 doi:10.20944/preprints202211.0093.v1

Article

Automatic XML extraction from Word and formatting of e-book formats: Insight into the Open Source Academic Publishing Suite (OS-APS)

Carsten Borchert [1], Roberto Cozatl [2], Frederik Eichler [3], Astrid Hoffmann [4], and Markus Putnings [5,*]

[1] SciFlow GmbH, Altensteinstraße 40, 14195 Berlin, Germany; carsten.borchert@sciflow.net
[2] University and State Library Saxony-Anhalt, Martin-Luther University, Halle-Wittenberg, 06098 Halle, Germany; roberto.cozatl@bibliothek.uni-halle.de
[3] SciFlow GmbH, Altensteinstraße 40, 14195 Berlin, Germany; frederik.eichler@sciflow.net
[4] Friedrich-Alexander-Universität Erlangen-Nürnberg, Universitätsbibliothek Erlangen-Nürnberg, Universitätstr. 4, 91054 Erlangen, Germany; astrid.hoffmann@fau.de
[5] Friedrich-Alexander-Universität Erlangen-Nürnberg, Universitätsbibliothek Erlangen-Nürnberg, Universitätstr. 4, 91054 Erlangen, Germany; markus.putnings@fau.de

* Correspondence: markus.putnings@fau.de, Tel.: +49 9131 8527835

Abstract: Due to resource constraints, most Diamond Open Access journals publish less than 25 articles per year, and 75% of journals are not able to provide their content in XML and HTML, primarily providing only PDFs (Bosman et al., 2021, p. 7-8). In order to keep up with larger commercial publishers, a high degree of automation and streamlining of processes is necessary. The Open Source Academic Publishing Suite (OS-APS) project funded by the German Federal Ministry of Education and Research aims to achieve this. OS-APS automatically extracts the underlying XML from Word manuscripts and offers optimization and export options in various formats (PDF, HTML, EPUB). The professional corporate design, e.g., of the PDFs, is managed automatically by using templates or creating one's own using a Template Development Kit. OS-APS will also connect to scholarly-led and community-driven publishing platforms such as Open Journal Systems (OJS), Open Monograph Press (OMP), and DSpace: the software will be able to be integrated into a wide range of publication processes, whether at small, low-resource commercial Open Access Publishers, or institutional and Diamond Open Access Publishers.

References: Bosman, J., Frantsvåg, J. E., Kramer, B., Langlais, P.-C., & Proudman, V. (2021). Oa Diamond Journals Study. Part 1: Findings. https://doi.org/10.5281/zenodo.4558703

Keywords: automatic typesetting; media-neutral publishing; open access; open source; scholarly publishing; XML/HTML conversion

1. Introduction

The 2021 OA Diamond Journals Study [1] has compiled a representative overview of Diamond Open Access journal operators in its "Part 1: Findings". For example, 53% of journals are operated by less than 1 Full-time equivalent (FTE), and 60% of journals rely heavily on volunteers. Due to these resource constraints, most Diamond Open Access journals publish less than 25 articles per year, and 75% of journals are not able to provide their content in XML and HTML, primarily providing only PDFs.

In order to keep up with larger commercial publishers and their professionalized content offerings, a high degree of automation and streamlining of processes is necessary. The Open Source Academic Publishing Suite (OS-APS, https://os-aps.de/en/) project

 doi:10.20944/preprints202211.0093.v1

funded by the German Federal Ministry of Education and Research aims to achieve this. For this purpose, an open source software is to be developed by means of research (especially requirements analysis) and development, with which

1. XML is automatically extracted from (e.g. *.docx Word) author manuscripts;
2. the XML can be processed and e.g. supplemented with semantic information;
3. and various e-journal or e-book output formats (e.g. XML, HTML, EPUB, and PDFs) can be generated in a corporate design.

In addition, OS-APS will also connect to widely-used, scholarly-led and community-driven publishing platforms such as Open Journal Systems (OJS), Open Monograph Press (OMP), and Open Access repositories (e.g. DSpace). The software will be able to be integrated into a wide range of publication processes, whether at small, low-resource commercial Open Access Publishers, or institutional and Diamond Open Access Publishers.

To understand the requirements of these heterogeneous publishers, a practical advisory board and scientific advisory board with representatives from the different publication sectors accompany the OS-APS project. In addition, an extensive survey [2] was conducted across various publishing houses and demo days with corresponding feedback opportunities are held regularly (https://os-aps.de/demo/).

