

# Enhancing scientific publishing: automatic conversion to JATS XML

---

**Jertec Musap, Ljiljana**

*Source / Izvornik:* **European Science Editing, 2023, 49**

**Journal article, Published version**

**Rad u časopisu, Objavljena verzija rada (izdavačev PDF)**

<https://doi.org/10.3897/ese.2023.e114977>

*Permanent link / Trajna poveznica:* <https://urn.nsk.hr/urn:nbn:hr:102:888344>

*Rights / Prava:* [Attribution 4.0 International](#)/[Imenovanje 4.0 međunarodna](#)

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



*Repository / Repozitorij:*

[Digital repository of the University Computing Centre \(SRCE\)](#)



*Viewpoint*

Received: 31 Oct 2023

Accepted: 21 Nov 2023

Published: 22 Dec 2023

**Declaration of Interests**

The author has no conflict of interest to declare.

**Funding**

No funding was received for this study.

# 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](mailto:ljiljana.jertec.musap@srce.hr)

[orcid.org/0000-0002-0059-9899](https://orcid.org/0000-0002-0059-9899)



This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0).

## Citation

Cite this article as: Jertec Musap L. Enhancing scientific publishing: automatic conversion to JATS XML. *Eur Sci Ed.* 2023;49:e114977.

<https://doi.org/10.3897/ese.2023.e114977>

## 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.

## Keywords:

DOCX conversion, full-text JATS XML, HRČAK, Journal Article Tag Suite, open formats

## 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.<sup>1</sup>

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.<sup>2</sup> 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.<sup>3</sup>

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.<sup>4</sup> 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.<sup>5</sup> Furthermore, a 2022 survey indicated that there has been no growth at all in full-text XML article production since 2020.<sup>6</sup>

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

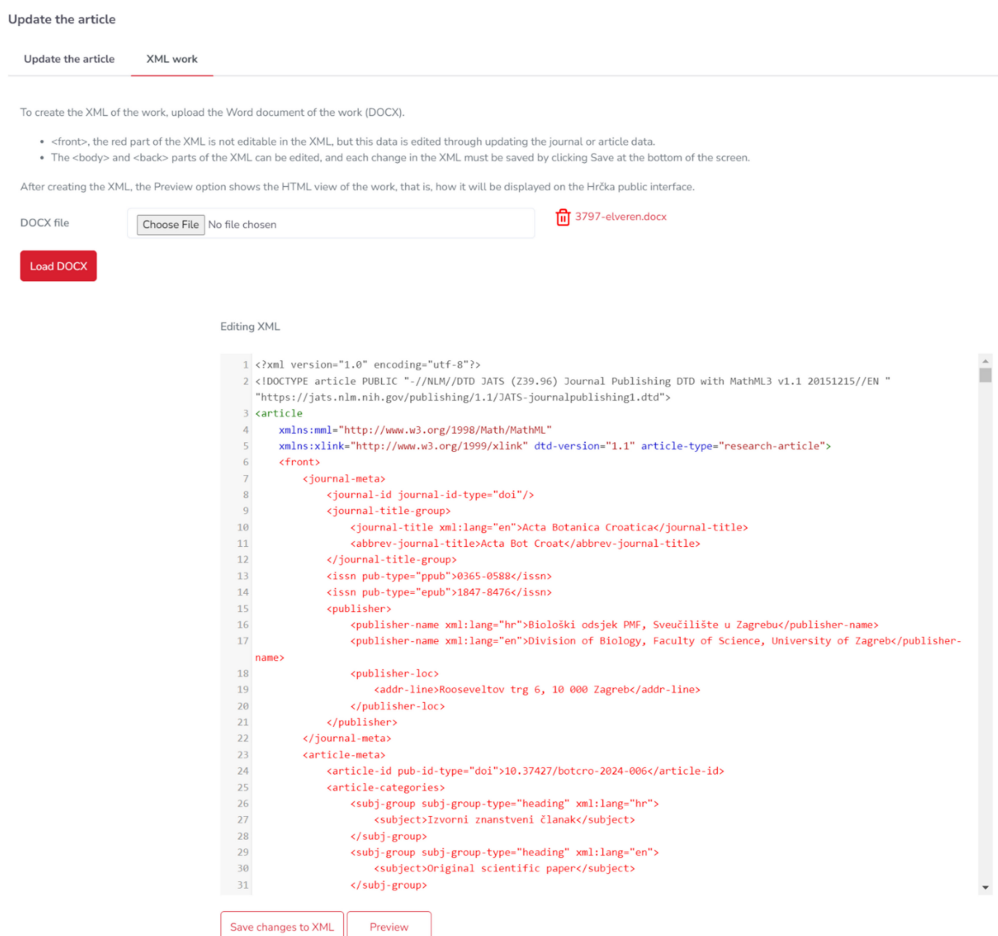
remained relatively low. To make a shift, it was clear to SRCE team that this major obstacle could only be addressed by developing the available alternative to the existing commercial tools.

