Edition 1.1.1

Author date: 2025-08-25

Archive date: 2025-08-25

Citation:
E. Castedo Ellerman (20??) "Baseprint Document Format (BpDF)" https://perm.pub/DPRkAz3vwSj85mBCgG49DeyndaE/1.1.1

Baseprint Document Format (BpDF)

E. Castedo Ellerman (castedo@castedo.com)

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 small subset of JATS XML [1] 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 also 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 2.2 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 [2], 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) [3]. 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 JATS XML [2] and is inspired and influenced by JATS4R [4,5]. 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

Tables, math, images, and footnotes

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

Citation style

Citations and references in Baseprint JATS XML are styled by viewer 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>`, viewer software must choose appropriate styling. If the `<element-citation>` “looks” like a journal reference, then viewer software should style it like a journal reference. This means it is challenging for viewer software to exactly match popular citation styles. Software can approximate these styles or rely on services like Crossref to get additional reference metadata absent from the document snapshot data.

Citation elements

Another notable restriction is `<xref ref-type="bibr">` XML elements being inside 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]`), and not necessarily superscripted text.

External Metadata

Metadata in a typical JATS XML document is sourced from both authors and journal publishers. In this most common scenario, JATS XML serves as a vehicle for multiple sources of metadata. Baseprint JATS XML differs from typical JATS XML documents in two ways:

1. a Baseprint document is designed for self-archiving/self-publishing by authors, and
2. Baseprint JATS XML is contained within an *immutable* document *snapshot*.

With respect to a document snapshot, some metadata is *internal*, to be included in Baseprint JATS XML, while other metadata is *external* and intentionally not included. For example, the JATS element `/article/front/article-meta/title-group/article-title` is internal metadata that is sourced from an author and thus appropriately included in Baseprint JATS XML. In contrast, the JATS element `/article/front/article-meta/history` is external metadata, and does not make sense to store inside an immutable Baseprint document snapshot.

Formal Specification

Terminology

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. 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.

Element

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

Content

The contents of XML elements fall into four categories:

empty
content of an empty XML element (e.g., `<break/>`)

text-only
content of non-whitespace text with no child elements

element-only
content of only child elements (and optional whitespace text)

mixed

content of both text and child elements

Whitespace is in the narrow sense of the ASCII characters tab (9), linefeed (10), vertical tab (11), formfeed (12), carriage return (13), and space (32).

Element Varieties

Some XML elements with the same tag have differing semantics depending on their location within an XML document tree. For this reason, some criteria in this specification are specified in terms of *element varieties*. Specifically, the elements <bold>, <italic>, <monospace>, <p>, <sub>, <sup>, and <xref> have multiple *varieties*. For XML documents that satisfy the criteria of this specification, these elements will unambiguously belong to exactly one of their varieties.

Elements with the following tags may be of the following varieties based on the criteria of this specification.

```
ELEMENT TAG   ELEMENT VARIETIES
-----------   -----------------
<bold>        ~HYPER    ~HYPO
<italic>      ~HYPER    ~HYPO
<monospace>   ~HYPER    ~HYPO
<p>           ~HTML     ~WRAPPER
<sub>         ~HYPER    ~HYPO
<sup>         ~HYPER    ~HYPO  ~CITE
<xref>        ~DEFAULT  ~CITE
```

The notation <foo>~BAR is used to denote a <foo> element of the ~BAR variety. The notation of <foo> without any variety means a <foo> element of *any* variety.

In theory, element varieties could be made unnecessary by using different XML tags. But due to backward compatibility with archived JATS XML files, changing XML tags is not an option.

As an example, consider the <sup> elements of 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>
```

Based on the criteria of this specification, the <sup> elements are of the following varieties:

- the first <sup> is of the variety <sup>~HYPER,
- the second <sup> (nested inside the first) is of the variety <sup>~HYPO, and
- the last <sup> is of the variety <sup>~CITE.

Snapshot Directory Encoding

Criterion #14435: The directory is encoded such that its computed hash interoperates with Git software as a Git tree hash.

Criterion #16289: 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 #12743: 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 #14763: The directory (Git tree) entry for `article.xml` has a normal file mode in Git and does not have the executable bit set.

XML Basics

Criterion #15719: The file `article.xml` is "well-formed" per the XML 1.0 W3C recommendation.

Criterion #13799: There is no parsing dependency on any external XML DTD (not even a dependency on an official JATS DTD).

Criterion #10192: The XML prefix `ali` is used for any and all elements and attributes using the XML namespace `http://www.niso.org/schemas/ali/1.0/` by relying on the declaration

