Information Services & Use 43 (2023) 367–372
DOI 10.3233/ISU-230205
IOS Press

NISO's Content Profile/Linked Document standard: A research communication format for today's scholarly ecosystem

Bill Kasdorf[*]
Principal, Kasdorf & Associates, LLC; Founding Partner, Publishing Technology Partners, Ann Arbor, MI, USA

Abstract. This paper is based upon the author's presentation at the 2023 NISO Plus Conference in which he discussed the purpose and importance of the NISO Content Profile/Linked Document standard. The need for the standard is obvious: users demand the delivery of contextualized, targeted content delivered as a natural part of their workflow and publishers aspire to produce machine-actionable FAIR (Findable, Accessible, Interoperable, Reusable) materials, but many publishing workflows are focused on articles, often published after research is concluded. This standard is an application of HTML5 and JSON-LD to create semantic relationships between data elements in scholarly publishing workflows and express machine actionable content, to ease reuse and interchange of scholarly research information. The format description defines a set of rules that outline the minimal characteristics of documents (Linked Documents) that conform to the standard and a mechanism to define more detailed Content Profiles that extend and refine the rules for specific use cases.

Keywords: Linked document standard, NISO CP/LD standard, HTML, JSON-LD

1. Introduction

Today's researchers require the ability to communicate their work in ways that go beyond the traditional scholarly article. While the formally published article will, of course, continue to be essential to the scholarly record, research needs to be communicated today in other forms, and more frequently than only by the article that documents its results. This needs to include not only text and image content, but also semantics, data, and other resources as well - ideally using current web technologies. And it needs to be done in arbitrarily small chunks throughout the research process in order for the research to fully participate in the broad research ecosystem, especially to meet the increasing demand for open science. NISO's new Content Profile/Linked Document (CP/LD) standard [1] is designed to enable this.

[*] E-mail: kasdorf.bill@gmail.com.

2. Standardizing journal article content: A brief history

For centuries, the scholarly article has been the canonical method for scholars and researchers to communicate their findings. And for decades, especially in science, the content of articles has been encoded in XML (and before that, in SGML, XML's progenitor). Initially, publishers, aggregators, and journal hosts used their own proprietary content models called DTDs (Document Type Definitions) [2]. It quickly became apparent that the resulting Tower of Babel was unsustainable for the effective communication of research; two seminal initiatives were launched in the early 2000s to address this.

PubMed Central (PMC), the online repository of articles in the life sciences from the U.S. National Library of Medicine (NLM), had developed a basic DTD, pmc-1.dtd, in 2000 to simplify access to full-text articles online. Realizing that a more robust model was needed to capture the full richness of the original articles, PMC contracted with Mulberry Technologies to analyze the DTDs used by major biomedical publishers and assess whether a common model could be created that could accommodate all of them [3]. Mulberry concluded that this was indeed possible and set about developing the pmc-2.dtd.

Separately, in 2001 the Harvard University Library E-Journal Archiving Project contracted with Inera to do a feasibility study of creating a DTD that could be used to archive all electronic journals (not just the biomedical journals archived by PMC). Inera analyzed the DTDs of ten major scholarly journal publishers and concluded that developing such a cross-publisher DTD was indeed possible. When Bruce Rosenblum of Inera saw the pmc-2.dtd being developed by Mulberry, he realized that it could form the basis for the master model that Inera and Harvard were looking for.

So, in 2002, Mulberry and Inera joined forces, releasing the NLM Archiving and Interchange Tag Suite in early 2003 that contained two DTDs, one designed to accommodate archiving and interchange and a closely related, but stricter version, for use by publishers. Those "NLM DTDs" became the *lingua franca* of scholarly publishing, being periodically updated and having evolved into today's JATS, the Journal Article Tag Suite, a NISO/ANSI standard first published as ANSI/NISO Z39.96-2012, JATS: Journal Article Tag Suite (version 1.0) in 2012 and currently in version 1.3 [4]. JATS XML is now used by almost all journal hosting platforms, scholarly and STM journal publishers, and the prepress and data conversion vendors that support them.

3. If we have JATS, why do we need CP/LD?

First, it is important to understand that JATS is not a rendering format. It is not intended to be "read" by people; it is intended to express the content according to a strictly-defined specification (the XML tag set) that enables it to be processed by computers and transformed as needed.

