Enhancing scientific publishing: automatic conversion to JATS XML

Jertec Musap, Ljiljana

Download date / Datum preuzimanja: **2024-05-14**

Enhancing scientific publishing: automatic conversion to JATS XML

Ljiljana Jertec Musap✉

SRCE - University of Zagreb University Computing Centre, Zagreb, Croatia

ljiljana.jertec.musap@srce.hr

orcid.org/0000-0002-0059-9899

Citation

European Science Editing

Abstract

JATS XML (Journal Article Tag Suite) is an XML-based format used for publishing scholarly content. It has multiple advantages over traditional publishing methods but faces adoption challenges due to the need for relatively expensive tools and/or manual work. In 2023, the HRČAK Portal's team enabled automatic full-text conversion from DOCX to JATS XML which does not require prior knowledge of XML nor additional tools. Created JATS facilitates content and reference mining as well as transformation to HTML. It also improves cross-device compatibility and produces interactive links for an enhanced reading experience.

Introduction

HRČAK is a central portal of Croatian professional and scientific journals that serves as a publishing platform for 530+ journals with more than 287,000 published full-text open access articles. In addition to the role of the central portal that provides open access to journals from all disciplines, the service also offers technical support to journal editors and promotes good practices in scientific publishing (e.g., usage of ORCID identifiers, publishing associated datasets, and linking papers to them). It is hosted and maintained by University of Zagreb University Computing Centre (SRCE), Croatia, and has been developed in cooperation with experts in the field of information and library science and representatives of the editors of prominent Croatian journals.

Since its launch in 2006, HRČAK initially required editors to publish only PDF files along with article metadata. In 2017, support for publishing in JATS XML format was implemented and, although it remained optional, HRČAK strongly encourages its use due to the numerous advantages it offers. This viewpoint aims to discuss the characteristics of JATS XML and highlight the implementation of a new feature within HRČAK – automated conversion of DOCX documents into JATS XML.

JATS XML

JATS (Journal Article Tag Suite) is an XML-based format used for publishing of contemporary scientific content, currently well on its way to become the standard in the field of scientific publishing. It includes a suite of XML elements and attributes that describes the content and metadata of journal articles – including research and nonresearch articles, letters, editorials, and book and product reviews – with the intent of providing a common format in which publishers and archives can exchange journal content.[1]

An increasing number of journal databases, such as PubMed Central or ScieLo, now either require or highly advise that articles are prepared in the JATS XML format. This format serves multiple purposes, such as direct deposit into Crossref.[2] Plan S strongly recommends availability for download of full text for all publications (including supplementary text and data) in a machine-readable community standard format such as JATS XML.[3]

The JATS XML format offers multiple advantages for scientific publishing over traditional publishing methods, including content and reference mining, possibility to convert into various formats including HTML and PDF, and interactive on-screen representation. JATS is semantic and declarative, meaning it conveys information about the structure and semantic meaning of article components, without specifying visual interpretation or styling, which offers the potential for enhanced digital display of content.[4] Its openness, flexibility, and ease of processing have put JATS to the focus of scientific publishing.

A complete JATS XML document includes not only the article's full text but also a range of crucial metadata. The metadata serve as interpreters, ensuring that each article is found and understood by humans but also by machines that aggregate, index, and use them in other ways. These metadata include multilanguage journal- and article-level information, DOIs, the significant dates such as acceptance, revision, and publication dates; copyright details; and information about the research funding.

Although the advantages of JATS XML are well communicated with the editors, with tutorials and educational materials available, the challenge in publishing JATS XML on HRČAK database lies in the fact that the

editors take on the responsibility of creating XML files on their own. They are obliged to prepare .ZIP archives that not only contain the article's XML but also include additional files, such as images. Due to the lack of free and user-friendly solutions for creating JATS XML, the adoption of this feature in HRČAK has been relatively low, with less than 10 journals incorporating it. In practice, it has been noted that the only journals using this format regularly for HRČAK were the ones that were already preparing XML files for the PubMed Central database, which mandates the use of JATS.

It is important to acknowledge that the situation is not unique to HRČAK nor Croatian journals – according to a 2020 survey by Scholastica, despite the many potential benefits of XML, fewer than half of the 63 surveyed publishers reported producing full-text XML article files.[5] Furthermore, a 2022 survey indicated that there has been no growth at all in full-text XML article production since 2020.[6]

While considering the possible causes of this stagnation, it should be mentioned that the process of creating JATS XML included the usage of the expensive professional tools or outsourcing the process. Editors therefore had no options other than to pay a significant amount, depending on the number of issues they published each year. The alternative for the editors was to learn the syntax rules of XML and JATS and manually rewrite articles from scratch using tools like Notepad++ or Oxygen XML Editor. However, this approach was not acceptable for journals due to its time-consuming nature, the need for dedicated personnel, and the editors' inability to recognize direct XML benefits. Given these challenges, combined with the fact that JATS XML was not mandatory in HRČAK, the percentage of articles that included XML format in the database

remained relatively low. To make a shift, it was clear to SRCE team that this major obstacle could only be addressed by developing the available alternative to the existing commercial tools.

Implementation of the New JATS XML Feature in HRČAK

In 2023, the team at SRCE, responsible for the HRČAK portal, started to develop the solution for automatic conversion to JATS XML from various formats. The first format they addressed was DOCX, as it is commonly used for writing articles.

