Matching Guidelines

**SPDX License List Matching Guidelines, v2.5**

This document provides guidelines to be used for the purposes of matching licenses and license exceptions against those included on the SPDX License List. There is no intent here to make a judgment or interpretation, but merely to ensure that when one SPDX user identifies a license as "BSD 3-clause," for example, it is indeed the same license as what someone else identifies as "BSD 3-clause" and the same license as what is listed on the SPDX License List. Examples of how to apply some of the matching guidelines to a license or exception are provided via templates.  Templates are comprised of technical markup within the master license xml file to provide further or specific guidance to SPDX document creators or tool makers.

**1. How These Guidelines Are Applied**

1.1 Purpose:  To ensure consistent results by different SPDX document creators when matching license information that will be included in the License Information in File field. SPDX document creators or tools may match on the license or exception text itself, the official license header, or the SPDX License List short identifier.

1.1.1 Guideline: Official License Headers  The same matching guidelines used for license and exception text apply to the official license headers. Where applicable, an official license header template file with xml markup is included with the SPDX License List. Official license headers are defined by the SPDX License List as specific text specified within the license itself to be put in the header of files. (see <http://spdx.org/spdx-license-list/license-list-overview> for more info). The markup for headers is <header> and </header>.

**2. Extensible Markup Language (XML)**

2.1 Purpose:  To enable automated matching and parsing of SPDX licenses by tools, a SPDX template is marked up with xml.

2.1.1 Guideline: XML Format The SPDX xml structure appears below:

<SPDX name="full name" identifier="SPDX short identifier" osi-approved="true/false">

<urls>

<url>URL for license</url>

<url>additional URL for license</url>

</urls>

<notes>notes on license, if any</notes>

<header>standard license header, if any</header>

<license>

<title>license name or title (Guideline 11) </title>

<copyright>Copyright notice (Guideline 10) </copyright>

<body> Substantive text of license - most matching guidelines apply here

<list>

<li><b>a bullet (Guideline 8) </b>text of the list item</li>

</list>

<p>some more text</p>

</body>

<optional>some text that is omitable (Guideline 3) </optional>

</license>

</SPDX>

* + 1. Guideline: XML Tags Recognized SPDX xml tags are below. The matching SPDX field is listed in parentheses.
* SPDX tag: root element tag; allows following attributes:
  + “name” (name) – Full license name, e.g., BSD 3-clause “New” or “Revised” License.
  + “identifier” (licenseId) – Short identifier, e.g., BSD-3-Clause.
  + “OSI-approved” (isOsiApproved) – “true” if OSI approved, otherwise blank
* URLs/URL tag (seeAlso): each url for the official text of the license or exception; URL is nested within a URLs tag.
* Notes tag (comment): Factual info about the license, e.g. date of release, license has been deprecated, links to official translations
* Header tag (standardLicenseHeader): the short form of a license that references the full license text.
* License tag (license): the root element that contains the license data
* Title tag: denotes an optional (as per Matching Guideline 11) section of the license content, with the implication that it contains the title/heading content that prefaces the substantive text
* Copyright tag: denotes an optional (as per Matching Guideline 10) section of the license content, with the implication that it contains copyright notices
* Body tag: substantive text of the license itself
* Optional tag: denotes otherwise optional (as per Matching Guideline 3) text content, such as appendices explaining how to use a license and/or defining the license's header form
* alt tag: denotes a small portion of the text that is specific to the content being licensed, such as the name of the copyright holder; allows following attribute:
  + "name" - ignored for matching purposes, but specifies a unique identifier to store the matched content.
  + "match" - a regular expression for text to match in lieu of the original text content
* syn tag: denotes meaningful substantive text that may have variations in spelling (regional or otherwise), for which alternate spellings are valid and have identical meaning; allows following attributes:
  + "identifier" - specifies name of set of synonymous words; tag wraps the original text as given in the license.
* List/Li tags: denotes a list per Matching Guideline 8; li denotes individual list items and must be nested in a list tag.
* B tag: denotes a bullet or section marker whose style and formatting is unimportant to the substantive content of the license text as per Matching Guideline 8; commonly found in li tags
* P tag: denotes a paragraph; does not affect matching.
* Br tag: denotes a line break; does not affect matching.