#### *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*,<sup>7</sup> is responsible for converting full-text DOCX files into JATS XML. The second tool, *AnyStyle*,<sup>8</sup> 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.<sup>9</sup>

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

hrčak

Home About HRČAK Journals Journals editors Authors

Enter search term...


srce


Acta Botanica Croatica, Vol. 82 No. 2, 2023.


Original scientific paper


<https://doi.org/10.37427/Acta-2023-008>


Sugar beet cells' and extracellular events taking place in response to drought and salinity

Delia Pavoni  [orcid.org/0000-0001-7186-6063](https://orcid.org/0000-0001-7186-6063) University of Zagreb, Faculty of Science, Department of Biology, Horvatićeva 167A, 10000 Zagreb, Croatia

Anta Križanec  [orcid.org/0000-0001-9426-1482](https://orcid.org/0000-0001-9426-1482) University of Zagreb, Faculty of Food Technology and Biotechnology, Department of Chemistry and Biochemistry, 10000 Zagreb, Croatia

Ingvald Torsvik  [orcid.org/0000-0001-7186-6063](https://orcid.org/0000-0001-7186-6063) University of Zagreb, Faculty of Science, Department of Biology, Horvatićeva 167A, 10000 Zagreb, Croatia

Milana Križanec  [orcid.org/0000-0001-7186-6063](https://orcid.org/0000-0001-7186-6063) University of Zagreb, Faculty of Science, Department of Biology, Horvatićeva 167A, 10000 Zagreb, Croatia

Milana Križanec  [orcid.org/0000-0001-7186-6063](https://orcid.org/0000-0001-7186-6063) University of Zagreb, Faculty of Science, Department of Biology, Horvatićeva 167A, 10000 Zagreb, Croatia

Full text: [request PDF \(412 KB\)](#) page 128-141 downloads 303

Download JATS file

Supplements: [10M\\_wwwscs\\_publications.pdf](#)

Abstract

Salt and drought stress are important abiotic factors that negatively affect plant growth and yield. To understand how these stress factors affect metabolism at the cellular level, we analyzed cation concentrations and expression of cellular and extracellular proteins, as well as their functions and types. Cells of the industrially important, halophyte sugar beet were exposed to 300 mM NaCl and 600 mM mannitol in dimethyl Gamborg B9 liquid nutrient medium (PDS). Severe stress altered the intracellular concentrations of most of the measured cations. The cellular proteins revealed that both stressors presented significant differential regulation of 130 cellular proteins. About 80% of the identified proteins were classified in metabolism, energy or cell rescue, defense and tolerance categories. We identified several novel proteins that respond to stress, including a member of the bZIP family of transcription factors, a member of the glycine-rich RNA-binding proteins, and the K<sup>+</sup> channel beta subunit. Among extracellular proteins we found previously unreported stress-responsive proteins, a beta-vetiolase and an isoform of chitinase. The obtained results indicate that salt and drought stress disturbed the concentrations of cellular cations and affected the expression of cellular and extracellular proteins in sugar beet cells.

Keywords

extracellular proteins; mannitol; osmotic stress; proteomics; salt stress

MeSH ID


302652

URI

<https://hrca.hr/acta/302652>

Publication date

1.10.2023.



Views: 327 \*

Article information

### Introduction

Salt stress and drought are major abiotic stressors that significantly affect all aspects of plant physiology, resulting in yield losses of more than 50% and a loss of more than \$10.3 billion per year (Ma et al. 2020). Future global scenarios, envisioned by the Intergovernmental Panel on Climate Change indicate a decrease in precipitation and an increase in evapotranspiration rates (Pattin et al. 2022). Initially, both stressors cause water deficit in plants, but under prolonged salinity, plants respond to hyper-ionic and hyper-osmotic stress in addition to dehydration (Chaves et al. 2009). Plant physiological responses to these stressors aim to minimize water deficit and restore ion homeostasis (Ma et al. 2020). Osmotic adjustments are achieved through the synthesis and accumulation of osmoprotective compounds (Chen and Hua 2002), while imbalances caused by excess sodium ions are remedied by changes in the activity and abundance of ionocoupling exchangers (Dekker et al. 2014). In addition, both salinity and drought can induce the production of reactive oxygen species (ROS) that can cause lipid peroxidation and DNA damage (Jain, Puri, and Dubey 2006), and activate defense responses to suppress disease (GOS), proteases (PDS), cellulase (CAT), and ascorbate peroxidase (APX) (Ajeet and Hira 2004).

