WORKING DRAFT

Baseprint Document Format (BpDF)

E. Castedo Ellerman (castedo@castedo.com)

Copyright:
© 2025, Ellerman et al
CC BY License
This document is distributed under a Creative Commons Attribution 4.0 International license.

Abstract

DOCUMENT TYPE: Living Technical Specification

The Baseprint Document Format is the digital encoding of a Baseprint document snapshot. These snapshots are immutable and are referenced using a SoftWare Hash IDentifier (SWHID). This format is designed for self-archived scientific and technical documents for long-term redistribution. The XML component of the format is a subset of the JATS XML Article Authoring Tag Set and approximates XML found in PubMed Central. After archiving, document snapshots are rendered into HTML pages and PDF files by independent websites using Baseprint-compatible software.

Feedback

In addition to email, feedback can also be communicated through the GitHub repository of the source files for this edition at github.com/castedo/bpdf-spec/. The online forum at https://baseprints.singlesource.pub is available for discussions related to Baseprint topics and specifications.

Interoperability

Websites such as https://lens.perm.pub and https://pilot.perm.pub use free open-source software, like the Python package Epijats, to render Baseprint document snapshots into HTML pages and PDF files.

This specification is for interoperability with reference software implementations. As of 2024, the only reference implementation is the Python package Epijats. For this edition of this specification, version 2.0 of Epijats is the reference software. Epijats is used by the authoring software Baseprinter, the Single-Page Application (SPA) BaseprintLens, and the website generation software BaseprintPress.

Snapshots vs. Successions

This specification of the Baseprint Document Format (BpDF) is for Baseprint document *snapshots* rather than Baseprint document *successions*. Snapshots are archived and redistributed as part of a *Baseprint document successions* whose digital encoding format is specified by the separate specification, the Document Succession Git Layout (DSGL). While snapshots are immutable, successions, in contrast, can be amended. The SWHID for a snapshot is distinct from the Document Succession Identifier (DSI) used to identify a succession.

Document Dates

Like JATS XML Article Authoring Tag Set [1], BpDF does not store a date for a document. Instead, dates for a Baseprint document are recorded in the Git commit records of the Document Succession Git Layout (DSGL) [2]. These dates are set when an author amends a succession in DSGL with a snapshot as a new edition.

JATS XML in a Directory

Technically, BpDF is not a file format but rather a format for a directory-like data structure. This structure is addressable as a SWHID version 1 directory (equivalent to a Git tree).

When generating BpDF data, it is temporarily stored in a file system directory. However, for long-term public storage, BpDF data is preserved in a SWHID addressable directory in the Software Heritage Archive (or an equivalent tree in a Git repository).

At the top level of a BpDF directory, a file named `article.xml` is encoded in a subset of the JATS Article Authoring Tag Set [1] and is inspired and influenced by JATS4R [3,4]. Most of BpDF is a specification of this JATS XML file format, which will be referred to as *Baseprint JATS XML*.

BpDF differs from the Manuscript Exchange Common Approach (MECA) in that a BpDF snapshot is automatically rendered into HTML pages and PDF files, and is not designed for a non-automated publishing process.

Restyling to JATS4R

Since the Baseprint JATS XML format is a small subset of JATS and is not intended for real journal articles, a Baseprint JATS XML file can be restyled by adding fictitious data to conform to the XML schema of a JATS4R validator or the PMC Style Checker.

The source code repository for Epijats includes an XSLT file for such restyling. Some of the information that must be added is fictitious, such as journal title. This restyling is for testing and facilitating possible interoperability with other JATS systems.

Notable Features/Limitations

XML elements for images and mathematics are absent from this edition of Baseprint JATS. These important features of JATS are planned for a future edition.

Citations and references in Baseprint JATS XML are styled by the application software that generates HTML pages and/or PDF files. Authors do not control the citation styling. References are in `<element-citation>` elements and not `<mixed-citation>`. Furthermore, there is no `publication-type` attribute. Depending on the bibliographic fields present inside the `<element-citation>`, the application software must choose appropriate styling. If the `<element-citation>` "looks" like a journal reference, then the software should style it like a journal reference. This means it is challenging for rendering software to exactly match popular citation styles. Software can approximate these styles or rely on services like Crossref to get addition reference metadata absent from the document snapshot data.

Another notable restriction is `<xref ref-type="bibr">` XML elements being in top-level `<sup>` elements, which are interpreted to have the semantic meaning of a group of citations to be styled together (e.g., `[7,11]`), not necessarily superscripted text.

Formal Specification

Terminology

Element

In this specification, the term “element” means a specific XML element within an XML document parse tree. The notation of `<foobar>` may refer to an XML tag or elements with that tag, depending on context. When an element “has a tag” it never refers a child element.