JATS is an incredibly rich and complex model, the result of many years of evolution in the context of representing many millions of journal articles, enabling very granular tagging to delineate the content of articles from any and all fields of scholarly and scientific study, as well as rich metadata to facilitate the discovery, management, and dissemination of the articles and their content. It is designed to be very flexible, offering an extensive suite of highly-structured and extensively documented elements and attributes (think of those as the nouns and adjectives) to accommodate the enormous range of content it must handle.

But in order for JATS-encoded content to be consumed, it must be rendered. The JATS XML of a given article is typically converted to HTML for online consumption or to PDF for print (including print-replica renderings online). Now, JATS can also be converted to CP/LD.

But CP/LD can express more than just the content of an article. Here's how the abstract of the CP/LD documentation from NISO describes it:

> CP/LD does not have to replace existing models used for journal articles, books, data sets, or semantic and metadata schemes. Instead, these can be transformed as needed to CP/LD, enabling the combination of arbitrary portions of content, data, semantics, and other resources from separate sources into a single, standards-based format optimized for interchange, search, and display.

Key is that CP/LD enables *just the appropriate portions* of the resources relevant to communicating the research at a given time or for a given purpose, to be packaged into *a single standards-based format* - as many portions as necessary and as many times as necessary throughout the research lifecycle. One very significant benefit of this is that the research can more easily be followed and engaged with by other researchers while the research is in progress, rather than only when the research is completed and the article documenting it is published.

It should also be noted that while the research is being done - and even when the article or articles resulting from it are written - the content is not yet in JATS. Most scholarly articles are written in Microsoft Word, with a significant number (and most articles in fields like physics and math) being written in TeX and LaTeX. The content in those formats can be expressed in CP/LD as well, before they are also converted to JATS for formal article publication.

4. What is a CP/LD document made of?

While CP/LD is designed to accommodate virtually any type of content, its initial implementation as a NISO standard is focused on scholarly articles. There are two fundamental parts to the standard, as described in the documentation of its specifications linked above.

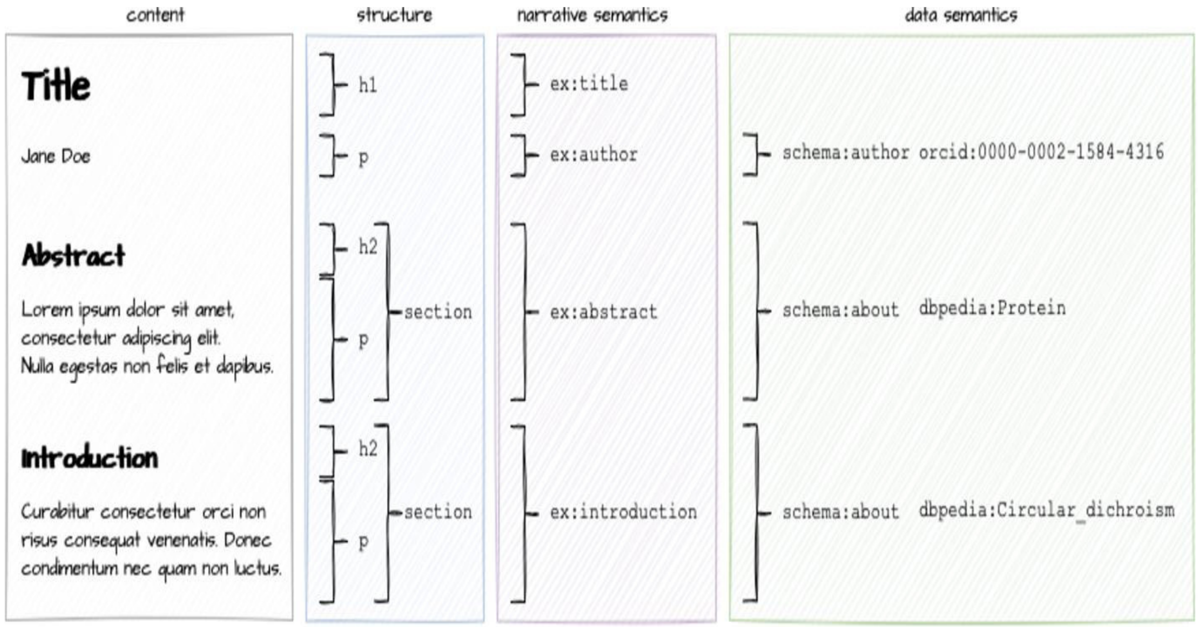
1. **Linked Documents**: a set of rules that outline the minimal characteristics of documents that conform to this standard. Such documents are called Linked Documents.