**3. Substantive Text**

3.1 Purpose:  To ensure that when matching licenses and exceptions to the SPDX License List, there is an appropriate balance between matching against the substantive text and disregarding parts of the text that do not alter the substantive text. Further guidelines of what can be disregarded or considered replaceable for purposes of matching are listed below here and in the subsequent specific guidelines. A conservative approach is taken in regards to rules about disregarded or replaceable text.

3.1.1 Guideline:  Verbatim Text  License and exception text should be the same verbatim text (except for the guidelines stated here). The text should be in the same order, e.g., differently ordered paragraphs would not be considered a match. The markup for text is <body> and </body>.

3.1.2 Guideline:   No Additional Text Matched text should only include that found in the vetted license or exception text. Where a license or exception found includes additional text or clauses, this should not be considered a match.

3.1.3 Guideline:  Replaceable Text  Some licenses include text that refers to the specific copyright holder or author, yet the rest of the license is exactly the same as a generic version. The intent here is to avoid the inclusion of a specific name in one part of the license resulting in a non-match where the license is otherwise an exact match (e.g., the third clause and disclaimer in the BSD licenses, or the third, fourth, and fifth clauses of Apache 1.1). In these cases, there should be a positive license match.

Text that can be considered replaceable for matching purposes is indicated in the SPDX License List template with xml markup and in the corresponding HTML pages with colored text. The text indicated as such can be replaced with similar values (e.g., a different name or generic term; different date) and still be considered a positive match. This rule also applies to text-matching in official license headers (see Guideline #1).

The markup for replaceable text is <alt> and </alt>. Available attributes are “name” and “match”. The name attribute is optional, but should specify a unique identifier. The match attribute specifies a regular expression of valid matches.

3.1.4 Guideline:  Omitable Text  Some licenses have text that can simply be ignored. The intent here is to avoid the inclusion of certain text that is superfluous or irrelevant in regards to the substantive license text resulting in a non-match where the license is otherwise an exact match (e.g., directions on how to apply the license or other similar non-substantive exhibits). In these cases, there should be a positive license match. Text that can be considered omitable for matching purposes is indicated in the SPDX License List template with markup and in the corresponding HTML pages with colored text. The license should be considered a match if the text indicated is present and matches OR the text indicated is missing altogether. The markup for omitable text is <optional> and </optional>. This tag may also be used to identify extraneous typos found in the license text such as an extra period or stray quotation mark if corrected versions of the license have also been found.

**4. Whitespace**

4.1 Purpose:   To avoid the possibility of a non-match due to different spacing of words, line breaks, or paragraphs.

4.1.1 Guideline:  All whitespace should be treated as a single blank space. Templates do not include markup for this guideline. While not used for matching purposes, <p> and </p> denote paragraphs and <br /> denote line breaks in the text.

**5. Capitalization**

5.1 Purpose:  To avoid the possibility of a non-match due to lower case or upper case letters in otherwise the same words.

5.1.1 Guideline:  All upper case and lower case letters should be treated as lower case letters*.*Templates do not include markup for this guideline.

**6. Punctuation**

6.1 Purpose:  Because punctuation can change the meaning of a sentence, punctuation needs to be included in the matching process. License template files do not include markup for this guideline.

6.1.1 Guideline:  Punctuation  Punctuation should be matched, unless otherwise stated in these guidelines.

6.1.2 Guideline:  Hyphens, Dashes  Any hyphen, dash, en dash, em dash, or other variation should be considered equivalent.

6.1.3 Guideline:  Quotes  Any variation of quotations (single, double, curly, etc.) should be considered equivalent.

**7. Code Comment Indicators**

7.1 Purpose:  To avoid the possibility of a non-match due to the existence or absence of code comment indicators placed within the license text, e.g. at the start of each line of text.

7.1.1 Guideline:  Any kind of code comment indicator or prefix which occurs at the beginning of each line in a matchable section should be ignored for matching purposes. Templates do not include markup for this guideline.

**8. Bullets and Numbering**

8.1 Purpose:  To avoid the possibility of a non-match due to the otherwise same license using bullets instead of numbers, number instead of letter, or no bullets instead of bullet, etc., for a list of clauses.

8.1.1 Guideline:  Where a line starts with a bullet, number, letter, or some form of a list item (determined where list item is followed by a space, then the text of the sentence), ignore the list item for matching purposes. The markup for bullets and numbering is <list> and </list> with nested <li> and </li>.