The project is also in line with the recommendations of the OA Diamond Study and its urgent call for cOAlition S Funders and Infrastructures: “Support the development of generic tools to generate structured content in XML and HTML” [3]. This will also be a prerequisite for creating new, dynamic and machine-processable media formats, for example in terms of accessibility and screen readers.

The Open Source software could be thus a significant improvement for smaller, independent Open Access Publishers. It offers the possibility to increase the effectiveness and efficiency of their processes to create, for example, new e-journal article or e-book formats such as HTML and EPUB. These developments contribute to a higher bibliodiversity and may help independent OA-publishers to become more viable and sustainable in the long term.

2. Materials and Methods

2.1. Materials

In terms of materials, the OS-APS project team has so far produced insights into the project’s progress via presentations, posters, and articles [[2],[4],[5],[6],[7],[8]], various software development sprints documented on GitLab [9], and a demo [10] to provide hands-on testing and feedback on the developments to date.

2.2. Methods

Methodologically, the project work is divided into four milestones. In the first, the requirements for the software were analyzed. In the second, all technology components, interfaces and intended workflows for connecting e.g. OJS, OMP and DSpace were developed on this basis. In the third, existing journals and book series at the publishing services of the project partners will be iteratively converted to the Open Source Academic Publishing Suite production workflow for the purpose of practical testing and proof of implementation. In the fourth, a release of all open source software development results will take place; the OS-APS software can be downloaded free of charge and installed on the publishers’ own servers (all components are browser-based).

The following sections describe the used methods within the milestones in more detail.

The entire OS-APS project is accompanied by two advisory boards, which consist of publishers as well as institutions and libraries active in publishing. The Scientific Advisory Board is responsible for strategic and methodological advice and the User Advisory Board for discussions on practical procedures and requirements.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 November 2022 doi:10.20944/preprints202211.0093.v1

2.2.1. Software requirements analysis

Interviews were conducted with the publishers and publishing services of both the OS-APS project partners and advisory boards. The various publishing operations were mapped graphically via miro (https://miro.com). Miro provides a range of different options and templates for software development purposes (https://miro.com/templates/developers/). Subsequently, a review and statistical evaluation and classification of the workflows took place, whether they are, e.g., Word-, InDesign- or LaTeX-based and which export formats are generated.

After obtaining a structured overview of various publishing processes through these interviews, a broader online survey was designed on this basis to reach a lot of more experts from the publishing community. This online survey was methodologically designed with Typeform (https://www.typeform.com), which is an online software specialized in creating dynamic surveys with logic flows.

The survey was sent to the mailing lists of the German AG Universitätsverlage (https://ag-univerlage.de), The Association of European University Presses (https://www.aeup.eu), the Enable! community (https://enable-oa.org), Peergroup Produktion of IG Digital (https://www.boersenverein.de/interessengruppen/ig-digital/die-peergroups-der-ig/#accordion-23919), GeSIG (https://gesig.org), Library Publishing Coalition (https://librarypublishing.org), Association of University Presses (https://aupresses.org), Open Access Scholarly Publishing Association (https://oaspa.org), ACUP / APUC - Association of Canadian University Presses (http://acup-apuc.ca), The Association of Japanese University Presses (https://www.ajup-net.com) as well as to cooperation partners of OA-STRUKTKOMM (https://oa-struktkomm.htwk-leipzig.de), DEval Communication and Publications Office (https://www.deval.org/de/publikationen), Center for Digital Systems Berlin (https://www.cedis.fu-berlin.de) and in forums such as the Open Access Books Network (https://openaccessbooksnetwork.hcommons.org) and the German PKP Community Forum (https://forum.pkp.sfu.ca/c/regional-networks/german-topics/13).

The results were evaluated, processed in a structured manner [2] and had a significant impact on some project decisions. Thus, it was decided to initially focus on Word manuscripts for XML extraction, since e.g. LaTeX or other manuscript format submissions were rather rare.

