Network Working Group
Request for Comments: 1874
Category: Experimental

E. Levinson
Accurate Information Systems, Inc.

December 1995

SGML Media Types

Status of this Memo

This memo defines an Experimental Protocol for the Internet community. This memo does not specify an Internet standard of any kind. Discussion and suggestions for improvement are requested. Distribution of this memo is unlimited.

Abstract

This document proposes new media sub-types of Text/SGML and Application/SGML. These media types can be used in the exchange of SGML documents and their entities. Specific details for the exchange or encapsulation of groups of related SGML entities using MIME are currently being considered by the mimesgml Working Group <sgml-internet@ebt.com>.

1. Introduction

A need exists for the transfer the elements of documents constructed using the Standard Generalized Markup Language (SGML) [ISO-8879]. While the specific details of such transfers are being considered general agreement exists on the need to register basic media types for the SGML entities not covered by existing types.

The Standard Generalized Markup Language (SGML) is used to encode document structure and a rigorous description of it is left to [ISO-8879]. The terms used in the present document attempt to be consistent with SGML terminology and usage.

2. The SGML Media-Types

There are two media-types for SGML parsable entities, Text/SGML and Application/SGML. Both have the same optional parameters. Text/SGML provides a fallback to Text/Plain for those without SGML capability. Senders should base the choice between text and application media-types on the entity's content. Text is suggested for entities that would be meaningful to a human being without SGML processing. Application/SGML is recommended for all others.

2.1. Text/SGML

MIME type name:	Text
MIME subtype name:	SGML
Required parameters:	none
Optional parameters:	charset, SGML-bctf, SGML-boot
Encoding considerations:	may be encoded
Security considerations:	see section 4 below
Published specification:	ISO 8879:1986
Person and email address	to contact for further information: E. Levinson <ELevinson@Accurate.com>

The Text/SGML media-type can be employed when the contents of the SGML entity is intended to be read by a human and is in a readily comprehensible form. That is the content can be easily discerned by someone without SGML display software. Each record in the SGML entity, delimited by record start (RS) and record end (RE) codes, must correspond to a line in the Text/SGML body part.

SGML entities that do not meet the above requirements should use the Application/SGML media-type.

See section 2.3 for a description of the parameters.

2.2. Application/SGML

MIME type name:	Application
MIME subtype name:	SGML
Required parameters:	none
Optional parameters:	SGML-bctf, SGML-boot
Encoding considerations:	may be encoded
Security considerations:	see section 4 below
Published specification:	ISO-8879
Person and email address	to contact for further information: E. Levinson <ELevinson@Accurate.com>

Use the Application/SGML media-type for SGML text entities that are not appropriate for Text/SGML. When used, each record start (RS) and record end (RE) character shall be explicitly represented by the bit combination specified in the SGML declaration.

The parameters are described in the next section.

2.3. SGML Sub-type Parameters

The parameters for the Text/ and Application/SGML subtypes are defined below.

charset — The charset parameter for Text/SGML is defined in [RFC-1521], the valid values and their meaning are registered by the Internet Assigned Numbers Authority (IANA) [RFC-1590]. The default charset value for all Text content-types is "us-ascii" [RFC-1521].

The charset parameter is provided to permit non-SGML capable systems to provide reasonable behavior when Text/SGML defaults to Text/Plain. SGML capable systems will use the SGML-bctf parameter.

SGML-bctf — The SGML-bctf (SGML bit combination transformation format) parameter describes the method used to transform the entity's sequence of constant width binary numbers (called "bit combinations" in [ISO 8879, 4.24]) into the octet stream contained in the MIME body part.

Valid values for SGML-bctf are the BCTF notation names defined in Annex C of [ISO-10744] and are reproduced for convenience in the Appendix. The default value is "identity", i.e. perform no transformation.

SGML-boot — The SGML-boot parameter value is the content-ID of a MIME body part (Application/Octet-stream) that satisfies the requirements of the boot attribute in [ISO-10744]. The Appendix contains a summary of those requirements. The SGML-boot parameter is only applicable if the SGML entity is a document entity.

3. Security Considerations

SGML entities contain information to be parsed and processed by the recipient's SGML system. Those entities may contain and such systems may permit explicit system level commands to be execute while processing the data. To the extent that an SGML system will execute arbitrary command strings recipients of SGML entities may be at risk.