Criterion

The formal part of this specification is defined in terms of *criteria* and does not prescribe what criteria XML files must or should satisfy. However, to accurately claim that an XML file “satisfies all the criteria” of this specification, the XML file MUST satisfy all the criteria of this specification.

Each formal criterion is a true or false statement for a given XML file. Each criterion is documented to facilitate communication about which criteria might not be satisfied in particular contexts. Depending on the context, it might or might not make sense to satisfy specific criteria. In general, the more criteria that are satisfied by an XML file, the higher the level of interoperability it will achieve with the reference software of this specification.

Constraints

Due to the complexities of where XML elements can appear within an XML document tree, some criteria are specified in terms of *constraints*. XML elements in a document may have zero, one, or more constraints. As an example, consider the following JATS paragraph:

```
<p>
    <sup>
        <xref rid="definition-e">
            e<sup>x</sup>
        </xref>
    </sup>
    <sup>
         <xref rid="r1" rid-type="bibr">1</xref>
    </sup>
</p>
```

The above example satisfies the criteria of this specification because constraints can be assigned in the following manner:

- the first `<sup>` has constraint `$HYPERTEXT`,
- the second `<sup>` (nested inside the first) has constraint `$HYPOTEXT` (in addition to `$HYPERTEXT`), and
- the last `<sup>` has constraint `$CITATION`.

Definitions/Symbols

Definition: Criteria of this specification imply XML elements in an XML document tree must have zero, one, or more of the following *constraints*:

```
$CITATION
$HYPERTEXT
$HYPOTEXT
$P_CHILD
$P_LEVEL
```

Definition: Symbol {TYPO_TAG} means any one of the following tags:

<bold>
<italic>
<monospace>
<sub>
<sup>

Snapshot Directory Encoding

Criterion: The directory is encoded such that its computed hash interoperates with Git software as a Git tree hash.

Criterion: The directory is encoded such that its computed hash interoperates with the hash following the `swh:1:dir:` prefix of a SWHID (SoftWare Hash IDentifier).

Criterion: The directory is encoded such that its computed hash interoperates with Git software as a Git tree hash.

Criterion: There is only one file in the directory and its filename is `article.xml`. This file is in the Baseprint JATS XML format described in this specification.

Criterion: The tree (directory) entry for `article.xml` has normal file mode in Git and does not have the executable bit set.

XML Basics

Criterion: "Well-formed" per the XML 1.0 W3C recommendation.

Criterion: No dependency on any external XML DTD (not even a dependency to an official JATS DTD).

Criterion: The following elements contain only optional whitespace between start tag, any child elements, and end tag:

- <article-meta>
- <article>
- <back>
- <contrib-group>
- <contrib>
- <date-in-citation>
- <disp-quote>
- <element-citation>
- <front>
- <license>
- <permissions>
- <person-group>
- <ref-list>
- <ref>
- <sec>

- <table-wrap>
- <table>
- <tbody>
- <thead>
- <title-group>
- <tr>

Minimal attributes

Criterion: The following elements have no attributes:

- <abstract>
- <article-meta>
- <back>
- <body>
- <bold>
- <code>
- <comment>
- <contrib-group>
- <copyright-statement>
- <def-item>
- <def-list>
- <disp-quote>
- <element-citation>
- <fpage>
- <front>
- <isbn>
- <issn>
- <issue>
- <italic>
- <license-p>
- <license>
- <list-item>
- <lpage>
- <monospace>
- <name>

- <permissions>
- <preformat>
- <publisher-loc>
- <publisher-name>
- <ref-list>
- <string-name>
- <sub>
- <sup>
- <table-wrap>
- <table>
- <tbody>
- <thead>
- <title-group>
- <tr>
- <uri>
- <volume>

Criterion: The elements with the following tags only have the following possible attributes:

Tag	Possible Attributes
<article>	lang=
<contrib>	contrib-type= id=
<date-in-citation>	content-type=
<ext-link>	href= ext-link-type=
<license_ref>	content-type=
<list>	list-type=
<person-group>	person-group-type=
<pub-id>	pub-id-type=
<sec>	id=
<td>	align=
<th>	align=

<article> element

Criterion: <article> is the root element of the XML document.

Criterion: The <article> attribute lang= (if present) has value "en".

Criterion: <article> has a sequence of child elements matching regular expression:

(<front>) (<body>) (<back>)?

<front> element tree

Criterion: <front> has exactly one child element <article-meta>.

Criterion: <article-meta> has a sequence of child elements matching regular expression:

(<title-group>) (<contrib-group>) (<permissions>)? (<abstract>)

Criterion: <title-group> has exactly one child element <article-title>.

Criterion: <article-title> has constraint $HYPERTEXT.

Criterion: <contrib-group> has only child elements with tag <contrib>.