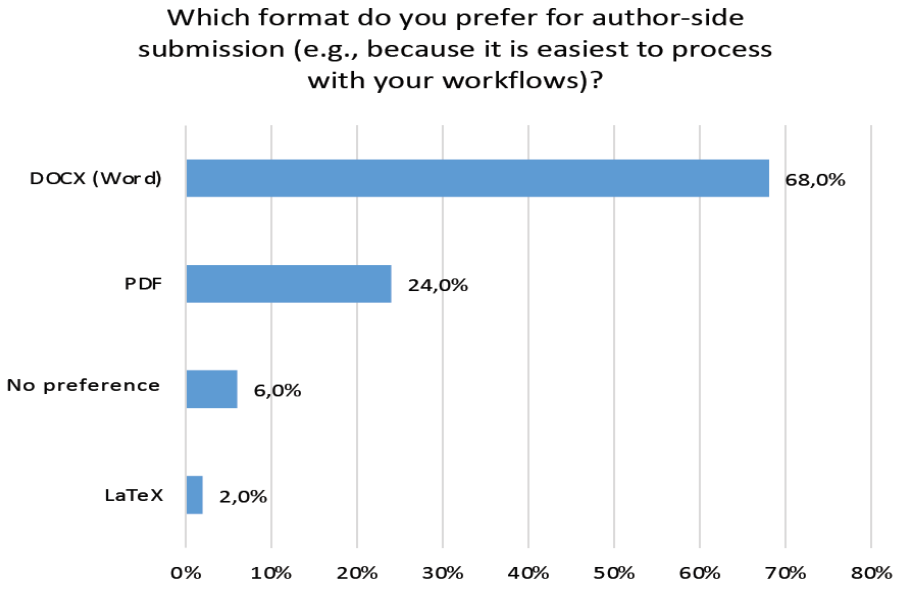

Figure 1. Manuscript acceptance preferences of the surveyed publishers (own representation, translated from [2]).

2.2.2. Open source development of the technology components

 doi:10.20944/preprints202211.0093.v1

Methodologically, it was planned to build on existing open source software wherever possible. In several cases it was also possible to build on existing code of the project partner SciFlow, which offers an online platform for collaborative scientific writing and automatic formatting according to the format specifications of renowned academic publishers (cf. https://www.sciflow.net/en/sciflow-free-researchers). SciFlow has extracted the relevant components from its platform and made them available as open source. Additional software development parts in the project context were that

- Word *.docx manuscripts can be fed in the OS-APS browser-based importer;
- the XML is automatically extracted from them and is displayed in an editable browser interface; for this, many components were reused from SciFlow's collaborative writing and editing platform;
- options for optimizing and semantically enriching the XML can be provided;
- corporate design templates depending on the publisher or its content, for example for different book or journal series, can be used;
- and that the publishing user should be able to control this corporate design "look" himself using the planned template development kit.

The Open Source software is currently based on Pandoc, Docker, paged.js, and components extracted by SciFlow from their own platform: https://gitlab.com/sciflow/development/-/milestones.

Pandoc (https://pandoc.org) is a free, GPL-licensed (https://www.gnu.org/licenses/gpl-3.0.html) converter and parser software. It is used to convert one document-based markup and file format to another.

Docker (https://www.docker.com) is an open platform for the running of applications. In this project, it is used to streamline the development for our OJS, OMP and DSpace platforms and to ease the deployment of ready-to-go code from our test environments onto our production systems.

Paged.js (https://pagedjs.org) is an open source library for displaying paginated content in the browser and then creating PDFs and their designs using e.g. HTML and CCS.

2.2.3. Proof of implementation and application of the Open Source Academic Publishing Suite

At the University and State Library in Sachsen-Anhalt (ULB-SA) a testbed was created which can serve the purpose of implementing a number of the software tools developed in the course of the project. The library team supports several publishers' teams in their efforts to publish a wide range of journals spanning across topics such as social geography, transnational economic law, ecology, geosciences. Out of this selection of journals, monographs and series, it has been possible to choose specific examples which have allowed us to not only test specific modules of the OS-APS developed tools but also the connection and integration of our publication tools OJS and OMP systems to a DSpace based publications repository.

