# W3C Recommendations

## W3C Documents

Parsing the XML Schema primer and specs at:

[XML Schema Part 0: Primer](https://www.w3.org/TR/2004/REC-xmlschema-0-20041028/primer.html)

[XML Schema Part 1: Structures](https://www.w3.org/TR/2004/REC-xmlschema-1-20041028/structures.html)

[XML Schema Part 2: Datatypes](https://www.w3.org/TR/2004/REC-xmlschema-2-20041028/datatypes.html)

# Internal links in MS Word

An internal link, in a MS document, is implemented with a **bookmark**.

A bookmark is created on the target location, it has a *name*, with some constraints (no special characters, max 40 characters).

A hyperlink that references a bookmark can be created anywhere, it will need the *bookmark name*, and it also has its own *link text*, the part of the document that is *active* as a link, and that MS Word styles in a specific way (blue, underlined).

## Different kinds of internal references in W3C documents

We find many different kinds of internal links in the W3C documents. The links are established by matching the **ref** or **def** attribute on the reference element to the **id** attribute on the target.

We use the common value of the **id** and (**ref** or **def**) attributes as the *bookmark name*.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Reference element** | **Target identification** | **Target** | **Identifier** | **Link text in the reference** | **Tgt?** |
| <specref> | "**ref**" attribute | <div> element | "**id**" attribute | The <div> element has a <head> child element, whose text content is used as the heading text in the <div>; that same text is used as the ***link text***. | Yes |
|  |  | <constraintnote> | "**id**" attribute |  |  |
|  |  | <p> element | "**id**" attribute | The target paragraph's section number (e.g., §2.7). This is not in the XML text. | No  **(\*)** |
|  |  | <schemaComp> | "**id**" attribute | The <schemaComp> element has a <head> child element, whose text content is used as the table heading/title; that same text, concatenated with the section number, is used as the ***link text***. | Yes |
|  |  | <constraintnote> | "**id**" attribute | The <constraintnote> element has a <head> child element, whose text content is used as the table heading/title; that same text, concatenated with the section number, is used as the ***link text***. | Yes |
|  |  | <note> | "**id**" attribute | The target paragraph's section number (e.g., §3.4.6). This is not in the XML text. | No  **(\*)** |
| <bibref> | "**ref**" attribute | <bibl> | "**id**" attribute | **[***key\_value***]**, where *key\_value* is the value of the **key** attribute on the <bibl> element. | Yes |
| <termref> | "**def**" attribute | <termdef> | "**id**" attribute | The <termref> element's inner *text* is the ***link text***. | No |
| <eltref> | "**ref**" attribute | <reprelt> | "**eltname**" attribute | The value of the <eltref> element's **ref** attribute, in between angle brackets **< >**. | No |
| <clauseref> | "**ref**" attribute | <item> | "**id**" attribute | A clause number, which I don't have. It also needs to be generated on the <item> rendering. | Yes |
| <nt> | "**def**" attribute | <prod> | "**id**" attribute | ***Not implemented !!!*** |  |

**(\*)** Note that if we do want to use the section number as link text, we will need info from the target.

**TBD: enforcing bookmark name rules, who/where ?**

# Tags

Tags currently handled by the python code (i.e., that have a do\_XYZ() method).

Nodes like <abstract> or <blist> only exist to organize the structure, but they don't actually contribute text or formatting to the document (they have sub-elements, but no text nodes).

An element such as <loc> which we classify as inline can sometimes appear by itself, this means the enclosing element (<publoc> or <altlocs> for example) must create the required paragraph.

I should do a first pass to eliminate all <phrase> elements (keep add/chg, remove del). They describe document history, how we got to this version, but I just want to work with the latest revision.

<emph> should be (inline, simple, simple), and it's not just because of one <phrase>.

Inline has nothing to do with the element's XML structure, only with how it's going to be rendered in the output.