Criterion: <contrib> elements have an attribute of contrib-type= and it has value the "author".

Criterion: <contrib> elements have only child elements with a tag of any one of:

- <name> (exactly one)
- <contrib-id> (zero or one)
- <email> (zero or one)

Criterion: <name> has child elements <surname> and/or <given-names> (zero or one of each).

Criterion: <surname> and <given-names> have string content with no child elements.

Criterion: <contrib-id> has exactly one attribute with value contrib-id-type="orcid".

Criterion: <contrib-id> has just string content of a valid an ORCID including the https://orcid.org/ prefix.

Criterion: <permissions> has only child elements <copyright-statement> and <license> (zero or one each).

Criterion: <copyright-statement> has constraint $HYPERTEXT.

Criterion: <license> has only child elements <license-p> and <license_ref>.

Criterion: <license-p> has constraint $HYPERTEXT.

Criterion: <license_ref> tags are in the XML namespace:

"http://www.niso.org/schemas/ali/1.0/".

Criterion <license_ref> element content is just a string (URL).

Criterion: Attributes values of content-type= of element <license_ref> are any one of the following:

```
"cc0license"
"ccbylicense"
"ccbysalicense"
"ccbynclicense"
"ccbyncsalicense"
"ccbyndlicense"
"ccbyncndlicense"
```

Criterion: If the non-whitespace string contents of <license_ref> has one of the following prefixes:

```
"https://creativecommons.org/publicdomain/zero/"
'https://creativecommons.org/licenses/by/"
'https://creativecommons.org/licenses/by-sa/"
'https://creativecommons.org/licenses/by-nc/"
'https://creativecommons.org/licenses/by-nc-sa/"
'https://creativecommons.org/licenses/by-nd/"
'https://creativecommons.org/licenses/by-nc-nd/"
```

then the `content-type=` value, if present, must equal the corresponding respective value:

```
"cc0license"
"ccbylicense"
"ccbysalicense"
"ccbynclicense"
"ccbyncsalicense"
"ccbyndlicense"
"ccbyncndlicense"
```

<body> element tree

Criterion: Elements <body> and <abstract> contain a sequence of child elements with tags matching the regular expression:

```
(<p>)* (<sec>)*
```

Criterion: <sec> elements contain a sequence of child elements with tags and constraints matching the regular expression:

```
(<title>)? ($P_LEVEL)* (<sec>)*
```

Definition: Elements with constraint $P_LEVEL have any one of the following tags:

```
<code>
<disp-quote>
<list>
<p>
<preformat>
<table-wrap>
```

<back> element tree

Criterion <back> has exactly one child element <ref-list>.

Criterion: The <ref-list> element contain a sequence of child elements with tags matching the regular expression:

```
(<title>)? (<ref>)*
```

Criterion: <ref> elements have one attribute and it is id=.

Criterion: <ref> elements have exactly one child element of <element-citation>.

Criterion: <element-citation> elements only have child elements with any one the following tags:

<article-title>
<comment>
<date-in-citation>
<edition>
<fpage>
<isbn>
<issn>
<issue>
<lpage>
<person-group>
<pub-id>
<publisher-loc>
<publisher-name>
<source>
<uri>
<volume>

Criterion: For <element-citation> elements, there are one or zero child elements for each possible tag, with the exception of <pub-id>, which can appear more than once.

Criterion: For <element-citation> elements, all child elements with tag <pub-id> have different values for attribute pub-id-type=.

Criterion: <person-group> elements have an attribute person-group-type= with either value "author" or "editor".

Criterion: <person-group> elements have zero or more child elements with either tag <name> or <string-name>.

Criterion: The following elements have string content only:

<comment>
<fpage>
<isbn>
<issn>
<issue>
<lpage>
<publisher-loc>
<publisher-name>
<string-name>
<uri>
<volume>

Criterion: <date-in-citation> attribute content-type= equals value "access-date".

Criterion: <date-in-citation> have all child elements with tag <year>, <month>, or <day> and at most once for each tag.

Criterion: <date-in-citation> elements contain child element <year>.

Criterion: <date-in-citation> has child element <month> only if <year> is also present.

Criterion: <date-in-citation> has child element <day> only if <month> is also present.

Criterion: <edition> elements have just an integer as content and do not have any non-digit characters.

Criterion: <pub-id> pub-id-type= attributes have values "doi" or "pmid".

Criterion: <pub-id> elements with attribute value pub-id-type="doi" have string content that starts with “10.” and not “http”.

HTML-like content

Criterion: Elements with constraint $HYPOTEXT have tag {TYPO_TAG}.

Criterion: Elements with constraint $HYPOTEXT have all child elements also with constraint $HYPOTEXT.

