



Comparison of SGML and XML

World Wide Web Consortium Note 15-December-1997

This version:

<http://www.w3.org/TR/NOTE-sgml-xml-971215>

Author:

James Clark <jjc@jclark.com>

Status of this document

This document is a NOTE made available by the W3 Consortium for discussion only. This indicates no endorsement of its content, nor that the Consortium has, is, or will be allocating any resources to the issues addressed by the NOTE. Errors or omissions in this document should be reported to the [author](#).

Abstract

This document provides a detailed comparison of SGML (ISO 8879) and XML.

Comparison of SGML and XML

Version 1.0

Table of Contents

- [1. Differences Between XML and SGML](#)
- [2. Transforming SGML to XML](#)
- [3. SGML Declaration for XML](#)

1. Differences Between XML and SGML

XML allows only documents that use the SGML declaration in this note. This declares all the following SGML features as NO:

- DATATAG
- OMITTAG
- RANK
- LINK (SIMPLE, IMPLICIT and EXPLICIT)
- CONCUR
- SUBDOC
- FORMAL

Note that it differs from the reference concrete syntax in a number of ways:

- It also declares no short reference delimiters; it follows that `SHORTREF` and `USEMAP` declarations cannot occur in XML
- The `PIC` (processing instruction close) delimiter is `?>`
- Quantities and capacities are effectively unlimited

- Names are case sensitive (NAMECASE GENERAL is NO)
- Underscore and colon are allowed in names
- Names can use Unicode characters and are not restricted to ASCII

The following constructs which are permitted in SGML when SHORTTAG is YES are not allowed in XML:

- Unclosed start-tags
- Unclosed end-tags
- Empty start-tags
- Empty end-tags
- Attribute values in attribute specifications entered directly rather than as literals
- Attribute specifications that omit the attribute name

NET delimiters can be used only to close an empty element. In SGML without the Web SGML Adaptations Annex, the NET delimiter is declared as />. With this approach, XML is not allowing null end-tags and is allowing net-enabling start-tags only for elements with no end-tag. In SGML with the Web SGML Adaptations Annex, there is a separate NESTC (net-enabling start tag close) delimiter. This allows the XML <e/> syntax to be handled as a combination of a net-enabling start-tag <e/ and a null end-tag >. With this approach, XML is allowing a net-enabling start-tag only when immediately followed by a null end-tag.

XML imposes the following restrictions not in SGML:

- Entity references
 - Entity references must be closed with a REFC delimiter
 - References to external data entities in content are not allowed
 - General entity references in content are required to be synchronous
 - External entity references in attribute values are not allowed
 - Parameter entity references are allowed in the internal subset only within a declaration separator (that is, at a point where a markup declaration could occur)
- Character references
 - Character references must be closed with a REFC delimiter
 - Named character references are not allowed
 - Numeric character references to non-SGML characters are not allowed
- Entity declarations
 - A #DEFAULT entity cannot be declared
 - External SDATA entities are not allowed
 - External CDATA entities are not allowed
 - Internal SDATA entities are not allowed
 - Internal CDATA entities are not allowed
 - PI entities are not allowed
 - Bracketed text entities are not allowed
 - External identifiers must include a system identifier
 - Attributes cannot be specified for an entity
 - The replacement text of general text entities and external parameter entities is required to be well-formed
 - An ampersand in a parameter literal must be followed by a syntactically valid entity reference or numeric character reference
- Attribute definition list declarations
 - Associated element type in attribute definition list declarations cannot be a name group
 - Attributes cannot be declared for a notation
 - CURRENT attributes are not allowed
 - Content reference attributes are not allowed
 - NUTOKEN(S) declared values are not allowed
 - NUMBER(S) declared values are not allowed
 - NAME(S) declared values are not allowed
 - A name token group must use the or connector
 - Attribute values specified as defaults in attribute definition list declarations must be literals (SGML allows them not to be even when SHORTTAG is NO)
- Element type declarations
 - Associated element type in element type declaration cannot be a name group
 - In an element declaration, a generic identifier cannot be specified as a rank stem and rank suffix (SGML allows this even when the RANK feature is NO)
 - Minimization parameters in element declarations are not allowed
 - RCDATA declared content are not allowed
 - CDATA declared content are not allowed
 - Content models cannot use the and connector
 - Content models for mixed content have a restricted form
 - Inclusions are not allowed
 - Exclusions are not allowed
- Comments
 - A parameter separator cannot contain comments; this means that markup declarations (other than comment declarations) cannot contain comments
 - Empty comment declarations (<!> in the reference concrete syntax) are not allowed