Maybe write the code to create this table from the XML tree, make sure the content model is right.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tag** | **Block / inline** | **Type of type** | **Content model** | **Description** |
| abstract | n/a | complex | complex | Structural node. |
| back | n/a | complex | complex | Structural node. |
| bibl | block | complex | mixed |  |
| bibref | inline | complex | empty | Reference, link text to be computed. |
| blist | n/a | complex | complex | Structural node. |
| body | n/a | complex | complex | Structural node. |
| cell (th, td) | ***n/a or block ?*** |  |  |  |
| code | inline | simple | simple | Fixed-width (terminal) font |
| compdef | n/a | complex | complex | Structural node. |
| div | n/a | complex | complex | Structural node. |
| **head** (in a <div>) | inline | simple | simple | Formatted as heading |
| **head** (in a <constraintnote>) | inline | complex | mixed |  |
| eg | block | complex | simple |  |
| eltref | inline | complex | empty | Reference, link text to be computed. |
| emph |  |  |  | Emphasis, typically *italic* text. Formatting must be transmitted to the children, i.e. phrase, and indirectly to loc, code, term, refs, … |
| header | n/a | complex | complex | Structural node. |
| loc | inline | complex | simple | Reference, link text in the text node. |
| local | inline | simple | simple | Apparently just **bold** text. |
| note | n/a | complex | complex | Structural node. |
| orglist | n/a | complex | complex | Structural node. |
| p | block | complex | mixed |  |
| phrase | inline | complex | mixed |  |
| propdef | block | complex | mixed |  |
| proplist | n/a | complex | complex | Structural node. |
| propmap | block | complex | mixed |  |
| propref | inline | complex | empty | Reference, link text to be computed. |
| pt | inline | simple | simple | Apparently a no-op |
| reprcomp | block+ | complex | complex | More than just block, table ? |
| reprdef | n/a | complex | complex | Structural node. |
| reprelt | block+ | complex | empty | More than just block, XML repr/grammar ? |
| spec | n/a | complex | complex | Structural node. |
| specref | inline | complex | empty | Reference, link text to be computed. |
| status | n/a | complex | complex | Structural node. |
| table |  |  |  |  |
| tbody |  |  |  |  |
| term | inline | simple | simple | Inside a definition: the term being defined (currently **bold**). |
| termdef | inline | complex | complex | A term definition. |
| termref | inline | complex | empty | Reference, link text to be computed. |
| tr |  |  |  |  |
| xpropref | inline | complex | simple | Reference, link text in the text node. |
| xspecref | inline | complex | simple | Reference, link text in the text node. |
| glist | n/a | complex | complex | Structural node. |
| gitem | n/a | complex | complex | Structural node. |
| label | block+ | complex | mixed |  |
| def | n/a | complex | complex | Structural node. |
| ulist | n/a | complex | complex | Structural node. |
| olist | n/a | complex | complex | Structural node (with attribute) |
| item | n/a | complex | complex | Structural node. |
| quote | inline | simple | simple | Enclosed by double-quotes |

xxx

# Styles

We want to assign MS Word styles to the different tpes of formatting we find in the W3C documents. Each format corresponds to some XML tag or combination of tags.

There are 100 diferents tags in [XML Schema Part 1: Structures](https://www.w3.org/TR/2004/REC-xmlschema-1-20041028/structures.html), we're going to classify them as inline/block and assign a style to each one.

|  |  |
| --- | --- |
| **Tag** | **Type** |
| abstract |  |
| affiliation |  |
| all |  |
| altlocs |  |
| any |  |
| authlist |  |
| author |  |
| back |  |
| bibl | block |
| bibref |  |
| blist | n/a |
| body |  |
| choice |  |
| clauseref |  |
| code |  |
| compdef |  |
| constraintnote |  |
| day |  |
| def |  |
| div1 |  |
| div2 |  |
| div3 |  |
| div4 |  |
| ednote |  |
| edtext |  |
| eg |  |
| elt |  |
| eltref |  |
| email |  |
| emph |  |
| errataloc |  |
| gitem |  |
| glist |  |
| graphic |  |
| head |  |
| header |  |
| iiName |  |
| imagemap |  |
| item |  |
| label |  |
| language |  |
| langusage |  |
| latestloc |  |
| lhs |  |
| loc |  |
| local |  |
| member |  |
| month |  |
| name |  |
| note |  |
| nt |  |
| olist |  |
| orglist |  |
| p |  |
| phrase |  |
| preverrataloc |  |
| prevlocs |  |
| prod |  |
| propdef | block |
| proplist |  |
| propmap |  |
| propref |  |
| pt |  |
| pubdate |  |
| publoc |  |
| pubstmt |  |
| pvlist |  |
| pvpair |  |
| quote |  |
| reprcomp |  |
| reprdef |  |
| reprelt |  |
| restrict |  |
| restrictCases |  |
| revisiondesc |  |
| rhs |  |
| role |  |
| schemaComp |  |
| scrap |  |
| seq |  |
| sitem |  |
| slist |  |
| sourcedesc |  |
| spec |  |
| specref |  |
| status |  |
| term |  |
| termdef |  |
| termref |  |
| title |  |
| translationloc |  |
| ulist |  |
| version |  |
| w3c-designation |  |
| w3c-doctype |  |
| xnt |  |
| xpropref |  |
| xspecref |  |
| xtermref |  |
| year |  |