**9. Varietal Word Spelling**

9.1 Purpose:  English uses different spelling for some words. By identifying the spelling variations for words found or likely to be found in licenses, we avoid the possibility of a non-match due to the same word being spelled differently. This list is not meant to be an exhaustive list of all spelling variations, but meant to capture the words most likely to be found in open source software licenses.

9.1.1 Guideline:  The words in the following columns are considered equivalent and interchangeable. The markup for equivalent words are <syn> and </syn>. The "identifier" attribute specifies the set of synonymous words that may match and the tag wraps the original text as given in the license proper.

|  |  |
| --- | --- |
| Column 1 | Column 2 |
| 1. Acknowledgement  2. Analog  3. Analyze  4. Artifact  5. Authorization  6. Authorized  7. Caliber  8. Canceled  9. Capitalizations  10. Catalog  11. Categorize  12. Center  13. Emphasized  14. Favor  15. Favorite  16. Fulfill  17. Fulfillment  18. Initialize  19. Judgement  20. Labeling  21. Labor  22. License  23. Maximize  24. Modeled  25. Modeling  26. Offense  27. Optimize  28. Organization  29. Organize  30. Practice  31. Program  32. Realize  33. Recognize  34. Signaling  35. Sublicense  36. Sub-license  37. Utilization  38. While  39. Wilfull  40. Noncommercial  41. Percent  42. Copyright holder | 1. Acknowledgment  2. Analogue  3. Analyse  4. Artefact  5. Authorisation  6. Authorised  7. Calibre  8. Cancelled  9. Capitalisations  10. Catalogue  11. Categorise  12. Centre  13. Emphasised  14. Favour  15. Favourite  16. Fulfil  17. Fulfilment  18. Initialise  19. Judgment  20. Labelling  21. Labour  22. Licence  23. Maximise  24. Modelled  25. Modelling  26. Offence  27. Optimise  28. Organisation  29. Organise  30. Practise  31. Programme  32. Realise  33. Recognise  34. Signalling  35. Sub-license  36. Sub license  37. Utilisation  38. Whilst  39. Wilful  40. Non-commercial  41. Per cent  42. Copyright owner |

**10. Copyright Symbol**

10.1 Purpose:  By having a rule regarding the use of "©", "(c)", or "copyright", we avoid the possibility of a mismatch based on these variations.

10.1.1 Guideline:  "©", "(c)", or "Copyright" should be considered equivalent and interchangeable. Copyright should be identified with the syn identifier attribute “”.

**11. Copyright Notice**

11.1 Purpose:  To avoid a license mismatch merely because the copyright notice (usually found above the actual license or exception text) is different. The copyright notice is important information to be recorded elsewhere in the SPDX file, but for the purposes of matching a license to the SPDX License List, it should be ignored because it is not part of the substantive license text.

11.1.1 Guideline:  Ignore copyright notices. A copyright notice consists of the following elements, for example: "2012 Copyright, John Doe. All rights reserved." or "(c) 2012 John Doe." The markup for copyright notices is <copyright> and </copyright>.

**12. License Name or Title**

12.1 Purpose:  To avoid a license mismatch merely because the name or title of the license is different than how the license is usually referred to or different than the SPDX full name. This also avoids a mismatch if the title or name of the license is simply not included.

12.1.1 Guideline:  Ignore the license name or title for matching purposes, so long as what ignored is the title only and there is no additional substantive text added here. The markup for license names or titles is <title> and </title>

**13. Extraneous Text At the End of a License**

13.1 Purpose:  To avoid a license mismatch merely because extraneous text that appears at the end of the terms of a license is different or missing. This also avoids a mismatch if the extraneous text merely serves as a license notice example and includes a specific copyright holder's name.

13.1.1 Guideline:  Ignore any text that occurs after the obvious end of the license and does not include substantive text of the license, for example: text that occurs after a statement such as, "END OF TERMS AND CONDITIONS," or an exhibit or appendix that includes an example or instructions on to how to apply the license to your code. Do not apply this guideline or ignore text that is comprised of additional license terms (e.g., permitted additional terms under GPL-3.0, section 7). The markup for extraneous text is <optional> and </optional>.