- A comment declaration cannot contain more than one comment
- In a comment declaration, an S separator is not allowed before the final MDC
- Processing instructions
 - Processing instructions must start with a name (the PI target)
 - A processing instruction whose PI target is `xml` can only occur at the beginning of an external entity and must be an XML declaration if it occurs in the document entity, and otherwise a text declaration
 - A PI target must not match `[Xx]` `[Mm]` `[Ll]` unless it is `xml`
- Marked sections
 - In marked section declarations, `TEMP` status keyword is not allowed
 - `RCDATA` marked sections are not allowed
 - `INCLUDE/IGNORE` marked sections are not allowed in the document instance
 - In a marked section declaration, a status keyword specification that contains no status keywords is not allowed
 - In a marked section declaration, a status keyword specification cannot contain more than one status keyword
 - Marked sections are not allowed in the internal subset
 - Parameter separators are not allowed in status keyword specifications in the document instance; in particular, parameter entity references are not allowed
- Other
 - Names beginning with `[Xx]` `[Mm]` `[Ll]` are reserved
 - The SGML declaration must be implied and cannot be explicitly present in the document entity
 - When `<` and `&` occur as data, they must be entered as `<` and `&`
 - A parameter separator required by the formal syntax must always be present and cannot be omitted when it is adjacent to a delimiter

XML predefines the semantics of the attributes `xml:space` and `xml:lang`. It also reserves all attribute, element type and notation names beginning with `[Xx]` `[Mm]` `[Ll]`.

XML requires that an SGML parser use an entity manager that behaves as follows:

- Lines are terminated by newline (Unicode code #X000A) rather than being delimited by RS and RE as with a typical SGML entity manager
- System identifiers are treated as URLs
- The entity manager must support entities encoded in UTF-16 and UTF-8, and must be able automatically to detect which encoding an entity uses based on the presence of the byte order mark
- The entity manager should be able to recognize the encoding declaration in the XML declaration and encoding PI and use it to determine the encoding of entity

XML imposes requirements on the information that a parser must make available to an application.

XML depends on the following changes to SGML made by Web SGML Adaptations Annex:

- `HCRO` delimiter (for hex numeric character references); for XML this is `&#x`
- `EMPTYNRM` feature that allows elements declared `EMPTY` to have end-tags
- `NESTC` delimiter
- Duplicate enumerated attribute tokens are allowed
- Relaxation of rules on use of parameter entity references inside groups
- Multiple `ATTLIST` declarations for a single element type
- `ATTLIST` declarations which don't declare any attributes
- `KEEPRSRE` feature that turns off SGML's rules for ignoring RSs and REs
- Fully-tagged SGML documents; a document that is fully-tagged but not type-valid is a conforming SGML document; this makes all XML documents, including those that are well-formed but not valid, conforming SGML documents
- Predefined data character entities in the SGML declaration (for `lt`, `amp` and so on)
- Unlimited capacities and quantities

The Web SGML Adaptations Annex also enables some XML restrictions to be enforced in SGML:

- `SHORTTAG` is unbundled, so the SGML declaration can allow attribute defaulting and `NET` without allowing other `SHORTTAG` constructs
- The SGML declaration can assert that a document is integrally stored, which disallows improperly nested entity references in content

2. Transforming SGML to XML

For most restrictions in XML that go beyond SGML, it is possible to transform an SGML document automatically into a document that meets the restrictions, and is equivalent in the sense that it has the same ESIS. There are a number of restrictions for which this is not the case:

External `SDATA` entities, external `CDATA` entities

These could be transformed into `NDATA` entities.

Subdocument entities

These could be converted into `NDATA` entities with a notation that indicates that they are SGML or XML.

References to external data entities in content

These could be transformed into an empty element with an attribute whose declared value is `ENTITY`.

Data attributes

Since an external data entity can only be used in an ENTITY or ENTITIES attribute on an element, these could be transformed into other attributes on the element.

Internal SDATA entities

References could be transformed into numeric character references to the appropriate Unicode character; if used in an entity or entities attribute, the entity will have to be made external.

Internal CDATA entities

If used in an ENTITY or ENTITIES attribute, the entity will have to be made external (references to CDATA entities are not part of ESIS).

PI entities

If they contain ?>, they cannot be converted into an XML PI. It could be an application convention that entity references are replaced in PIs. Also if they do not start with a name, they cannot be converted into a well-formed XML PI.

names

An SGML document can have a concrete syntax which allows characters in names that XML does not allow in names.