To implement this feature, the team integrated two tools into HRČAK's editorial interface. The first tool, *Pandoc.org*,[7] is responsible for converting full-text DOCX files into JATS XML. The second tool, *AnyStyle*,[8] is used to identify references and parse them into separate XML elements. While both tools functioned acceptably, some adjustments were undertaken to optimize their results. One of them was using machine learning techniques, offering the potential to enhance the performance of AnyStyle tool, especially in terms of addressing the unique characteristics of the Croatian language.

After the implementation of the new feature, all journals publishing on HRČAK can now upload their articles in DOCX format and automatically create JATS XML files, even without any prior knowledge of the standard (Figure 1). The front section of the JATS is filled with journal and article metadata that is already stored in HRČAK. The body (full-text) section of the JATS is created from the DOCX file, as well as some elements in the back section that include information about appendices or supplements. The majority of the back section consists of a list of references that are tagged using the AnyStyle *tool*.

Figure 1. The new feature in HRČAK administrative interface consisting of the DOCX upload field and the built-in XML editor.

Once created, JATS is transformed into the HTML on the HRČAK's public interface, allowing the entire article to be accessed and read without the need to download a PDF (Figure 2). This transformation also ensures compatibility with various devices and introduces interactive links (both within the text and externally) that enhance the reader's experience.

The Results and Plan for the Future
To achieve the best results from this feature, there are specific rules to follow when creating a DOCX template for articles. These rules include using the Microsoft Word styles throughout the whole text, placing the figures in the desired positions, using the Equation feature in Word for creating formulas and using standardized citation styles for referencing, such as Vancouver, APA (American Psychological Association), MLA (Modern Language Association), and more.[9]

The new JATS XML conversion feature in HRČAK was released in September 2023, and the results have been visible already in the first month. Seventeen journals out of 536 have adopted the feature, resulting in the publication of 150+ articles in JATS XML format. The HRČAK team has been receiving a lot of feedback from the journal editors, demonstrating a growing interest in the new feature as well as their growing awareness of the JATS. To support the community and facilitate the usage, HRČAK team hosts monthly consultations on JATS and is actively working on positioning the format as the future of scientific publishing. However, it is
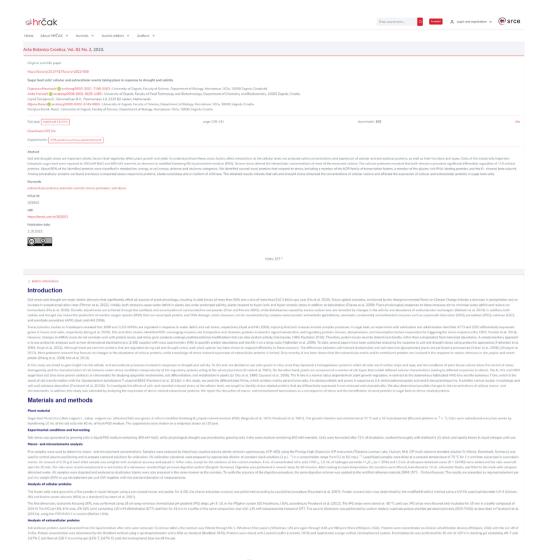
Figure 2. The example of the article on HRČAK published using JATS XML.

important to note that the access to this new feature is currently limited to the editors of journals already included in HRČAK.

To expand the impact, in 2024, HRČAK team will develop a stand-alone tool for a broader academic and scientific user base. The stand-alone tool is expected to include essential features such as a user interface with authorization and authentication functions, a form for inputting metadata, and a built-in XML editor for editing created XML files. The development is expected to be finished by the Autumn 2024 when the service for automatic conversion to JATS XML will be provided through the EOSC Marketplace,

a centralized online platform within the European Open Science Cloud (EOSC) offering access to diverse research tools, services, and datasets from various European service providers. The presence of this tool in the EOSC Marketplace will allow other European journal editors to use it as well.

References

1. American National Standards Institute/National Information Standards Organization. *Z39.96-2021, JATS: Journal Article Tag Suite*. version 1.3. [CrossRef] Accessed 18 October 2023.

2. Crossref. documentation. Available at: https://www.crossref.org/documentation/register-ma

intain-records/web-deposit-form/ Accessed 18 October 2023.

3. cOAlition S. Principles and implementation: plan S. Available at: https://www.coalition-s.org/addendum-to-the-coalition-s-guidance-on-the-implementation-of-plan-s/principles-and-implementation/ Accessed 18 October 2023.

4. JATS XML. Everything a publisher needs to know. Available at: https://typeset.io/resources/jats-xml-everything-a-publisher-needs-to-know/ Accessed 18 October 2023.

5. Scholastica. *The State of Journal Production and Access 2020: Society and University Publishers*. Available at: https://lp.scholasticahq.com/ journal-production-access-survey/ Accessed 18 October 2023.

6. Scholastica. *The State of Journal Production and Access 2022: Report on Survey of Independent Academic Publishers*. Available at: https://lp.scholasticahq.com/journal-production-access-survey-2022/ Accessed 18 October 2023.

7. *Pandoc, a Universal Document Converter*. Available at: https://pandoc.org/ Accessed 18 October 2023.

8. AnyStyle. Available at: https://anystyle.io/ Accessed 18 October 2023.

9. Smjernice za pripremu DOCX dokumenta. Available at: https://wiki.srce.hr/x/NHrXC Accessed 25 November 2023.

ease / publications

ese / European Science Editing

European Science Editing is an official publication of EASE. It is an open access peer-reviewed journal that publishes original research, review and commentary on all aspects of scientific, scholarly editing and publishing.

https://ese.arphahub.com/
https://www.ease.org.uk
https://twitter.com/Eur_Sci_Ed
https://www.linkedin.com/company/easeeditors/