2. **Content Profiles**: a mechanism, building on the W3C Publication Manifest profile recommendation, to define more detailed requirements above and beyond the ones defined by this standard, for specific types of content and use cases.

Just as journal hosts create specifications for JATS - e.g., for Atypon, you must tag your JATS XML to the Atypon spec, and for Silverchair, you need to tag your JATS XML to the Silverchair spec, though both are fundamentally JATS - a Content Profile defines how to create Linked Documents in CP/LD for a specific purpose or for a specific kind of content. An article in CP/LD from Livermore Labs might have a different set of requirements than an article from Harvard Medical School. But both of them are still Linked Documents conforming to the CP/LD spec.

Linked Documents distinguish between content structure and the narrative and data semantics associated with that content. The content structure is expressed in HTML, the standard structural markup for the web, used both in websites and in EPUBs. The narrative and data semantics are expressed in JSON-LD, a method for encoding linked data using JSON, a format commonly used by developers. HTML and JSON-LD are both widely-used web standards, governed by the Internet Engineering Task Force (IETF) [5] and the Worldwide Web Consortium (W3C) [6].

This diagram from the CP/LD specification illustrates the relationships between these aspects of a Linked Document for a scholarly article:


content
structure
narrative semantics
data semantics
Title
Jane Doe
Abstract
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nulla egestas non felis et dapibus.
Introduction
Curabitur consectetur orci non risus consequat venenatis. Donec condimentum nec quam non luctus.
h1
p
h2
p
section
h2
p
section
ex:title
ex:author
ex:abstract
ex:introduction
schema:author orcid:0000-0002-1584-4316
schema:about dbpedia:Protein
schema:about dbpedia:Circular_dichroism

Each Linked Document must also have a unique identifier, called the document IRI (technically, an HTTPS IRI).

In addition to the required HTML and JSON-LD, a Linked Document may also have these optional components:

- Any ancillary Cascading Style Sheet [7] (CSS - which specifies the rendering of the HTML), image, or other files associated with the Linked Document (for example, data files or even software).
- Additional HTML and JSON-LD files that pertain to the same document.
- A W3C Publication Manifest [8].

The Publication Manifest is required if the Linked Document is composed of multiple parts. Like CP/LD in general, the Publication Manifest is expressed in JSON-LD. Any style, layout, or positioning must be specified by CSS (not JSON-LD) to ensure interoperability. Likewise, JSON-LD should not be used to specify structural relationships in the content; that's the job of the HTML.

A Linked Document can also include external data. A simple example is given in the diagram above: the identity of the author is specified by their ORCID ID. Importantly for open science, this can also include research data, for example by using a DataCite DOI to link to a dataset stored in a repository such as Figshare [9], Zenodo [10], Mendeley [11], or Data Dryad [12]. It can even provide provenance information via the W3C Provenance Recommendation [13]. And CP/LD enables referencing and linking to arbitrary ranges of text using the W3C Annotations Recommendation [14].

To maintain maximum flexibility, there is no formal schema for a Linked Document, although NISO provides very thorough documentation with helpful examples. And finally, Content Profiles can be created that specify stricter requirements for Linked Documents created for specific purposes or in specific contexts. The documentation of the CP/LD standard published by NISO and approved by ANSI contains an appendix that provides a Content Profile for a scholarly article.

5. CP/LD complements, rather than replaces, other formats

It is important not to think that CP/LD is a replacement for any other format or formats used in the scholarly ecosystem. It is an entirely new format, designed to work and play well with its counterparts.

It was created to address an unmet need: to be able to convey content, data, metadata, structure, and semantics in a single package. That package can be arbitrarily small or large. It is entirely based on web standards that are well-known and widely-used. It aligns with the continuing evolution of those standards. And it's expected to be able to accommodate new standards as they emerge.