3. SGML Declaration for XML

The following SGML declaration takes advantage of the Extended Naming Rules Technical Corrigendum to ISO 8879, but does not make use of the Web SGML Adaptations Annex:

```

<!SGML -- SGML Declaration for XML --
  "ISO 8879:1986 (ENR) "

  CHARSET
    BASESET
      "ISO Registration Number 176//CHARSET
      ISO/IEC 10646-1:1993 UCS-4 with implementation
      level 3//ESC 2/5 2/15 4/6"
    DESCSET
      0      9      UNUSED
      9      2      9
      11     2      UNUSED
      13     1      13
      14     18     UNUSED
      32     95     32
      127    1      UNUSED
      128    32     UNUSED
      160    55136  160
      55296  2048   UNUSED -- surrogates --
      57344  8190   57344
      65534  2      UNUSED -- FFFE and FFFF --
      65536  1048576 65536

  CAPACITY SGMLREF
    -- Capacities are not restricted in XML --
    TOTALCAP 99999999
    ENTCAP   99999999
    ENTCHCAP 99999999
    ELEMCAP  99999999
    GRPCAP   99999999
    EXGRPCAP 99999999
    EXNMCAP  99999999
    ATTCAP   99999999
    ATTCHCAP 99999999
    AVGRPCAP 99999999
    NOTCAP   99999999
    NOTCHCAP 99999999
    IDCAP    99999999
    IDREFCAP 99999999

```

MAPCAP 99999999
 LKSETCAP 99999999
 LKNMCAP 99999999

SCOPE DOCUMENT

SYNTAX

SHUNCHAR NONE
 BASESET "ISO Registration Number 176//CHARSET
 ISO/IEC 10646-1:1993 UCS-4 with implementation
 level 3//ESC 2/5 2/15 4/6"

DESCSET
 0 1114112 0

FUNCTION
 RE 13
 RS 10
 SPACE 32
 TAB SEPCHAR 9

NAMING

LCNMSTRT ""
 UCNMSTRT ""
 NAMESTRT
 58 95 192-214 216-246 248-305 308-318 321-328
 330-382 384-451 461-496 500-501 506-535 592-680
 699-705 902 904-906 908 910-929 931-974 976-982
 986 988 990 992 994-1011 1025-1036 1038-1103
 1105-1116 1118-1153 1168-1220 1223-1224
 1227-1228 1232-1259 1262-1269 1272-1273
 1329-1366 1369 1377-1414 1488-1514 1520-1522
 1569-1594 1601-1610 1649-1719 1722-1726
 1728-1742 1744-1747 1749 1765-1766 2309-2361
 2365 2392-2401 2437-2444 2447-2448 2451-2472
 2474-2480 2482 2486-2489 2524-2525 2527-2529
 2544-2545 2565-2570 2575-2576 2579-2600
 2602-2608 2610-2611 2613-2614 2616-2617
 2649-2652 2654 2674-2676 2693-2699 2701
 2703-2705 2707-2728 2730-2736 2738-2739
 2741-2745 2749 2784 2821-2828 2831-2832
 2835-2856 2858-2864 2866-2867 2870-2873 2877
 2908-2909 2911-2913 2949-2954 2958-2960
 2962-2965 2969-2970 2972 2974-2975 2979-2980
 2984-2986 2990-2997 2999-3001 3077-3084
 3086-3088 3090-3112 3114-3123 3125-3129
 3168-3169 3205-3212 3214-3216 3218-3240
 3242-3251 3253-3257 3294 3296-3297 3333-3340
 3342-3344 3346-3368 3370-3385 3424-3425
 3585-3630 3632 3634-3635 3648-3653 3713-3714
 3716 3719-3720 3722 3725 3732-3735 3737-3743
 3745-3747 3749 3751 3754-3755 3757-3758 3760
 3762-3763 3773 3776-3780 3904-3911 3913-3945
 4256-4293 4304-4342 4352 4354-4355 4357-4359
 4361 4363-4364 4366-4370 4412 4414 4416 4428

4430 4432 4436-4437 4441 4447-4449 4451 4453
 4455 4457 4461-4462 4466-4467 4469 4510 4520
 4523 4526-4527 4535-4536 4538 4540-4546 4587
 4592 4601 7680-7835 7840-7929 7936-7957
 7960-7965 7968-8005 8008-8013 8016-8023 8025
 8027 8029 8031-8061 8064-8116 8118-8124 8126
 8130-8132 8134-8140 8144-8147 8150-8155
 8160-8172 8178-8180 8182-8188 8486 8490-8491
 8494 8576-8578 12295 12321-12329 12353-12436
 12449-12538 12549-12588 19968-40869 44032-55203

LCNMCHAR ""

UCNMCHAR ""

