

Communication

The Importance of Free and Open Source Software and Open Standards in Modern Scientific Publishing

Michael L. Wilson ^{1,*} and Vakhtang Tchantchaleishvili ²

¹ Centre for Injury Prevention and Community Safety, PeerCorps Trust Fund, 352 Makunganya Street, Co-Architecture Building 4th Floor, P. O. Box 22499, Dar es Salaam, Tanzania

² Division of Cardiac Surgery, University of Rochester Medical Center, Box Cardiac Surgery, 601 Elmwood Avenue, Rochester, NY 14642, USA;
E-Mail: vakhtang_tchantchaleishvili@urmc.rochester.edu

* Author to whom correspondence should be addressed; E-Mail: michael.wilson@peercorpstrust.org; Tel.: +255-754-636963.

Received: 12 April 2013; in revised form: 7 June 2013 / Accepted: 19 June 2013 /

Published: 26 June 2013

Abstract: In this paper we outline the reasons why we believe a reliance on the use of proprietary computer software and proprietary file formats in scientific publication have negative implications for the conduct and reporting of science. There is increasing awareness and interest in the scientific community about the benefits offered by free and open source software. We discuss the present state of scientific publishing and the merits of advocating for a wider adoption of open standards in science, particularly where it concerns the publishing process.

Keywords: open source; open access; electronic publishing; free software

1. The State of Scientific Publishing

It is near universal practice that scientific journals require authors to submit prepared manuscripts in proprietary file formats. By extension, this also means that as authors we are to some degree restricted to using proprietary software. The most commonly used formats by journals in the peer-review, editorial and publication processes are DOC/DOCX for written text and XLS/XLSX for graphs and tables [1]. PPT/PPTX files are sometimes requested for graphs or embedded images [2]. These formats

have a number of issues associated with them which ultimately make science less open, less transparent and the scientific authorship process less accessible.

Firstly, in order to read, edit and create documents in these formats with complete compatibility, we are required to license proprietary software at often great expense. This has a discriminative impact on researchers with modest means because they are forced to purchase proprietary software when free and open source software (FOSS) alternatives are widely available. While many FOSS alternatives such as LibreOffice do allow reading, editing and exporting into proprietary formats such as DOC/DOCX [1], the Microsoft Corporation has not allowed the release of full documentation on its formats for optimal compatibility [3]. Indeed it would represent a potential threat to its business model to do so [4].

Thus the reverse engineering of proprietary formats has become necessary by FOSS developers to enable compatibility. While the results of reverse engineering have generally been acceptable, it is far from optimal. This pay-to-participate model in science is undesired and should not be a feature of an inclusive scientific publishing environment. Furthermore, because the Microsoft Corporation makes slight changes to its file formats with each version release, users become locked into a proprietary ecosystem of software upgrades [4]. Not only does the company make these changes, it does so to (1) force users to buy new versions of the software and (2) create a moving target for the competition that must reverse engineer each new version, again increasing the chances that users will need to buy the commercial proprietary software rather than use the alternatives to remain compatible. Because of these often subtle changes to the software, the competition, at a minimum, will lag behind and, at worst case, never support certain versions/features.

Second, using proprietary formats can pose important security and confidentiality risks. Because MS-Word allows full macro-scripting it has become a common carrier for computer viruses [5]. This means that embedded within an DOC/DOCX file can be a malicious computer program which runs without the recipient's permission each time they view the file on their computer. Also, due to the way in which MS-Word stores its version changes, it has been possible for recipients to see prior drafts of the sender's document that may contain confidential information [6].

Thirdly, storing important data in proprietary file formats puts that very data at risk of being lost. Computers and software have made the storage of data convenient and safer in many instances. However, data that is stored in a proprietary file format today, may not be readable in the future. The very programs which are used to record data and the file formats in which they are stored can become obsolete over time. Furthermore, because those same programs and file formats may be the property of a corporation, if the company goes bankrupt and the software is pulled from the market, data stored in these formats may be lost as well [7]. If the source code for these programs and file formats were made available, such as is the case with FOSS under an appropriate license, a programmer could with some effort resurrect the original software to read and recover the data. Because of these reasons, the continued use of proprietary formats for archiving scientific data not only represents a hindrance to scientific openness and reproducibility, it could be harming the very conduct of science.

2. What Does Free and Open Source Mean?

The Free Software Foundation, which champions the use of free software, defines free software as respecting a user's freedom to run, copy, distribute, study, change and improve the software. The

organization goes on to further state that “when users don’t control the program, the program controls the users. The developer controls the (proprietary) program and through it controls the users” [8]. Having access to the program’s underlying source code is a precondition for the above.

3. The Benefits of FOSS and the OpenDocument Format

There is increasing interest in the scientific community of the benefits offered by FOSS and also increasingly “open source hardware” to make scientific tools [9,10]. Free open source operating systems such as the GNU/Linux system and BSD variants offer open, stable and scalable features. These features include parallel computing [11,12], multi-core processing [13], and portability to small and embedded devices [14]. FOSS operating systems run most of the world’s web servers. They are responsible for high performance scientific computing at centers such as CERN [1,15], where mathematical simulations are carried out under Linux environments using open source tools such as GNU Octave and Scilab, largely supplanting the proprietary MATLAB [16].

