JATS4R Subject and Keyword Guidelines

Status: Full Version
Version: 1.0
Date: July 19, 2019
Provenance: JATS4R Sub-group [Helen Alexander, Emerald Publishing (prev.); Franziska Buehring, De Gruyter; Rachel Carriere, EBSCO; Mike Eden, Cambridge University Press; Kara Handren, University of Toronto; Kevin Lawson, Sheridan (chair); Vincent Lizzi, Taylor & Francis; Jonathan Watson, Emerald Publishing]
History: Working Draft (begin): November 18, 2017
   Steering Committee Review (1): December 6, 2018
   Steering Committee Review (2): January 29, 2019
   Steering Committee Review (3): March 5, 2019
   Public Review: March 22, 2019
   Steering Committee Review (4): June 25, 2019
   Full Version: July 19, 2019:
   Steering Committee Approval: September 9, 2019
   NISO Information Creation & Curation Topic Committee Approval: October 4, 2019
Published: March 2, 2020

Description
This set of recommendations contains best practices for tagging keywords and subject data using NISO JATS 1.1. It treats keywords as distinct from subjects, the former of which pertain to indexing rather than generation of an eTOC or other display-related purpose. A keyword, in JATS usage, is a subject term, key phrase, abbreviation, indexing word, taxonomic structure, or other expression that is associated with the entire article for the purpose of searching and indexing. A subject, in JATS usage, is a term describing a document’s content, or a component document’s content, for the purpose of grouping documents for display or print for the purpose of browsing. Both subjects and keywords can use uncontrolled or controlled vocabularies.

The sub-group has reviewed the new elements and attributes in the NISO Journal Article Tag Suite (JATS) v1.2; any recommendations that involve elements or attributes from this set are identified in these recommendations for the readers’ consideration.

Objectives
Objectives of these recommendations:
1. To minimize usage crossover between keywords and subjects (e.g., capturing keywords in subject groups that are intended for indexing, etc.)
2. Enable machines to clearly identify text that has been captured for display rather than indexing and vice versa
3. To establish clear guidelines and validation rules for keywords and subjects

Context
Keywords
- Elements = <kwd-group>, <kwd>, <compound-kwd>, <compound-keyword-part>, and <nested-kwd>. The element <unstructured-kwd-group> is only in the Journal Archiving and Interchange Tag Set.
- Attributes = @kwd-group-type, @specific-use, @xml:lang, @content-type. There are four vocabulary attributes that were introduced in JATS 1.2: @vocab, @vocab-identifier, @vocab-term, and @vocab-term-identifier.
Subjects
- Elements = `<subj-group>`, `<subject>`, `<compound-subject>`, `<compound-subject-part>`
- Attributes = `@subj-group-type`, `@specific-use`, `@xml:lang`, `@content-type`. There are four vocabulary attributes that were introduced in JATS 1.2: `@vocab`, `@vocab-identifier`, `@vocab-term`, and `@vocab-term-identifier`.

References

Recommendations
1. Language for Keywords and Subjects
   1.1. Summary: `@xml:lang` should be present on `<kwd-group>` or `<subj-group>` if the language is different than `@xml:lang` used at the `<article>` level.
   1.2. Examples
      (a) Specifying different languages in keyword groups
         ...
         <kwd-group xml:lang="fr">
           <title>Mots-clés</title>
           <kwd>conservation de la diversité biologique</kwd>
           <kwd>développement communautaire durable</kwd>
           <kwd>gestion communautaire</kwd>
           <kwd>races indigènes</kwd>
           <kwd>ressources zoogénétiques</kwd>
         </kwd-group>
         <kwd-group xml:lang="es">
           <title>Palabras clave</title>
           <kwd>conservación de la biodiversidad</kwd>
           <kwd>desarrollo sostenible de la comunidad</kwd>
           <kwd>gestión basada en la comunidad</kwd>
           <kwd>razas autóctonas</kwd>
           <kwd>recursos zoogenéticos</kwd>
         </kwd-group>
         ...
      
      (b) Specifying language for subject group
         ...
         <subj-group subj-group-type="primary" xml:lang="en">
           <subject>Chemistry</subject>
           <subj-group subj-group-type="secondary">
             <subject>Industrial Chemistry/Chemical Engineering</subject>
           </subj-group>
         </subj-group>
         ...