NAMECHAR

45-46 183 720-721 768-837 864-865 903 1155-1158
 1425-1441 1443-1465 1467-1469 1471 1473-1474
 1476 1600 1611-1618 1632-1641 1648 1750-1764
 1767-1768 1770-1773 1776-1785 2305-2307 2364
 2366-2381 2385-2388 2402-2403 2406-2415
 2433-2435 2492 2494-2500 2503-2504 2507-2509
 2519 2530-2531 2534-2543 2562 2620 2622-2626
 2631-2632 2635-2637 2662-2673 2689-2691 2748
 2750-2757 2759-2761 2763-2765 2790-2799
 2817-2819 2876 2878-2883 2887-2888 2891-2893
 2902-2903 2918-2927 2946-2947 3006-3010
 3014-3016 3018-3021 3031 3047-3055 3073-3075
 3134-3140 3142-3144 3146-3149 3157-3158
 3174-3183 3202-3203 3262-3268 3270-3272
 3274-3277 3285-3286 3302-3311 3330-3331
 3390-3395 3398-3400 3402-3405 3415 3430-3439
 3633 3636-3642 3654-3662 3664-3673 3761
 3764-3769 3771-3772 3782 3784-3789 3792-3801
 3864-3865 3872-3881 3893 3895 3897 3902-3903
 3953-3972 3974-3979 3984-3989 3991 3993-4013
 4017-4023 4025 8400-8412 8417 12293 12330-12335
 12337-12341 12441-12442 12445-12446 12540-12542

NAMECASE

GENERAL NO

ENTITY NO

DELIM

GENERAL SGMLREF

NET "/>"

PIC "?>"

SHORTREF NONE

NAMES

SGMLREF

QUANTITY SGMLREF

-- Quantities are not restricted in XML --

ATTCNT 99999999

```
ATTSPLEN      99999999
-- BSEQLEN    not used --
-- DTAGLEN    not used --
-- DTEMPLLEN  not used --
ENTLVL        99999999
GRPCNT        99999999
GRPGTCNT     99999999
GRPLVL        99999999
LITLEN        99999999
NAMELEN       99999999
-- no need to change NORMSEP --
PILEN         99999999
TAGLEN        99999999
TAGLVL        99999999
```

FEATURES

MINIMIZE

```
DATATAG NO
OMITTAG NO
RANK NO
SHORTTAG YES -- SHORTTAG is needed for NET --
```

LINK

```
SIMPLE NO
IMPLICIT NO
EXPLICIT NO
```

OTHER

```
CONCUR NO
SUBDOC NO
FORMAL NO
```

```
APPINFO NONE
```

>

The following SGML declaration takes advantage of the Web SGML Adaptations Annex to ISO 8879:

```

<!SGML -- SGML Declaration for XML --
  "ISO 8879:1986 (WWW) "

CHARSET
  BASESET
    "ISO Registration Number 176//CHARSET
    ISO/IEC 10646-1:1993 UCS-4 with implementation
    level 3//ESC 2/5 2/15 4/6"
  DESCSET
    0          9          UNUSED
    9          2          9
    11         2          UNUSED
    13         1          13
    14         18         UNUSED
    32         95         32
    127        1          UNUSED
    128        32         UNUSED
    160        55136     160
    55296      2048      UNUSED -- surrogates --
    57344      8190      57344
    65534      2          UNUSED -- FFFE and FFFF --
    65536      1048576  65536

CAPACITY NONE

SCOPE DOCUMENT

SYNTAX
  SHUNCHAR NONE
  BASESET "ISO Registration Number 176//CHARSET
  ISO/IEC 10646-1:1993 UCS-4 with implementation
  level 3//ESC 2/5 2/15 4/6"
  DESCSET
    0 1114112 0
  FUNCTION
    RE    13
    RS    10
    SPACE 32
    TAB   SEPCHAR 9

NAMING
  LCNMSTRT ""
  UCNMSTRT ""
  NAMESTRT
    58 95 192-214 216-246 248-305 308-318 321-328
    330-382 384-451 461-496 500-501 506-535 592-680
    699-705 902 904-906 908 910-929 931-974 976-982
    986 988 990 992 994-1011 1025-1036 1038-1103
    1105-1116 1118-1153 1168-1220 1223-1224
    1227-1228 1232-1259 1262-1269 1272-1273
    1329-1366 1369 1377-1414 1488-1514 1520-1522
    1569-1594 1601-1610 1649-1719 1722-1726
    1728-1742 1744-1747 1749 1765-1766 2309-2361
    2365 2392-2401 2437-2444 2447-2448 2451-2472