Transcriptomics studies in *Arabidopsis* revealed that 1008 and 1123 mRNAs are regulated in response to water deficit and salt stress, respectively (Fujita and Hira 2004), implying that both stressors involve complex processes. In sugar beet, an experiment with salinization and alkalization identified 4773 and 2291 differentially expressed genes in leaves and roots, respectively (Jiang et al. 2010). This and other studies identified ROS-scavenging enzymes, ion transporters and channels, proteins involved in signal transduction, and regulatory proteins, kinases, phosphatases, and transcription factors responsible for triggering the stress response (Zhu 2002; Yoshida et al. 2014). However, changes in mRNA levels do not correlate well with protein levels, and many gene products undergo posttranslational modifications that can alter protein activity (Vishnukumar 1996; Dreyhues 2010). Therefore, protein levels must be determined directly rather than extrapolated from transcript abundances. A complementary approach is to use proteomic analysis such as two-dimensional electrophoresis (2-DE) coupled with mass spectrometry (MS) to quantify protein abundance and identify it on a large scale (Friedland et al. 2009). To date, several papers have been published analyzing the response to salt and drought stress using proteomic approaches (Friedland et al. 2009; Singh et al. 2022). Although there are common proteins that are regulated during salt and drought stress, each plant species has been shown to respond differently to these stressors. The differences between salt-tolerant (halophytes) and salt-sensitive (glycophytes) plants are particularly pronounced (Kumar et al. 2008; Zhang et al. 2013). Most proteomic research has focused on changes in the abundance of cellular proteins, while knowledge of stress-induced expression of extracellular proteins is limited. Only recently, it has been shown that the extracellular matrix and its constituent proteins are involved in the response to various stressors in rice, papaya and sweet potato (Zhang et al. 2020; Kim et al. 2021).

In this study we directed our insight into the cellular and extracellular processes involved in responses to drought and salinity. To this end, we decided to use cells grown in vitro, since they represent a homogeneous system in which all cells are of similar origin and type, and the conditions of plant tissue culture allow the control of stress homogeneity and the choice of cell behavior under stress conditions independently of the regulatory systems acting at the whole plant level (Prasad et al. 2003). On the other hand, plants are composed of a number of cell types that exhibit different cellular characteristics leading to different responses to stimuli. The N, H<sub>2</sub>O and H<sub>2</sub>O<sub>2</sub> sugar beet cell lines have proven useful as *in vitro* models for studying epigenetic mechanisms, cell differentiation, and metabolism in plants (La Dity et al. 1990; Casanova et al. 2006). The N line is a normal culture dependent on plant growth regulators, in contrast to the autotrophic habituated H<sub>2</sub>O line and the tumorous T line, which is the result of cell transformation with the *Agrobacterium tumefaciens* T<sub>1</sub> plasmid BR2 (Friedland et al. 2012a). In this study, we used the differentiated N line, which contains mainly parenchyma cells, is photosynthetic and grows in response to 2,4-dichlorophenoxyacetic acid and 6-benzylaminopurine. It exhibits normal nuclear morphology and cell wall cellulose deposition (Friedland et al. 2012a). To investigate the effects of salt- and mannitol-induced stress at the cellular level, we sought to identify stress-related proteins that are differentially expressed in non-stressed and stressed cells. We also determined possible changes in the concentrations of cellular macro- and microelements. In addition, this study was extended by analyzing the expression of stress-related extracellular proteins. We report the disruption of macro- and microelement homeostasis as a consequence of stress and the identification of novel proteins in sugar beet as stress-related proteins.

### Materials and methods

#### Plant material

Sugar beet N cell line (Beta vulgaris L., subsp. vulgaris var. affinis [D0]) was grown in vitro in modified Gamborg B9 liquid nutrient medium (PDS) (Bjorklund et al. 1970; Piskovcic et al. 2007). The growth chamber was maintained at 22 °C and a 16 h photoperiod (80 µmol photons m<sup>-2</sup> s<sup>-1</sup>). Cells were subcultured every two weeks by transferring 10 mL of the old cells into 40 mL of fresh PDS medium. The suspensions were shaken on a rotapost shaker at 125 rpm.