1.3. Validator:
   a) `<(kwd|subj)-group>` has the same `@xml:lang` as the article [WARNING]
   b) There should not be two `<(kwd|subj)-group>` without a `@(kwd|subj)-group-type` or `@xml:lang` attribute. [WARNING]

   Examples:

   ```xml
   kwd-group[@kwd-group-type]
   kwd-group[@xml-lang]
   kwd-group
   kwd-group[@kwd-group-type]
   kwd-group[@kwd-group-type]
   kwd-group[@kwd-group-type]
   kwd-group[@xml-lang]
   kwd-group[@xml-lang]
   kwd-group[@kwd-group-type]
   kwd-group
   kwd-group
   ```

2. Location of Keywords

   2.1. Summary: For optimal system use, keywords should be included in `<article-meta>` if they apply to the article as a whole. However, in cases when keywords are intended only to apply to a particular area rather than to the entire article, it is acceptable to capture keywords in that area only. When keywords are captured in a specific object (figure, table, section), they can be duplicated in `<article-meta>`. It is acceptable for `<kwd-group>` to be contained in any location that is allowed by the DTD. Keywords that appear in specific objects should follow the same tagging practices as keywords that appear at the article level.

   2.2. Example

   ```xml
   ...<fig>
   <label>Figure 6</label>
   <caption>Experimental set-up of corrosion rate determination</caption>
   <kwd-group><kwd>null hypothesis</kwd></kwd-group>
   </kwd-group>
   </fig>
   ...
   ```

2.3. Validator: N/A

3. Multi-part Keywords

   3.1. Summary: If a scenario exists where there are multiple types of keywords or multiple keyword parts, care should be taken to represent them in a standard way. There are several options for representing multi-part keywords depending on how the parts are related and what is intended for display versus indexing. JATS 1.1 and 1.2 offer dedicated elements for compound keywords that can be used for representing code-term pairs, or abbreviations and expansions (for more information see recommendation 4). There are also dedicated elements for nested keywords that represent hierarchically related terms (see recommendation 8 and sample c). For multi-part keywords where one part is a label for display, there are a number of ways to represent this, but the recommendations...
are samples a and b below, with (a) being the preferred option because fewer keyword groups will be needed; all can be contained within one group. The @content-type attribute should be used whenever a descriptor is available. Finally, sample (c) is an illustration of a representation to avoid. The terms modeled in sample (c) would be best represented using nested-kwd elements as described in recommendation 8.

3.2. Examples
(a) *PREFERRED* Defining parts by using @content-type(s)

```xml
<kwd-group kwd-group-type="author-generated">
  <kwd content-type="subjects">cancer</kwd>
  <kwd content-type="subjects">cells</kwd>
  <kwd content-type="subjects">therapeutic development</kwd>
  <kwd content-type="materials">mice</kwd>
  <kwd content-type="materials">vincristine</kwd>
  <kwd content-type="materials">procarbazine</kwd>
  <kwd content-type="methods">data analysis</kwd>
  <kwd content-type="methods">statistical analysis</kwd>
  <kwd content-type="methods">in vivo experiments</kwd>
</kwd-group>
<kwd-group kwd-group-type="publisher">
  <kwd content-type="subjects">oncology</kwd>
  <kwd content-type="subjects">cells</kwd>
  <kwd content-type="subjects">chemotherapy</kwd>
  <kwd content-type="materials">live specimens</kwd>
  <kwd content-type="materials">leurocristine</kwd>
  <kwd content-type="materials">alkylating agents</kwd>
  <kwd content-type="methods">data analysis</kwd>
  <kwd content-type="methods">statistical analysis</kwd>
  <kwd content-type="methods">in vivo experiments</kwd>
</kwd-group>
```

(b) Defining parts by using @kwd-group-type(s)

```xml
<kwd-group kwd-group-type="subjects">
  <label>Author Provided Keywords</label>
  <kwd>cancer</kwd>
  <kwd>cells</kwd>
  <kwd>therapeutic development</kwd>
</kwd-group>
<kwd-group kwd-group-type="materials">
  <label>Author Provided Keywords</label>
  <kwd>mice</kwd>
  <kwd>vincristine</kwd>
  <kwd>procarbazine</kwd>
</kwd-group>
<kwd-group kwd-group-type="methods">
  <label>Author Provided Keywords</label>
  <kwd data analysis</kwd>
  <kwd statistical</kwd>
</kwd-group>
```
<kwd>in vivo experiments</kwd>
</kwd-group>
<kwd-group kwd-group-type="subjects">
  <label>Publisher Provided Keywords</label>
  <kwd>oncology</kwd>
  <kwd>cells</kwd>
  <kwd>chemotherapy</kwd>
</kwd-group>
<kwd-group kwd-group-type="materials">
  <label>Publisher Provided Keywords</label>
  <kwd>live specimens</kwd>
  <kwd>leurocristine</kwd>
  <kwd>alkylating agents</kwd>
</kwd-group>
<kwd-group kwd-group-type="methods">
  <label>Publisher Provided Keywords</label>
  <kwd>data analysis</kwd>
  <kwd>statistical</kwd>
  <kwd>in vivo experiments</kwd>
</kwd-group>