In the first case, one journal, the "Hallesches Jahrbuch für Geowissenschaften" (the Yearbook of Geosciences in Halle) and the ULB-SA's own series "Schriften zum Bibliotheks- und Büchereiwesen in Sachsen-Anhalt" (series on librarianship studies in Saxony-Anhalt) have been selected to be enhanced and given new layouts via the usage of the OS AP suite. In this particular case word*.docx templates are uploaded into the OS-APS environment and specific output formats can be generated for importing into the OJS of the ULB-SA. This process streamlines the template generation process of editorial teams, increases its level of automatization, and generally contributes to an increase in citations rate and visibility. These actions are in line with specific Open Science principles which aim at improving the accessibility and reusability of research outputs in fields where these issues may still need attention such as in some areas of the digital humanities. Scholars in these fields have recognized these endeavors as key components that can promote new research opportunities and can have a great societal value impact [11].

 doi:10.20944/preprints202211.0093.v1

Regarding our connection to our publication tools, a number of journals and series (see for example MLU Human Geography Working Paper Series and the Policy Papers on Transnational Economic Law) are now fully integrated into our OJS/OMP publication cluster and have been exported to our DSpace repository. In this process, all articles have been issued with persistent identifiers (DOIs) and have thus gained a higher visibility and findability given the high data discoverability advantages that the DSpace platform offers. An ongoing migration is taking place so that a total of 13 journals will be migrated in the scope of this project.

As for the technical connection, it has been done in a way that modular scripts are independently available to suit the different needs of our prospective end users. This means that the developed scripts can be implemented as a full set of scripts or just individually depending on the specifications of the environment where the tools are to be deployed. This modular approach has also meant that our developments do not compromise the native code and functionality of the publication tools in a way that further system upgrades or updates are compromised.

2.2.4. Release of the open source software development results

The open source software can be downloaded free of charge from https://os-aps.de and a suitable repository, presumably GitLab, after the end of the project (31.12.2022, if necessary the project will be extended cost-neutrally, then possibly also later in spring 2023). Alternatively, a paid hosting and support offering from SciFlow can be used: https://www.sciflow.net. The business and fee model for this will also be announced in parallel with the phase-out of the project.

Accompanying documentation of the software is of course also provided. The OJS and OMP to DSpace connection scripts and a series of quality control and validation scripts as well as documentation on how these publishing tools have been setup under Docker will be fully available as open source code as part of the project's integral code materials. As part of our project commitments towards open science and transparency and reproducibility, we have already published some of the scripts (in a none-finalized and openly available for scrutiny and feedback version) over the Github repository of the University and State Library in Sachsen Anhalt (explore for instance, our OJS/OMP2DSpace connecting script, and our scripts for the dockerisation of OJS and OMP).

3. Results

The workflow extracted from the perceptions and requirements of the surveyed publishing group is shown in Fig. 2 (see also 2.2.2).

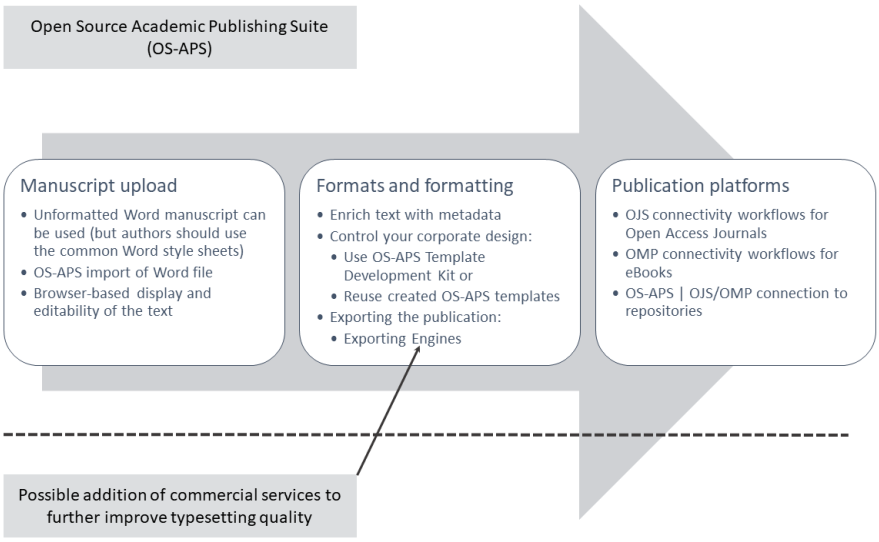

Figure 2. Overview over the Open Source Academic Publishing Suite functionalities

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 November 2022 doi:10.20944/preprints202211.0093.v1

3.1. OS-APS importer and editor

Manuscripts can be imported into the programmed OS-APS editor. By extracting XML structures, elements such as column titles, page breaks, tables, etc. are recognized. In the editor, it is possible to change the text as well as the formatting, if elements were not recognized correctly. If necessary, more metadata (e.g. with regard to accessibility) and semantic references can be added.