```
xmlns:ali="http://www.niso.org/schemas/ali/1.0/"
```

Note: A similar restriction is specified by NISO JATS [2].

Criterion #11855: The XML prefix `xlink` is used for any and all elements and attributes using the XML namespace `http://www.w3.org/1999/xlink` by relying on the declaration

```
xmlns:xlink="http://www.w3.org/1999/xlink"
```

Note: A similar restriction is specified by NISO JATS [2].

HTML-like content

Hypertext

Definition: The element set {HYPERTEXT} consists of the elements:

```
<bold>
<ext-link>
<italic>
<monospace>
<sub>
<sup>
<xref>~DEFAULT
```

Criterion #19521: The following elements contain mixed content with all child elements from the set {HYPERTEXT}:

```
<bold>~HYPER
<italic>~HYPER
<monospace>~HYPER
<sub>~HYPER
<sup>~HYPER
```

Criterion #18455: The following elements do not have any attributes:

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

Hypotext

Definition: The element set {HYPOTEXT} consists of the elements:

<bold>~HYPO
<italic>~HYPO
<monospace>~HYPO
<sub>~HYPO
<sup>~HYPO

Criterion #16382: Elements from the set {HYPOTEXT} contain mixed content with all child elements from the set {HYPOTEXT}.

Hyperlinking elements

<ext-link>

Criterion #13099: <ext-link> has an xlink:href= attribute with a URL as an attribute value.

Criterion #14614: Every <ext-link> attribute ext-link-type= takes the value "uri" (if present).

Criterion #17431: <ext-link> has no attributes other than xlink:href= and (optional) ext-link-type=.

Criterion #19236: <ext-link> contains mixed content with all child elements from the set {HYPOTEXT}.

<xref>~DEFAULT

Criterion #17683: <xref>~DEFAULT elements have exactly one attribute, and it is rid=.

Criterion #12342: <xref>~DEFAULT contains mixed content with all child elements from the set {HYPOTEXT}.

Other elements

<break>

Criterion #12430: <break> elements have no attributes and contain empty content.

<code>

Criterion #13634: <code> elements have no attributes.

Criterion #15943: <code> elements contain mixed content with child elements from the set {HYPERTEXT}.

<p>~HTML and <p>~WRAPPER (paragraph)

<p> elements in JATS may contain block elements (e.g., <list>). This is inconsistent with HTML where <p> contains only “phrasing content”, which does not allow block elements. In Baseprint XML, the element varieties <p>~HTML and <p>~WRAPPER distinguish between <p> elements like HTML, which may only contain phrasing content, and <p> elements that contain block elements.

Criterion #13912: <p>~HTML elements have no attributes.

Criterion #14762: <p>~HTML elements contain mixed content with all child elements from the set {HYPERTEXT}.

Criterion #16648: <p>~WRAPPER elements have exactly one attribute, and it is specific-use="wrapper".

Criterion #17818: <p>~WRAPPER elements contain element-only content with only child elements from the set:

<code>
<def-list>
<disp-quote>
<list>
<preformat>

<preformat>

Criterion #10279: <preformat> elements have no attributes.

Criterion #16819: <preformat> elements contain mixed content with child elements from the set {HYPERTEXT}.

List elements

<list>

Criterion #14304: <list> elements have no attributes or only an attribute of list-type=.

Criterion #17495: <list> element attributes of list-type= have the value "bullet" or "order".

Criterion #13090: <list> elements contain element-only content with only <list-item> child elements.

<list-item>

Criterion #18148: <list-item> elements have no attributes.

Criterion #12420: <list-item> elements contain element-only content with only child elements from the set:

<p>
<list>
<def-list>

<def-list>

Criterion #18543: <def-list> elements have no attributes.

Criterion #14530: <def-list> elements contain element-only content with only <def-item> child elements.

<def-item>

Criterion #13583: <def-item> elements have no attributes.

Criterion #10045: <def-item> elements contain element-only content with child elements of either <term> or <def>.

<term>

Criterion #11829: <term> elements have no attributes.

Criterion #13735: <term> elements contain mixed content with child elements from the set {HYPERTEXT}.

<def>

Criterion #14358: <def> elements have no attributes.

Criterion #15807: <def> elements contain element-only content with only <p> child elements.

High-level structural elements

Highest-level elements

<article>

Criterion #15199: <article> is the root element of the XML document.

Criterion #10864: <article> has no attributes (apart from pseudo-attributes for XML namespaces).

Criterion #16641: <article> contains element-only content with a sequence of child elements matching the regular expression:

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

<front>

Criterion #14001: <front> has no attributes.

Criterion #12640: <front> contains element-only content with exactly one child element <article-meta>.

<article-meta> element tree

Criterion #13284: <article-meta> has no attributes.

Criterion #11553: <article-meta> contains element-only content with a sequence of child elements matching the regular expression:

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

<back> element tree

Criterion #11019: <back> has no attributes.

Criterion #18947: <back> contains element-only content with exactly one child element <ref-list>.

Content elements

Definition: {P_LEVEL} denotes the set of elements:

```
<code>
<def-list>
<disp-quote>
<list>
<p>~HTML
<preformat>
```

<disp-quote>

Criterion #18135: <disp-quote> elements have no attributes.

Criterion #18442: <disp-quote> elements contain element-only content with only <p> child elements.

<abstract>

Criterion #14631: <abstract> has no attributes.

Criterion #10926: <abstract> contains element-only content with a sequence of child elements matching the regular expression:

```
({P_LEVEL})* (<sec>)*
```

<body>

Criterion #19029: <body> has no attributes.

Criterion #18521: <body> contains element-only content with a sequence of child elements matching the regular expression:

```
({P_LEVEL})* (<sec>)*
```

<sec>

Criterion #12620: <sec> elements have no attributes or an id= attribute.

Criterion #18933: <sec> elements contain element-only content with a sequence of child elements matching the regular expression:

```
(<title>)? ({P_LEVEL})* (<sec>)*
```

<title>

Criterion #15129: <title> elements have no attributes.

Criterion #16981: <title> elements contain mixed content with each child element either <break> or from the set {HYPERTEXT}.

Metadata elements

Article title

<title-group>

Criterion #15574: <title-group> has no attributes.

Criterion #19365: <title-group> contains element-only content with exactly one child element <article-title>.

<article-title>

Criterion #17019: <article-title> has no attributes.

Criterion #16217: <article-title> contains mixed content with child elements of set {HYPER-TEXT}.

Contributors

<contrib-group>

Criterion #10923: <contrib-group> has no attributes.

Criterion #17698: <contrib-group> contains element-only content with only child elements of <contrib>.

<contrib>

Criterion #17181: <contrib> elements have exactly one attribute of contrib-type= with value "author".

Criterion #19818: <contrib> elements contain element-only content with child elements:

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

<name>

Criterion #15691: <name> has no attributes.

Criterion #12424: <name> contains element-only content with child elements from the set:

<surname>
<given-names>
<suffix>

and at most one child element for each tag.

<surname>, <given-names>, and <suffix>

Criterion #17569: <surname>, <given-names>, and <suffix> have no attributes.

Criterion #17289: <surname>, <given-names>, and <suffix> contain text-only content.

<contrib-id>

Criterion #13828: <contrib-id> has exactly one attribute with value contrib-id-type="orcid".

Criterion #12150: <contrib-id> contains text-only content of a valid ORCID including the https://orcid.org/ prefix.

Permissions and licensing

<permissions>

Criterion #19885: <permissions> has no attributes.

Criterion #11010: <permissions> contains element-only content of only child elements <copyright-statement> and <license> (zero or one of each).

<copyright-statement>

Criterion #13932: <copyright-statement> has no attributes.

Criterion #13317: <copyright-statement> contains mixed content with child elements of set {HYPERTEXT}.

<license>

Criterion #19618: <license> has no attributes.

Criterion #19475: <license> contains element-only content with child elements <license-p> and/or <ali:license_ref>.

<license-p>

Criterion #10671: <license-p> has no attributes.

Criterion #11028: <license-p> contains mixed content with child elements of set {HYPERTEXT}.

<ali:license_ref>

Criterion #16170: <ali:license_ref> contains text-only content of a URL.

Criterion #16811: <ali:license_ref> has no attribute or an attribute of content-type= with any one of the following values:

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

Criterion #11510: If the non-whitespace text contents of <ali:license_ref> have 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"
```

Bibliographic elements

Citation

<xref>~CITE

Criterion #14740: <xref>~CITE elements have exactly two attributes: rid= and ref-type=.