(c) *NOT RECOMMENDED* Defining parts within the captured text

...<kwd-group kwd-group-type="keywords">
  <kwd>methods: numerical</kwd>
  <kwd>methods: analytical</kwd>
  <kwd>galaxies: clusters: general</kwd>
  <kwd>galaxies: evolution</kwd>
  <kwd>cosmology: miscellaneous</kwd>
  <kwd>stars: planetary</kwd>
  <kwd>stars: low mass</kwd>
</kwd-group>
...

3.3. Validator: If <kwd> is present multiple times within <kwd-group>, and there is more than one <kwd-group>, and @content-type is present on at least one <kwd>, @content-type should be present on all <kwd>s. [WARNING]

4. Compound Keywords & Subjects
4.1. Summary: The compound keyword structure should be used for code and term pairs, which are representative of the same concept. Compound keywords can also be used to capture abbreviations and their expansions when both parts are intended to be treated as keywords for indexing purposes. Capturing an abbreviation list in a keyword group is discouraged unless both abbreviations and expansions are required for indexing. Instead, a definition list should be placed where the list needs to be displayed. See Section 5 for examples representing abbreviations and expansions. [Note: The online JATS documentation shows an abbreviation and expansion example for compound keywords.]
4.2. **Examples**

(a) Compound keywords showing a single code and term pair

```xml
<compound-kwd>
  <compound-kwd-part content-type="ISO-463-code">863</compound-kwd-part>
  <compound-kwd-part content-type="ISO-463-text">Icelandic sagas</compound-kwd-part>
</compound-kwd>
```

(b) Compound subjects showing multiple code and term pairs

```xml
<subj-group subj-group-type="topic">
  <compound-subject>
    <compound-subject-part content-type="code">02_0260</compound-subject-part>
    <compound-subject-part content-type="label">Energy and Materials</compound-subject-part>
  </compound-subject>
  <compound-subject>
    <compound-subject-part content-type="code">02_0840</compound-subject-part>
    <compound-subject-part content-type="label">Surface engineering</compound-subject-part>
  </compound-subject>
</subj-group>
```

(c) Example of structure that is recommended only when both parts of keywords are to be indexed

```xml
<kwd-group>
  <compound-kwd>
    <compound-kwd-part content-type="abbrev">AODM</compound-kwd-part>
    <compound-kwd-part content-type="expansion">adult onset diabetes mellitus</compound-kwd-part>
  </compound-kwd>
  <compound-kwd>
    <compound-kwd-part content-type="abbrev">DI</compound-kwd-part>
    <compound-kwd-part content-type="expansion">diabetes insipidus</compound-kwd-part>
  </compound-kwd>
  <compound-kwd>
    <compound-kwd-part content-type="abbrev">DKA</compound-kwd-part>
    <compound-kwd-part content-type="expansion">diabetic ketoacidosis</compound-kwd-part>
  </compound-kwd>
</kwd-group>
```

4.3. **Validator:** Check if there is only one `<compound-kwd-part>/<compound-subject-part>`. If so, then [WARNING].
5. Abbreviations and Expansions

5.1. Summary: Abbreviations should not be tagged as keywords unless there is a specific need to index them because there are already <abbrev>, <glossary>, and <def-list> elements. Abbreviations are used in articles to make the text less cumbersome to read and are spelled out, per journal style, either in the text or in a list so readers know what the abbreviations mean. Keywords are added to articles to make the article more searchable online. One would not necessarily want to include all of their in-text abbreviation terms in their Keywords. See sample (a) below for example of display-related usage. If there is a need for indexing-related usage, then see sample (b) below, which utilizes the compound keyword structure described in Recommendation #4.