#### Experimental conditions and harvesting

Salt stress was generated by growing cells in liquid PDS medium containing 300 mM NaCl, while physiological drought was provided by growing cells in the same medium containing 600 mM mannitol. Cells were harvested after 72 h of incubation, washed thoroughly with distilled H<sub>2</sub>O, dried, and rapidly frozen in liquid nitrogen until use.

#### Macro- and microelement analysis

Five samples were used to determine macro- and microelement concentrations. Samples were analyzed by inductively coupled plasma atomic emission spectroscopy (ICP-AES) using the Prodigy High Dispersion ICP instrument (Teledyne Leeman Labs, Hudson, NH). ICP multi-element standard solution IV (Merck, Darmstadt, Germany) was used to control plasma positioning and to prepare standard solutions for calibration. All calibration standards were prepared by appropriate dilution of standard stock solutions (1 g L<sup>-1</sup>) in a concentration range from 0.1 to 5.0 mg L<sup>-1</sup>. Lyophilized samples were dried at a constant temperature of 70 °C for 1 h and then pulverized in a porcelain mortar. An amount of 0.15 g of each dried sample was weighed with analytical accuracy and placed in Teflon vials, except for the solutions of the nutrient medium. 4 mL of concentrated nitric acid (HNO<sub>3</sub>, 1.0 mL of hydrogen peroxide (H<sub>2</sub>O<sub>2</sub>, 30% w/w) and 0.2 mL of ultrapure deionized water (H<sub>2</sub>O) were added and the vials were left open for 30 min. The vials were sealed and placed in a microwave-assisted high-pressure digestion system (Biospec, Germany). Digestion was performed in several steps for 40 minutes. After cooling to room temperature, the solutions were filtered, transferred to 50 mL volumetric flasks, and filled to the mark with ultrapure deionized water. All samples were digested and analyzed as duplicates; blanks were also prepared in the same manner as the samples. To verify the accuracy of the digestion procedure, the same digestion scheme was applied to the certified reference material (SRM 1571 – Orchard leaves). The results are presented as mg macroelement per g of dry weight (DW) or µg microelement per g of DW together with the standard deviation of measurements.

#### Analysis of cellular proteins

The frozen cells were ground to a fine powder in liquid nitrogen using a pre-cooled mortar and pestle. For 2-DE, the phenol extraction protocol was performed according to a published procedure (Friedland et al. 2007). Protein concentration was determined by the modified Bradford method using a UV/VIS spectrophotometer UV-160U (Shimadzu, UK) and bovine serum albumin (BSA) as a standard (Friedland et al. 2007).

The first dimension, isoelectric focusing (IEF), was performed using 18 cm long non-linear, immobilized pH gradient (iPG) strips, pH 3–10, in the iPGher system (GE Healthcare, USA), according to Piskovcic et al. (2012). The IEF strips were stained at 80 °C until use. iPG strips were thawed and incubated for 15 min in a buffer composed of 0.05 M Tris-HCl, pH 8.8, 0.4 M urea, 2% SDS (w/v) containing 120 mM dithiothreitol (DTT) and then for 5 min in a buffer of the same composition, but with 120 mM isobutanol instead of DTT. The second dimension was performed by sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) as described by Friedland et al. (2012a), using the PROTEOMAN 3-ri system (BioRad, USA).

#### Analysis of extracellular proteins

Extracellular proteins were harvested from the liquid medium after cells were removed. To remove debris, the medium was filtered through No. 1 Whatman filter papers (Whatman, UK) and again through 0.45 µm Millipore filters (Millipore, USA). Proteins were concentrated on Amicon ultrafiltration devices (Millipore, USA) with the cut-off at 3 kDa. Protein concentration was determined by the Bradford method using a spectrophotometer and BSA as a standard (BioRad 1976). Proteins were mixed with Laemmli buffer (Laemmli 1970) and loaded onto a large vertical electrophoresis system. Electrophoresis was performed for 30 min at 100 V in stacking gel containing 4% T and 2.67% C and then at 220 V in running gel (12% T, 2.67% C until the bromophenol blue ran off the gel).

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://twitter.com/Eur_Sci_Ed)  
<https://www.linkedin.com/company/easeeditors/>



© 2023 the authors. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.