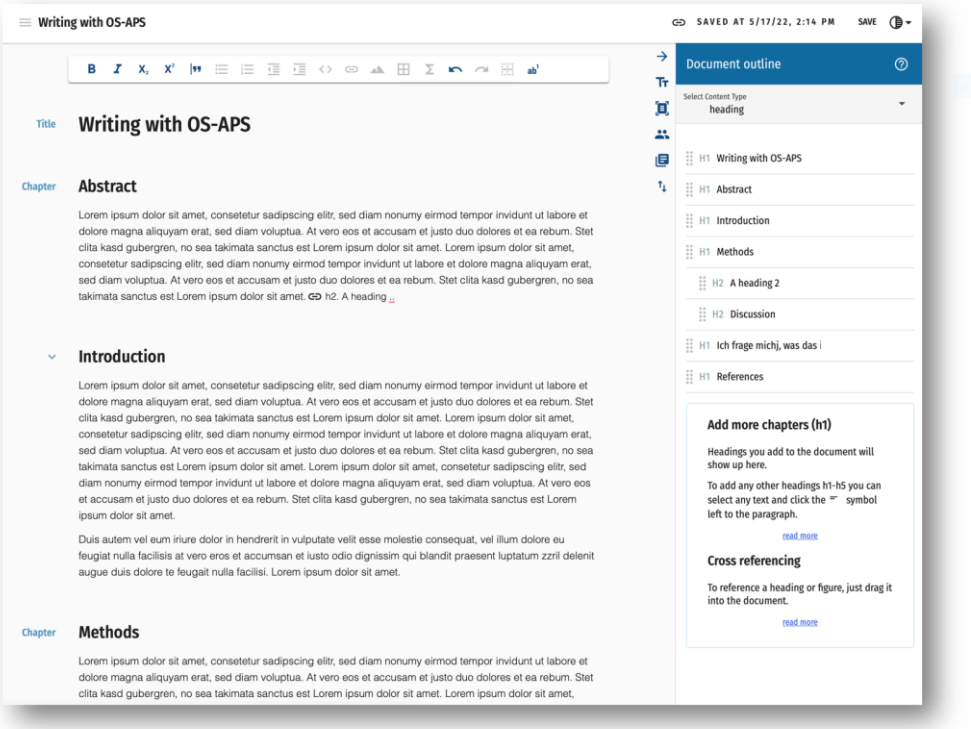

Figure 3. View of the OS-APS editor

3.2. Template Development Kit and re-usable templates

During export, the corporate design of the respective publisher is mapped via templates. Various standard templates are provided and can be reused.

Further templates and exports can be developed using the Template Development Kit. This is particularly interesting for publishers who have very clear format specifications and do not want to deviate from them.

With the help of the Template Development Kit, individual parameters in ready-made templates can be easily changed. It is also possible to create completely new templates, although this requires prior technical knowledge (esp. web programming). New exports can also be programmed in this way. The Template Development Kit is based on the open source software Pandoc and on SciFlow's own development.

Commercial non-open-source based tools can also be integrated during export for typesetting optimization, specifically e.g. Prince XML (https://www.princexml.com).

3.3. Connection to OJS, OMP and repositories such as DSpace

The OJS and OMP applications are deployed via a Docker environment; the OJS and OMP systems are connected to a DSpace repository (specifically a DSpace 6.3. version, in the case of our project partners). As part of the intended workflow, OJS and OMP data will be exported to DSpace with subsequent return of DOI information. The corresponding publications are displayed in OJS and OMP as well as in DSpace. In general, OJS and OMP are intended as presentation platforms and DSpace for long-term archiving.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 November 2022 doi:10.20944/preprints202211.0093.v1

The connection scripts as well as documentation on how these publishing tools have been setup will be fully available as open source code as part of the project's integral code (see section 2.2.4).

Fig. 4 gives an overview of the interfaces and data paths.