5.2. Examples

(a) Display-related usage


<article dtd-version="1.1">...
    <back>...
        <glossary>
            <def-list>
                <title>ABBREVIATIONS</title>
                <term-head>Abbreviation</term-head>
                <def-head>Expansion</def-head>
                <def-item><term id="G1">PAP I</term>
                    <def><p>poly(A)polymerase I</p></def>
                </def-item>
                <def-item><term id="G2">PNPase</term>
                    <def><p>polynucleotide phosphorylase</p></def>
                </def-item>
            </def-list>
        </glossary>
    </back>
</article>

(b) Indexing-related usage


...<article-meta>...
    <abstract>...
        <kwd-group kwd-group-type="author">
            <compound-kwd>
                <compound-kwd-part content-type="abbrev">AODM</compound-kwd-part>
                <compound-kwd-part content-type="expansion">adult onset diabetes mellitus</compound-kwd-part>
            </compound-kwd>
            <compound-kwd>
                <compound-kwd-part content-type="abbrev">DI</compound-kwd-part>
                <compound-kwd-part content-type="expansion">diabetes insipidus</compound-kwd-part>
            </compound-kwd>
            <compound-kwd>
                <compound-kwd-part content-type="abbrev">DKA</compound-kwd-part>
                <compound-kwd-part content-type="expansion">diabetic ketoacidosis</compound-kwd-part>
            </compound-kwd>
        </kwd-group>...
5.3. Validator: N/A

6. Unstructured Keywords

6.1. Summary: Unstructured keywords can be used to capture archival content (e.g., reconverting back content when there is no resource to tag each keyword individually). It is present in the Journal Archiving and Interchange (green) version only. Unstructured keywords are strongly discouraged except for special cases (e.g., initial pass of capturing data from PDFs where further passes would use structured keywords).

6.2. Example

```xml
<article-meta>...
  <abstract>...</abstract>
  <kwd-group kwd-group-type="author">
    <unstructured-kwd-group>molecular chaperones; surface plasmon resonance; dynamic light scattering; trypsin digestion; citrate synthase</unstructured-kwd-group>
  </kwd-group>
</article-meta>...

6.3. Validator: N/A

7. Primary and Secondary Classifications in Keywords and Subjects

7.1. Summary: If it is necessary to include this information in JATS 1.1., we recommend using @subj-group-type or @kwd-group-type. In these cases, group all subjects/keywords of the same type together, like the example below (sec. 7.2., sample a). If @subj-group-type or @kwd-group-type is already in use, another option is to use @specific-use to define primary and secondary terms in JATS 1.1 (sec 7.2, sample b).

In JATS 1.2, @vocab-term and @vocab-term-identifier can be used, which enables @content-type to define whether the term is primary or secondary. In the following JATS 1.2 example (sec 7.2., sample c), the intent is for the value of @vocab-term to not display on the target system.

We suggest not mixing different vocabularies in the same <kwd-group> or <subj-group> as a Best Practice, even though all four @vocab related attributes are possible on any <kwd> or <subject>.

7.2. Examples

a) Using @subj-group-type in JATS 1.1 to define priority with all similar types grouped together

```xml
  <subj-group subj-group-type="Primary">
    <subject>Agriculture</subject>
    <subject>Agricultural Methods</subject>
  </subj-group>
  <subj-group subj-group-type="Secondary">
    <subject>Marketing Cartels</subject>
    <subject>Output ceilings</subject>
    <subject>Agricultural Soil Science</subject>
  </subj-group>
```

```xml
  ...
```
b) Using @specific-use in JATS 1.1 to define priority if @subj-group-type is already in use

```xml
<subj-group subj-group-type="Heading" specific-use="Primary">
  <subject>Agriculture</subject>
  <subject>Agricultural Methods</subject>
</subj-group>
<subj-group subj-group-type="Heading" specific-use="Secondary">
  <subject>Marketing Cartels</subject>
  <subject>Output ceilings</subject>
  <subject>Agricultural Soil Science</subject>
</subj-group>
```

...  

c) In JATS 1.2, using @content-type to define priority with @vocab-term used as non-displaying term on target system

```xml
<kwd-group kwd-group-type="classification" vocab="INSPEC">
  <kwd content-type="primary-heading" vocab-term="B1265D">Memory circuits</kwd>
  <kwd content-type="primary-heading" vocab-term="C5320G">Semiconductor storage</kwd>
  <kwd content-type="secondary-heading" vocab-term="B1265A">Digital circuit design, modelling and testing</kwd>
  <kwd content-type="secondary-heading" vocab-term="C5210">Logic design methods</kwd>
</kwd-group>
```

...  

d) In JATS 1.2, using @content-type to define priority without any of the vocab-related attributes