Parsable SGML entities may also contain explicit processing instructions for a presentation or composition system; use of such instructions present concerns similar to those of Application/PostScript.

4. References

[ISO-8879]
Information processing -- 8-bit Single-Byte Coded Graphic Character Sets -- Part 1: Latin Alphabet No. 1, ISO 8859-1:1987.

[ISO-8879]
ISO 8879:1986, Information processing -- Text and office systems -- Standard Generalized Markup Language (SGML).

[ISO-10744]
ISO/IEC 10744:1992, Information technology -- Hypermedia/Time-based Structuring Language (HyTime) (as modified by First Proposed Technical Corrigendum, ISO/IEC JTC1/SC18 N5027)

[RFC-1521]
Borenstein, N., and N. Freed, "MIME (Multipurpose Internet Mail Extensions) Part One: Mechanisms for Specifying and Describing the Format of Internet Message Bodies", RFC 1521, Bellcore, Innosoft, September 1993.

[RFC-1590]
Postel, J., "Media Type Registration Procedure", RFC 1590, USC/Information Sciences Institute, March 1994.

[RFC-1642]
Goldsmith, D., and M. Davis, "UTF-7, A Mail-Safe Transformation Format of UNICODE", RFC 1642, Taligent, Inc., July 1994.

5. Author's Address

Ed Levinson
Accurate Information Systems, Inc.
2 Industrial Way
Eatontown, NJ 07724

EMail: ELevinson@Accurate.com

APPENDIX

ISO-10744 BCTF Values and Boot Attribute

A.1. Bit Combination Transformation Format (BCTF) Values

The following list of Bit Combination Transformation Format (BCTF) values is provided as a convenience. The authoritative source is [ISO-10744].

identity — Each bit combination is represented by a single octet; this BCTF can be used only for entities all of whose bit combinations have a value not exceeding 255.

fixed-2 — Each bit combination is represented by exactly 2 octets, with the more significant octet first; this BCTF can be used only for entities all of whose bit combinations have a value not exceeding 65535.

fixed-3 — Each bit combination is represented by exactly 3 octets, with a more significant octet preceding any less significant octets; this BCTF can be used only for entities all of whose bit combinations have a value not exceeding 16777215.

fixed-4 — Each bit combination is represented by exactly 4 octets, with a more significant octet preceding any less significant octets.

utf-8 — Each bit combination is represented by a variable number of octets according to UCS Transformation Format 8 defined in Annex P to be added by the first proposed drafted amendment (PDAM 1) to ISO/IEC 10646-1:1993.

utf-7 — Each bit combination is represented by a variable number of octets in the range 0 through 127 as described in [RFC-1642]; this BCTF can be used only for entities all of whose bit combinations have a value not exceeding 65535.

euc-jp — Each bit combination is treated as a pair of octets, most significant octet first, encoding a character using the Extended_UNIX_Code_Fixed_Width_for_Japanese charset, and is transformed into the variable length sequence of octets that would encode that character using the

Extended_UNIX_Code_Packed_Format_for_Japanese char-
set.

sjis Each bit combination is treated as a pair of octets, most significant octet first, encoding a character using the Extended_UNIX_Code_Fixed_Width_for_Japanese charset, and is transformed into the variable length sequence of octets that would encode that character using the Shift_JIS charset.

A.2. The Boot Attribute

The body part specified by the SGML-boot parameter contains a sequence of triplets of positive integers separated by white space. The triplets correspond to the described character set portion [IS0-8879, 13.1.1.2] of the SGML declaration. SGML-boot provides the capability to identify the character set of the document's SGML declaration when it uses significant SGML characters [ibid., 4.298] in the SGML reference concrete syntax [ibid., 13.4] that have a character number [ibid., 4.44] in the document's character set that differs from us-ascii. The default value is "0 128 0", all characters are us-ascii.

Notes: (1) The triplet, <dscn noc bscn> has the following meaning. Starting with character number dscn in the us-ascii character set, renumber noc characters starting at bscn and incrementing by one. Thus, 0 128 0, represents the identity mapping. (2) The document's declaration itself may also redefine the significant SGML characters; the boot attribute is intended to bootstrap the SGML system's parse of the declaration.