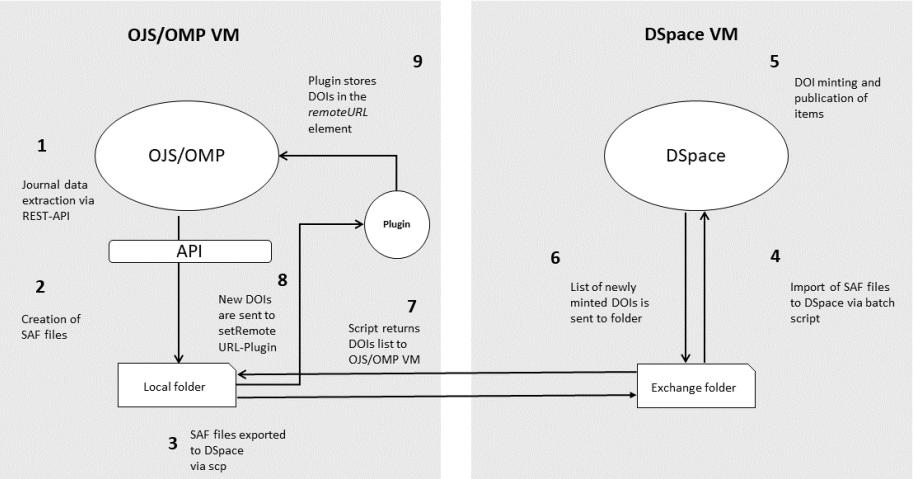

Figure 4. Schematic representation of the connection of OJS, OMP and DSpace to OS-APS

3.4. Test possibility of the current results

Every first Wednesday of the month, a "Demo Day" takes place, where interested parties are invited to test the current state of the OS-APS software and give feedback. The dial-in data for the video conference is published on the OS-APS website shortly before the event: https://os-aps.de/demo/.

For the final release of the software and documentation, see methodological announcements in 2.2.4.

4. Discussion

4.1. Possible necessary exceptions to the OS-APS workflow

The OS-APS software development project is currently on schedule with its planned milestones. The basic objectives described in the introduction are being achieved. However, the tests conducted so far show that not all special cases that might occur in manuscripts can be implemented graphically in e.g. PDFs in an ideal way.

This applies primarily but not exclusively to art volumes in which various figures must have exactly the same arrangement as in the original manuscript, grouped figures (e.g. as a block of four or six) with one caption, large rotated tables, nested tables with multiple content types (e.g. images in different cells of the table), Word text fields or images originally drawn in Word itself with multiple image elements, and much more.

In addition, there may be quality requirements from both publishers and authors that necessitate very thoughtful, small-scale, manual typesetting in InDesign, for example. Examples could include art and exhibition volumes. Here, too, the fully automated approach may not meet these individually high quality requirements.

4.2. Discussion about embedding the new output formats

What publishers or platforms do with the new output formats remains deliberately open and up for discussion. Those who previously only distributed PDFs via OJS, OMP or repositories (e.g. the university repository, in the case of university presses), must think about how and where they integrate the HTML or EPUB files when using OS-APS, e.g.,

 doi:10.20944/preprints202211.0093.v1

whether they provide viewers or corresponding plug-ins and whether they also archive them over the long term (or continue to only archive PDF/A).

In addition, they have to think about URL, DOI and, for eBooks, ISBN registration with regard to the new output formats. In the case of repositories and the use of one front door under which all formats hang, a single DOI could still be used, for example. However, according to the German "Verzeichnis lieferbarer Bücher" (https://vlb.de) as ISBN agency, each different e-book format needs its own ISBN.

4.3. Possible OS-APS platform extensions in the future

The OS-APS platform was developed as open source software. In addition, however, the project partner SciFlow will also offer hosting, then for a fee, for those who do not want to set up their own server to run the software or do not want to worry about support.

In the context of this support, further extensions are conceivable, for example with regard to special viewers, such as for EPUB or HTML, depending on the existing information infrastructure, or in terms of accessibility support. The project team is happy to enter into discussions.

5. Conclusions

Preparing manuscripts for various formats such as HTML or EPUB can pose challenges for small and medium-sized as well as non-commercial (e.g. university) academic publishers: A high level of professionalism often requires extensive technical expertise as well as the use of cost-intensive XML content management systems.

The third-party funded project "Open Source Academic Publishing Suite (OS-APS)" provides relief in this area. It is intended to enable academic publishers to publish in a media-neutral way by using XML-based workflows. The XML is automatically extracted from Word manuscripts and the corporate design of the exported PDFs can be controlled via templates. Institutions or publishers using OJS or OMP can also reuse the workflows and connections documented in the project. OS-APS is thus closely integrated into the open science landscape.

Supplementary Materials: The following supporting information can be downloaded at: www.mdpi.com/xxx/s1, Figure S1: title; Table S1: title; Video S1: title.