It is hoped that CP/LD will become a valuable asset at all phases of the research lifecycle. The standard was open for public comment through May 2023 [15] and now is under further review.

Acknowledgements

I wish to thank Suzanne BeDell (recently retired from Elsevier), co-chair of the NISO CP/LD Working Group, for her masterful leadership - I considered myself her sidekick as co-chair. Much gratitude as well to Rinke Hoekstra from Elsevier, who brought the proposal to NISO and who provided indispensable technical leadership as we refined the specification to become a truly global, non-proprietary, web-based standard, innovative yet based on widely used technologies. Finally, thanks to all the members of the CP/LD Working Group, too numerous to cite personally here, from the following organizations: Atypon, Brown University, Crossref, Elsevier, IEEE, IOP Publishing, the Taylor & Francis Group, the University of Florida Libraries, and the University of Washington Library. It was a pleasure to work with such a stellar group, who brought a variety of essential perspectives to the work. And finally, none of this could have been accomplished without NISO's Nettie Lagace, who managed to shepherd this complex work along with all the other important standards she helps create and maintain.

About the author

Bill Kasdorf is Principal at Kasdorf & Associates, LLC, Co-Founder, Publishing Technology Partners, and Co-Chair, NISO CP/LD Working Group. He is an expert in accessibility, XML/HTML/EPUB modeling, information infrastructure, editorial and production workflows, and standards alignment to future proof content and systems. Past President of the Society for Scholarly Publishing (SSP), Bill is a recipient of SSP's Distinguished Service Award, the IDEAlliance/DEER Luminaire Award, and the Book Industry Study Group's Industry Champion Award.

Active in many standards initiatives, Bill is the Global Publishing Evangelist for the Worldwide Web Consortium (W3C); he serves on the Steering Committee of the W3C publishing activity and is a member of the W3C Publishing Working Group developing the next generation of Web Publications, as well as the W3C's EPUB 3 Community Group and the EPUB 3 CG's Accessibility Task Force. He is a member of the Book Industry Study Group (BISG), which publishes resources for the book supply chain, with a special focus on accessibility and EPUB implementation, serving on BISG's Workflow Committee.

He is a member of the editorial board of the *Learned Publishing* journal and in his consulting practice, Bill has served clients globally, including large international publishers. E-mail: kasdorf.bill@gmail.com.

References

[1] See: https://doi.org/10.3789/ansi.niso.z39.105-2023.

[2] See: https://en.wikipedia.org/wiki/Document_type_definition, accessed September 30, 2023.

[3] See J. Beck, "Report from the Field: PubMed Central, an XML-based Archive of Life Sciences Journal Articles." Presented at International Symposium on XML for the Long Haul: Issues in the Long-term Preservation of XML, Montréal, Canada, August 2, 2010, in: *Proceedings of the International Symposium on XML for the Long Haul: Issues in the Long-term Preservation of XML*, Balisage Series on Markup Technologies, vol. 6, 2010. doi:10.4242/BalisageVol6.Beck01.

[4] This history is extensively documented in Beck, Jeff, "NISO Z39.96 The Journal Article Tag Suite (JATS): What Happened to the NLM DTDs?" *Journal of Electronic Publishing* **14**(1) *Standards*, Summer 2011. doi:10.3998/3336451.0014.106.

[5] See: https://ietf.org, accessed September 30, 2023.

[6] See: https://www.w3.org, accessed September 30, 2023.

[7] See: https://www.w3schools.com/html/html_css.asp, accessed October 1, 2023.

[8] See: https://www.w3.org/TR/Pub-manifest/, accessed October 21, 2023.

[9] See: https://figshare.com, accessed October 1, 2023.

[10] See: https://zenodo.org, accessed October 1 2023.

[11] See: https://mendeley.com. accessed October 1, 2023.

[12] See: https://datadryad.org, accessed October 1, 2023.

[13] See: https://www.w3.org/TR/prov-overview, accessed October 12, 2023.

[14] See: https://devopedia.org/web-annotation, accessed October 2, 2023.

[15] See: https://www.niso.org/standards-committees/cpld, accessed October 1, 2023.