

1. The string representing the XML-based container element is encoded using UTF-8 [RFC3629] without a byte-order mark (BOM), resulting in a sequence of octets.

2. The sequence of octets is encoded via an encoding method which results in text characters belonging to the character set of the XML document. For example, if the character set of the XML documents includes the ASCII character set [ASCII], the Base64 encoding method, as defined in section 4 of RFC 4648 [RFC4648], may be used. The choice of the encoding method is left up to the technical specification describing the markup language.
3. The text characters are included in the XML documents as text, as the value of an attribute, or as the content of a CDATA section.

An XML-based container element can include other XML-based container elements.

5. References

5.1. Normative References

- [RFC3629] Yergeau, F., "UTF-8, a transformation format of ISO 10646", STD 63, RFC 3629, DOI 10.17487/RFC3629, November 2003, <<https://www.rfc-editor.org/info/rfc3629>>.
- [Unicode] The Unicode Consortium, "The Unicode Standard", Version 5.0.0, (Boston, MA, Addison-Wesley, 2007. ISBN 0-321-48091-0), <<https://www.unicode.org/versions/Unicode5.0.0/>>.
- [XML] Bray, T., Paoli, J., Sperberg-McQueen, C., Maler, E., and F. Yergeau, "Extensible Markup Language (XML) 1.0 (Fifth Edition)", World Wide Web Consortium Recommendation REC-xml-20081126, November 2008, <<https://www.w3.org/TR/2008/REC-xml-20081126/>>.

5.2. Informative References

- [ASCII] American National Standards Institute (formerly United States of America Standards Institute), "USA Code for Information Interchange", ANSI X3.4-1968, 1968.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<https://www.rfc-editor.org/info/rfc2119>>.
- [RFC4648] Josefsson, S., "The Base16, Base32, and Base64 Data Encodings", RFC 4648, DOI 10.17487/RFC4648, October 2006, <<https://www.rfc-editor.org/info/rfc4648>>.

- [RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <<https://www.rfc-editor.org/info/rfc8174>>.
- [UAX15] The Unicode Consortium and K. Whistler, "Unicode Standard Annex #15: Unicode Normalization Forms", an integral part of The Unicode Standard, Version 14.0.0, Revision 51, August 2021, <<https://www.unicode.org/reports/tr15/tr15-51.html>>.