Author Contributions: Conceptualization, C.B., F.E., A.H., and M.P.; methodology, C.B., F.E., A.H., and M.P.; software, C.B., and F.E., developers at the ULB-SA; validation, C.B., F.E., and M.P.; formal analysis, C.B., F.E., A.H., and M.P.; investigation, C.B., F.E., A.H., and M.P.; resources, C.B., F.E., and M.P.; data curation, C.B., R.C. and F.E.; writing—original draft preparation, C.B., R.C., F.E., A.H., and M.P.; writing—review and editing, C.B., R.C., F.E., A.H., and M.P.; visualization, C.B., R.C., F.E., A.H., and M.P.; supervision, C.B., R.C., F.E., A.H., and M.P.; project administration, C.B., R.C., F.E., A.H., and M.P.; funding acquisition, C.B., F.E., and M.P.. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the German FEDERAL MINISTRY OF EDUCATION AND RESEARCH, grant numbers 16TOA017A (SciFlow), 16TOA017B (FAU), and 16TOA017C (ULB Sachsen-Anhalt).

Data Availability Statement: The data and software presented in this study are or will be made available on https://os-aps.de within the specified project deliverable times.

Acknowledgments: The authors acknowledge the support provided by the Members of the Scientific Advisory Board and the Members of the User Advisory Board (https://os-aps.de/en/participate/) on the OS-APS project and software development.

Conflicts of Interest: The authors declare no conflict of interest. The funder had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Bosman, Jeroen; Frantsvåg, Jan Erik; Kramer, Bianca; Langlais, Pierre-Carl; Proudman, Vanessa *OA Diamond Journals Study. Part 1: Findings*; Zenodo, 2021;
2. Borchert, C.; Eichler, F.; Söllner, K.; Putnings, M.; Hoffmann, A.; Berghaus-Sprengel, A.; Brenn, D. Open Source Academic Publishing Suite (OS-APS): OA-Publikationen medienneutral mit automatisiertem Corporate Design erstellen. **2022**.
3. Becerril, Arianna; Bosman, Jeroen; Bjørnshauge, Lars; Frantsvåg, Jan Erik; Kramer, Bianca; Langlais, Pierre-Carl; Mounier, Pierre; Proudman, Vanessa; Redhead, Claire; Torny, Didier *OA Diamond Journals Study. Part 2: Recommendations*; Zenodo, 2021;
4. Putnings, M.; Borchert, C.; Eichler, F. Project Open Source Academic Publishing Suite (OS-APS). 2021.
5. Söllner, Konstanze; Putnings, Markus; Hoffmann, Astrid Birgit; Berghaus-Sprengel, Anke; Brenn, Daniel; Borchert, Carsten; Eichler, Frederik Open Source Academic Publishing Suite (OS-APS): Medienneutrales OA-Publizieren Im Eigenen Corporate Design. **2021**, doi:10.5281/ZENODO.5526591.
6. Söllner, K.; Putnings, M.; Hoffmann, A.; Berghaus-Sprengel, A.; Cozatl, R.; Brenn, D.; Borchert, C.; Frederik, E. Publikationen medienneutral und automatisiert gemäß den eigenen Stilrichtlinien erstellen mit der Open-Source-Software „OS-APS“. Presented at the DINI-Jahrestagung, 2021.
7. Söllner, K.; Putnings, M.; Hoffmann, A.; Berghaus-Sprengel, A.; Borchert, C.; Eichler, F. Open Source Academic Publishing Suite (OS-APS): Simple, Media-Neutral OA Publishing with Automatic Typesetting. *SCS* **2021**, doi:10.7557/5.6188.
8. Putnings, M.; Borchert, C.; Cozatl, R. Ein Einblick in Das BMBF-Projekt Open Source Academic Publishing Suite (OS-APS). *ABI Technik* **2022**, *42*, 166–173, doi:10.1515/abitech-2022-0030.
9. SciFlow *Milestones*; SFO Development; 2022;
10. OS-APS Demo Available online: https://os-aps.de/demo/ (accessed on 24 August 2022).
11. Führ, F.; Bisset Alvarez, E. Digital Humanities and Open Science: Initial Aspects. In Proceedings of the Data and information in online environments; Bisset Álvarez, E., Ed.; Springer International Publishing: Cham, 2021; pp. 154–173.