```


2474-2480 2482 2486-2489 2524-2525 2527-2529
2544-2545 2565-2570 2575-2576 2579-2600
2602-2608 2610-2611 2613-2614 2616-2617
2649-2652 2654 2674-2676 2693-2699 2701
2703-2705 2707-2728 2730-2736 2738-2739
2741-2745 2749 2784 2821-2828 2831-2832
2835-2856 2858-2864 2866-2867 2870-2873 2877
2908-2909 2911-2913 2949-2954 2958-2960
2962-2965 2969-2970 2972 2974-2975 2979-2980
2984-2986 2990-2997 2999-3001 3077-3084
3086-3088 3090-3112 3114-3123 3125-3129
3168-3169 3205-3212 3214-3216 3218-3240
3242-3251 3253-3257 3294 3296-3297 3333-3340
3342-3344 3346-3368 3370-3385 3424-3425
3585-3630 3632 3634-3635 3648-3653 3713-3714
3716 3719-3720 3722 3725 3732-3735 3737-3743
3745-3747 3749 3751 3754-3755 3757-3758 3760
3762-3763 3773 3776-3780 3904-3911 3913-3945
4256-4293 4304-4342 4352 4354-4355 4357-4359
4361 4363-4364 4366-4370 4412 4414 4416 4428
4430 4432 4436-4437 4441 4447-4449 4451 4453
4455 4457 4461-4462 4466-4467 4469 4510 4520
4523 4526-4527 4535-4536 4538 4540-4546 4587
4592 4601 7680-7835 7840-7929 7936-7957
7960-7965 7968-8005 8008-8013 8016-8023 8025
8027 8029 8031-8061 8064-8116 8118-8124 8126
8130-8132 8134-8140 8144-8147 8150-8155
8160-8172 8178-8180 8182-8188 8486 8490-8491
8494 8576-8578 12295 12321-12329 12353-12436
12449-12538 12549-12588 19968-40869 44032-55203

LCNMCHAR ""

UCNMCHAR ""

NAMECHAR

45-46 183 720-721 768-837 864-865 903 1155-1158
1425-1441 1443-1465 1467-1469 1471 1473-1474
1476 1600 1611-1618 1632-1641 1648 1750-1764
1767-1768 1770-1773 1776-1785 2305-2307 2364
2366-2381 2385-2388 2402-2403 2406-2415
2433-2435 2492 2494-2500 2503-2504 2507-2509
2519 2530-2531 2534-2543 2562 2620 2622-2626
2631-2632 2635-2637 2662-2673 2689-2691 2748
2750-2757 2759-2761 2763-2765 2790-2799
2817-2819 2876 2878-2883 2887-2888 2891-2893
2902-2903 2918-2927 2946-2947 3006-3010
3014-3016 3018-3021 3031 3047-3055 3073-3075
3134-3140 3142-3144 3146-3149 3157-3158
3174-3183 3202-3203 3262-3268 3270-3272
3274-3277 3285-3286 3302-3311 3330-3331
3390-3395 3398-3400 3402-3405 3415 3430-3439
3633 3636-3642 3654-3662 3664-3673 3761
3764-3769 3771-3772 3782 3784-3789 3792-3801
3864-3865 3872-3881 3893 3895 3897 3902-3903
3953-3972 3974-3979 3984-3989 3991 3993-4013

4017-4023 4025 8400-8412 8417 12293 12330-12335
12337-12341 12441-12442 12445-12446 12540-12542

NAMECASE

GENERAL NO
ENTITY NO

DELIM

GENERAL SGMLREF
HCRO "&#x" -- 38 is the number for ampersand --
NESTC "/"
NET ">"
PIC "?>"
SHORTREF NONE

NAMES

SGMLREF

QUANTITY NONE

ENTITIES

"amp" 38
"lt" 60
"gt" 62
"quot" 34
"apos" 39

FEATURES

MINIMIZE

DATATAG NO
OMITTAG NO
RANK NO
SHORTTAG
STARTTAG
EMPTY NO
UNCLOSED NO
NETENABL IMMEDNET

ENDTAG

EMPTY NO
UNCLOSED NO

ATTRIB

DEFAULT YES
OMITNAME NO
VALUE NO

EMPTYNRM YES

IMPLYDEF

ATTLIST YES
DOCTYPE YES
ELEMENT YES
ENTITY YES
NOTATION YES

LINK

SIMPLE NO
IMPLICIT NO

EXPLICIT NO
OTHER

CONCUR NO
SUBDOC NO
FORMAL NO
URN NO
KEEPRSRE YES
VALIDITY TAG
ENTITIES

REF ANY
INTEGRAL YES

APPINFO NONE

SEEALSO "ISO 8879//NOTATION Application Requirements for XML//EN"

>