perm.pub/dsi:HKSI5NPzMFmgRlb4Vboi71OTKYo/1.2
Additional formats and editions available online.

Edition 1.2

Author date: 2023-05-16

Archive date: 2023-10-08

Citation:
E. Castedo Ellerman (2023) "What is a baseprint?" perm.pub https://perm.pub/dsi:HKSI5NPzMFmgRlb4Vboi71OTKYo/1.2

What is a baseprint?

E. Castedo Ellerman (castedo@castedo.com)

Abstract

A baseprint is a digitally encoded document created to disseminate research across independent websites in both PDF and HTML formats. Compared to PDF files, baseprints offer advantages for research communication by combining features of preprints and blog posts. A preliminary type of baseprint is based on JATS4R (JATS XML) and Software Heritage IDs. However, as of 2022, the technology for working with baseprints is new and still under development, and the tools for working with baseprints are not yet user-friendly.

AUDIENCE: Developers and early adopters of tools and services for research communication.

STAGE: Edition 2 planned. Feedback welcome.

Summary

Multiple independent websites can automatically generate both a PDF file and an HTML page from a baseprint digitally encoded in JATS XML. JATS XML is often used to encode journal articles published in both PDF and HTML formats, many of which appear on both the publisher's website and PubMed Central, a US government agency website for scientific literature.

Baseprints provide features of preprints and blog posts through the generation of PDF files and HTML pages. The following table summarizes the advantages of baseprints over PDF preprints and blog posts:

Feature	PDF preprint	Blog post	JATS XML	Baseprint
Designed for different displays		Yes	Yes	Yes
Clone documents across multiple websites	Yes			Yes
Document metadata like journal articles			Yes	Yes
Target of intrinsic identifier				Yes

PDF, HTML, and the format of record

Journals present research in both PDF and HTML formats, a benefit that is absent or very limited in preprint servers. Baseprints can also be used to present research in both formats, as they are rendered into both.

Millions of journal articles are encoded in JATS XML and rendered into HTML web pages at PubMed Central. Some journals render JATS XML into both PDF and HTML. PDF and HTML are presentation formats, whereas a *format of record* [1] is a digital encoding that is archived and

rendered into presentation formats. Baseprints are digitally encoded in a format of record, which can be JATS XML. However, baseprints can also be encoded in multiple formats, just as spreadsheets can be encoded in multiple formats.

Choice of websites

With baseprints, readers not only choose their preferred presentation format, but also their preferred website, just like they can choose between PubMed Central and a journal website for some articles. If articles were only available on journal websites, readers would miss out on the unique features of PubMed Central. Additionally, new websites can provide features that do not yet exist. Having multiple websites also reduces the risk of research becoming unavailable.

Baseprint identity

Bibliographic references should unambiguously identify research documents. A web address (URL) is not sufficient for identifying a baseprint, as multiple websites can present a baseprint and no single website is the authoritative source. Instead of referencing a web address, baseprints are referenced with an intrinsic identifier [2], such as a Software Heritage ID (SWHID) [3], [4]. The identity of a baseprint is determined by its exact digital encoding, which is a sequence of bytes if it consists of a single file. If a baseprint consists of a directory of files, the digital encoding can be a *git tree*. A SWHID identifies a baseprint with a cryptographic hash of its digital encoding.

Another identifier of research documents is the Digital Object Identifier (DOI). Both the DOI and the SWHID are persistent identifiers, but unlike the SWHID, the DOI is an extrinsic identifier [2] that depends on a DOI registrant and publisher. A baseprint may have both a SWHID and a DOI. Another type of persistent identifier is the Digital Succession Identifier (DSI) [5]. If a baseprint has been added to a digital succession, it can be identified by either its SWHID or a DSI with an edition number. A DSI can even identify baseprints that have yet to be created.

Research document metadata

Bibliographic references encoded in a computer-understandable format are an example of research document metadata. Other examples include titles, abstracts, contributors, email addresses, contributor identifiers, inline citations, copyright, and licensing terms. This metadata helps search engines, citation analysis, article discovery tools, and other computer applications. Baseprints include research document metadata, which can be encoded in JATS XML or other formats.

Conclusion

Baseprints are digitally encoded documents that are:

- presented across multiple websites,
- rendered in both PDF and HTML formats,
- encoded and archived in a format of record such as JATS XML,
- referenced with an intrinsic identifier such as a Software Heritage ID (SWHID), and
- include research document metadata (e.g., title, abstract, contributors, email addresses, contributor identifiers, inline citations, copyright, and licensing).

Acknowledgments

The OpenAI API with GPT 3.5 performed substantial copyediting of this document.

References

1. Bazargan K. XML – why it should be the “format of record”. 2022. Available: https://web.archive.org/web/20220920180031/https://rivervalley.io/format-of-record/
2. Heritage S. Intrinsic and extrinsic identifiers. 2020. Available: https://web.archive.org/web/20221019201056/https://www.softwareheritage.org/2020/07/09/intrinsic-vs-extrinsic-identifiers/
3. Cosmo RD, Gruenpeter M, Zacchiroli S. Referencing Source Code Artifacts: A Separate Concern in Software Citation. Computing in Science & Engineering. 2020;22: 33–43. doi: 10.1109/MCSE.2019.2963148
4. Di Cosmo R, Gruenpeter M, Zacchiroli S. Identifiers for Digital Objects: the Case of Software Source Code Preservation. iPRES 2018 - 15th International Conference on Digital Preservation. Boston, United States; 2018. pp. 1–9. Available: https://hal.archives-ouvertes.fr/hal-01865790
5. Ellerman EC. Digital succession identifiers. 2022. Available: https://perm.pub/1wFGhvmv8XZfPx0O5Hya2e9AyXo/1