Hypertext elements

Criterion: Elements with constraint $HYPERTEXT have any one of the following tags

```
<ext-link>
<xref>
{TYPO_TAG}
```

Criterion: Elements with constraint $HYPERTEXT and {TYPO_TAG} have all child elements with constraint $HYPERTEXT.

Criterion: Elements with tag <ext-link> have all child elements with constraint $HYPOTEXT.

Criterion: Every <ext-link> has an href= attribute from the XML namespace http://www.w3.org/1999/xlink.

Criterion: Every <ext-link> attribute ext-link-type= takes the value "uri" (if present).

Criterion: Elements with tag <xref> have all child elements with constraint $HYPOTEXT.

Criterion: Every element with tag <xref> and constraint $HYPERTEXT has an attribute rid=.

Criterion: Every element with tag <xref> and constraint $HYPERTEXT has no attribute other than rid=.

Paragraph elements

Criterion: Elements with constraint $P_CHILD have one of the following tags

```
<code>
<def-list>
<disp-quote>
<ext-link>
<list>
<preformat>
<xref>
{TYPO_TAG}
```

Criterion: <p> elements have all child elements with constraint $P_CHILD.

Criterion: Elements with constraint $P_CHILD and {TYPO_TAG} but not tag <sup> have constraint $HYPERTEXT.

Criterion: Elements with constraint $P_CHILD and tag <sup> have constraint $HYPERTEXT or $CITATION (but not both).

Citation elements

Criterion: Elements with constraint $CITATION and tag <sup> have text content of:

- optional whitespace before the first child element
- optional whitespace after the last child element, and
- a comma and optional whitespace between child elements.

Criterion: Elements with constraint $CITATION have all child elements with constraint $CITATION and tag <xref>.

Criterion: Elements with constraint $CITATION and tag <xref> have an attributes of ref-type= with value bibr.

Criterion: Elements with constraint $CITATION and tag <xref> have exactly two attributes of rid= and ref-type=.

Criterion: Elements with constraint $CITATION and tag <xref> have a value for rid= that matches the value of attribute id= in an element with tag <ref>.

Criterion: Elements with constraint $CITATION and tag <xref> have content of only a single integer, surrounded by optional whitespace, and no child elements. The integer corresponds to the position in <ref-list> of the <ref> element with an id= attribute value that equals the rid= attribute of the <xref> element.

List elements

Criterion: <list> element attributes list-type= have either value "bullet" or "order".

Criterion: <list> elements only have child elements with tag <list-item>.

Criterion: <list-item> elements only have child elements with either tag <p> or <list>.

Criterion: <def-list> elements only have child elements with tag <def-item>.

Criterion: <def-list> elements only have child elements with either tag <term> or <def>.

Criterion: <term> elements only have child elements with {TYPO_TAG} or {LINK_TAG}.

Criterion: <term> elements have all child elements with constraint $HYPERTEXT.

Criterion: <def> elements only have child elements with tag <p>.

Table elements

Criterion: <table-wrap> elements contains a single <table> element.

Criterion: <table> child elements are <thead> or <tbody> elements.

Criterion: <thead> and <tbody> child elements are <tr> elements.

Criterion: <tr> child elements are <th> or <td> elements.

Criterion: <th> and <td> elements have all child elements with constraint $P_CHILD.

Criterion: <th> and <td> attributes align= have values "left", "center", or "right".

Other elements

Criterion: <disp-quote> elements only contain child elements with tag <p>.

Criterion: <code> and <preformat> elements have all child elements with constraint $HYPERTEXT.

Interoperability Issues

- HTML <table> has <thead> before <tbody>
- HTML <table> probably (?) has exactly one <tbody> and at most one <thead>

- `<xref>` attribute `rid=` values that might not match an `id=` attribute of an element converted to HTML with the same `id=` (e.g., `<sec>`, `<ref>`).

References

1. U.S. National Library of Medicine (NLM). *JATS: Article Authoring Tag Set*. 2024, https://jats.nlm.nih.gov/articleauthoring/1.4/.
2. Ellerman, E. Castedo. *Document Succession Git Layout (DSGL)*. 2024, https://perm.pub/VGajCjaNP1Ugz58Khn1JWOEdMZ8.
3. Maloney, Chris, Alf Eaton, and Jeff Beck. "A client-side JATS4R validator using saxon-CE". *Balisage: The Markup Conference*, vol. 15, 2015, https://doi.org/10.4242/BalisageVol15.Beck01.
4. Beck, Jeffrey, Melissa Harrison, Stephen Laverick, Kevin Lawson, Kelly McDougall, Mary Seligy, and Lucie Senn. "What JATS4R can achieve, with a little help from its friends". *Journal Article Tag Suite Conference (JATS-Con)*, 2019, https://www.ncbi.nlm.nih.gov/books/NBK540949/.