Criterion #11027: <xref>~CITE elements have an attribute of ref-type= with the value bibr.

Criterion #12086: <xref>~CITE elements have a value for rid= that matches the value of the attribute id= of a <ref> element.

Criterion #10484: <xref>~CITE elements contain text-only content of a single integer (surrounded by optional whitespace). The integer corresponds to the ordered position in <ref-list> of the <ref> element with an id= attribute value matching the rid= attribute value of the <xref> element.

<sup>~CITE

Criterion #14278: <sup>~CITE elements only have child elements of <xref>~CITE.

Criterion #12352: <sup>~CITE elements have mixed content with text-only 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.

Bibliography

<ref-list>

Criterion #14165: `<ref-list>` has no attributes.

Criterion #12136: `<ref-list>` contains element-only content with a sequence of child elements matching the regular expression:

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

<ref>

Criterion #18652: `<ref>` elements have one attribute, and it is `id=`.

Criterion #15949: `<ref>` contains element-only content with exactly one child element of `<element-citation>`.

<element-citation>

Criterion #15660: `<element-citation>` elements have no attributes.

Criterion #14559: `<element-citation>` contains element-only content with child elements from the set:

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

Criterion #12492: 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 #13786: For `<element-citation>` elements, all child elements with the tag `<pub-id>` have different values for the attribute `pub-id-type=`.

Criterion #18428: The following elements under `<element-citation>` have no attributes and contain text-only content:

<comment>
<elocation-id>
<fpage>
<isbn>
<issn>
<issue>
<lpage>
<publisher-loc>
<publisher-name>
<source>
<uri>
<volume>

Criterion #10807: <article-title> elements, *when under* <element-citation>, contain text-only content.

Note: Criterion #10807 does not apply to <article-title> under <title-group>.

<person-group>

Criterion #18377: <person-group> elements have exactly one attribute, person-group-type=, with either the value "author" or "editor".

Criterion #17091: <person-group> contains element-only content with child elements from the set:

<name>
<string-name>
<etal>

<string-name>

Criterion #18187: <string-name> elements have no attributes and contain text-only content.

<etal>

Criterion #16837: <etal> elements have no attributes and contain empty content.

Criterion #14180: <person-group> elements have no more than one <etal/> child element.

<year>, <month>, and <day> elements

Criterion #13721: <year>, <month>, and <day> elements have no attributes.

Criterion #17289: <year>, <month>, and <day> elements contain just an integer as content and do not have any non-digit characters.

Criterion #10430: Child elements <year>, <month>, and <day> appear at most once under their parent element.

Criterion #14321: Child element <month> appears only if <year> is present as a sibling element.

Criterion #19206: Child element <day> appears only if <month> is present as a sibling element.

<date-in-citation>

Criterion #13166: <date-in-citation> has exactly one attribute, it is content-type=, and its value is "access-date".

Criterion #11337: <date-in-citation> contains element-only content with child elements from the set:

```
<year>
<month>
<day>
```

<edition>

Criterion #18615: `<edition>` elements have no attributes.

Criterion #11753: `<edition>` elements contain just an integer as content and do not have any non-digit characters.

<pub-id>

Criterion #14308: `<pub-id>` elements have exactly one attribute, it is `pub-id-type=`, and it has the value `"doi"` or `"pmid"`.

Criterion #15283: `<pub-id>` elements with the attribute value `pub-id-type="doi"` have text-only content that starts with “10.” and not “http”.

Criterion #10955: `<pub-id>` elements with the attribute value `pub-id-type="pmid"` have text-only content of a valid PubMed Identification Number.

References

1. U.S. National Library of Medicine (NLM). *Journal Article Tag Suite*. 2024, https://jats.nlm.nih.gov/.
2. U.S. National Library of Medicine (NLM). *JATS: Article Authoring Tag Set*. 2024, https://jats.nlm.nih.gov/articleauthoring/1.4/.
3. Ellerman, E. Castedo. *Document Succession Git Layout (DSGL)*. 2024, https://perm.pub/VGajCjaNP1Ugz58Khn1JWOEdMZ8.
4. 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.
5. 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/.