```xml
<kwd-group kwd-group-type="classification">
  <kwd content-type="primary-heading">Marketing</kwd>
  <kwd content-type="secondary-heading">Business-to-business marketing</kwd>
  <kwd content-type="secondary-heading">Business-to-consumer marketing</kwd>
  <kwd content-type="primary-heading">Business entities</kwd>
  <kwd content-type="secondary-heading">Small to medium-sized enterprises</kwd>
  <kwd content-type="secondary-heading">Family businesses</kwd>
</kwd-group>
```

...  

7.3. Validator: N/A
8. Nested Keywords and Nesting Subject Groups to Define Hierarchical Level

8.1. Summary: The <nested-kwd> structure should be used for indexing a set of keywords that are related to each other hierarchically, such as keywords from a taxonomy. For subject elements that are related hierarchically, the <subject> elements can be nested in the same structure as the <nested-kwd> elements. If the purpose of the subject group is to populate the display of a heading(s) on a table of contents, then @subj-group-type="heading" should be used on <subj-group>.

8.2. Examples:

a) Using nested keywords to define hierarchy for indexing

```xml
<kwd-group kwd-group-type="MeSH">
  <nested-kwd>
    <kwd>Diagnosis</kwd>
    <nested-kwd>
      <kwd>Diagnostic Techniques and Procedures</kwd>
      <nested-kwd>
        <kwd>Diagnostic Imaging</kwd>
      </nested-kwd>
    </nested-kwd>
  </nested-kwd>
</kwd-group>
```

b) Nesting subject groups to define hierarchy

```xml
<subj-group subj-group-type="Discipline-v3">
  <subject>Engineering and technology</subject>
  <subj-group>
    <subject>Electronics</subject>
    <subj-group>
      <subject>Capacitors</subject>
    </subj-group>
  </subj-group>
</subj-group>
```

c) JATS 1.1 example with hierarchical levels utilizing compound keywords and code/term pairs

```xml
<kwd-group kwd-group-type="hierarchical">
  <nested-kwd>
    <compound-kwd>
      <compound-kwd-part content-type="code">01</compound-kwd-part>
      <compound-kwd-part content-type="term">Mathematical Sciences</compound-kwd-part>
    </compound-kwd>
    <nested-kwd>
      <compound-kwd>
        <compound-kwd-part content-type="code">0101</compound-kwd-part>
        <compound-kwd-part content-type="term">Pure Mathematics</compound-kwd-part>
      </compound-kwd>
    </nested-kwd>
  </nested-kwd>
</kwd-group>
```
<compound-kwd>
  <compound-kwd-part content-type="code">010101</compound-kwd-part>
  <compound-kwd-part content-type="term">Algebra and Number Theory</compound-kwd-part>
</compound-kwd>

... 

d) JATS 1.2 example with hierarchical levels utilizing nested keywords
...
<kwd-group kwd-group-type="hierarchical" vocab="FoR" vocab-identifier="ANZSRC 2008">
  <nested-kwd>
    <kwd content-type="term" vocab-term-identifier="01">Mathematical Sciences</kwd>
    <nested-kwd>
      <kwd content-type="term" vocab-term-identifier="0101">Pure Mathematics</kwd>
      <nested-kwd>
        <kwd content-type="term" vocab-term-identifier="010101">Algebra and Number Theory</kwd>
      </nested-kwd>
    </nested-kwd>
  </nested-kwd>
</kwd-group>
...

8.3. Validator: N/A; unable to programmatically determine intended relationships or headings

9. Subject Groups - General Usage
9.1 Summary: Generally speaking, publishers use <subj-group> either to indicate the document’s type or its topical subject, and that is what is indicated by the @subj-group-type attribute (when it’s used). We do not prescribe specific values here, but for the purpose of helping to illustrate the two most common scenarios with Subject Groups, following are examples.


... 
<subj-group subj-group-type="heading">
  <subject>Original Article</subject>
</subj-group>
...

```xml
...<subj-group subj-group-type="subject">
    <subject>African studies</subject>
</subj-group>

<subj-group subj-group-type="section">
    <subject>VIRAL INFECTIONS</subject>
</subj-group>

<subj-group subj-group-type="Primary">
    <subject>S10</subject>
</subj-group>
<subj-group subj-group-type="Secondary">
    <subject>S8</subject>
</subj-group>

<subj-group subj-group-type="System Taxonomy">
    <subject>Drosophila</subject>
</subj-group>
...