

# Automatic XML extraction from Word and e-book formatting: an insight into the Open Source Academic Publishing Suite (OS-APS) project

## *Markus Putnings*

Friedrich-Alexander-Universität Erlangen-Nürnberg  
Erlangen, Bavaria, Germany  
shushaniksargsyan8@gmail.com

## *Carsten Borchert*

SciFlow  
Berlin, Germany  
carsten.borchert@sciflow.net

## *Frederik Eichler*

SciFlow  
Berlin, Germany  
frederik.eichler@sciflow.net

## ABSTRACT

According to the OA Diamond Journals Study of diamond open access journal operators, 53% of journals employ less than one full-time worker, 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 but only in PDF (Bosman et al., 2021, p. 7-8).

For them to keep up with larger commercial publishers and their professionalized content output, a high degree of process automation and streamlining is necessary. The Open Source Academic Publishing Suite (OS-APS, <https://os-aps.de/en/>) project funded by the German Federal Ministry of Education and Research aims to achieve this. Smaller and medium-sized publishers usually deliver Word manuscripts. OS-APS automatically extracts the underlying XML from these manuscripts, offers optimization, and, most importantly, export in various formats (PDF, HTML, XML, EPUB). Professional corporate design, such as that of the PDFs, is also managed automatically by reusing templates or by creating one's own with the OS-APS Template Development Kit.

In addition, OS-APS will also connect to scholarly 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 with small, low-resource, commercial, institutional, or diamond open access publishers. To understand the requirements of these heterogeneous publishers, the OS-APS project has a practical advisory and a scientific advisory board with representatives from the different publishing sectors. In addition, it regularly holds demo days with feedback opportunities. One such demo will be given at PUBMET2022.

The Open Source software could be a significant improvement for smaller, independent OA publishers to streamline their processes, create new e-book formats (such as HTML, EPUB), and secure their long-term presence and bibliodiversity. The project is in line with the recommendations of the OA Diamond Study and its call for cOAlition S funders and infrastructures to "support the development of generic tools to generate structured content in XML and HTML" (Becerril et al., 2021, p. 20).

## KEYWORDS

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

## REFERENCES

- ▶ Becerril, A., Bosman, J., Bjørnshauge, L., Frantsvåg, J. E., Kramer, B., Langlais, P.-C., Mounier, P., Proudman, V., Redhead, C., & Torny, D. (2021). Oa Diamond Journals Study. Part 2: Recommendations. <https://doi.org/10.5281/zenodo.4562789>
- ▶ 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>