Even the development process of FOSS resembles the peer-review process of scientific publishing. A software developer creates a piece of software, releases the source code to the community where other developers contribute improvements or voice their concerns over potential flaws. In this way security concerns or bugs in the software are generally fixed more rapidly than in proprietary environments [17]. In an open system, even end users have the ability to audit the underlying code of FOSS and have the final say on what that software does on their computer at any given time. These same benefits extend to open file formats.

Open file formats, such as the OpenDocument (ODT) format, rely on the input of an international multi-disciplinary consortium of standards organizations, information technology firms and even governments [18]. The creation of a usable international document standard that is open, free, backwards compatible and fully documented as to ensure legacy archival, is not only in the interest of data archivists, it is key to the conduct of science for the reasons outlined above. In Table 1 we provide a list of presently maintained and commonly used FOSS word processing packages, many of which use the ODT standard by default. In Table 2 we provide a list of FOSS graphing packages that use non-proprietary file formats by default.

Table 1. Overview of a selection of presently maintained and commonly used free and open source software (FOSS) word processing packages.

Word processing package	Native file formats	Operating system availability	Software license	Latest stable release	Cost	Obtainable from	Compatible with MS Word .doc format *
LibreOffice (Suite)	ODF	Linux, Mac OS X and Windows. Source code available for other platforms	LGPLv3	4.0.3 (2013-05-09)	Free	The Document Foundation (www.libreoffice.org)	Yes
Apache OpenOffice (Suite)	ODF	Linux, Mac OS X and Windows. Source code available for other platforms	Apache License v 2.0 and LGPLv3 (legacy versions)	3.4.1 (2012-08-23)	Free	Apache Software Foundation (www.apache.org)	Yes
NeoOffice (Suite)	ODF	Mac OS X	GPL	3.3 (2012-08-22)	Free (previous versions); latest version is donation-ware	NeoOffice (www.neooffice.org)	Yes
LaTeX (Application)	TEX	Linux (TeX Live), Mac OS X (MacTex) and Windows (proTeXt). Source code available for other platforms	LPPL	(2011-06-27)	Free	LaTeX (www.latex-project.org)	No
AbiWord (Application)	ABW	Linux and Windows. Source code available for other platforms	GPL	2.8.6 (2010-06-13)	Free	AbiWord (www.abisource.com)	Yes
Calligra	ODF	Linux and Windows. Source code available for other platforms	GPL, LGPL	2.6.2 (2013-03-13)	Free	Calligra (www.calligra.org)	Yes

* Can read and export files created by MS Word; ** Files of the open document format (ODF) include .ods (spreadsheets); .odt (text files); .odp (presentations).

Table 2. Overview of a selection of presently maintained and commonly used FOSS graphing applications.

Graphic package	File format	Operating system availability	Software license	Latest stable release	Cost	Obtainable from
R Statistical Environment	Reads .Rdata; SPSS, Stata, Excel and CSV. Exports to open image file formats for graphics	Linux, BSD, Mac OS X and Windows. Source code available for other platforms	GPL	3.0.1 (2013-05-16)	Free	R Project for Statistical Computing (www.r-project.org)
Gnuplot	Allows export of SVG drawings	Linux, BSD, Mac OS X and Windows. Source code available for other platforms	Permissively licensed	4.6.3 (2013-04-18)	Free	Gnuplot Homepage (www.gnuplot.info)
Gretl (Gnu Regression, Econometrics and Time-series Library)	ASCII, CSV, databank, EViews, Excel, Gnumeric, GNU Octave, JMulTi, ODF Spreadsheet, PcGive, RATS 4, SAS xport, SPSS, and Stata files. It can export to GNU Octave, R, Comma Separated Values, JMulTi, and PcGive file formats.	Linux, BSD, Mac OS X and Windows. Source code available for other platforms	GPLv3	1.9.12 (2013-03-15)	Free	Gretl Homepage (www.gretl.sourceforge.net)

4. Conclusions

This manuscript was prepared entirely using FOSS (Linux Mint, LibreOffice, Zotero). Unfortunately during the final preparation and submission for peer-review, we were required to export the final manuscript into the required DOC format. In a more open submission process, the final step of exporting to a proprietary file format, would have been prohibited for the reasons that we outline in this communication. It is hoped that through greater awareness of the current problem, that more journals might offer authors the ability to submit their works in a documented, ISO standard file format such as ODF which is accessible to all now and will be in the future. If science depends on openness and the collaborative pooling of ideas to solve big questions, then why should the very communication of scientific results be dictated by the use of closed corporate software models? In order to change the current paradigm, a critical mass of researchers have to be addressed via general science and engineering journals with the aim of informing them of the importance of using (and requesting) a wider adoption of formats in their work and publication. Publishers themselves also need to be addressed and made aware of the importance of requesting and even requiring open file-formats from authors. Newer versions of Microsoft Office will write to ODF and plug-ins are available for older versions. This makes a stronger case for publishers accepting ODF, since those who choose to use Microsoft products may continue to do so. This also prevents the opposite of what is being argued in this paper from occurring, that is locking out existing proprietary software users in favor of those using non-commercial tools. The ODF format is all encompassing, whereas Microsoft formats are not.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

The authors would like to thank the three anonymous reviewers for their helpful comments and constructive feedback on the draft versions of the manuscript.

References

1. Tchanchaleishvili, V.; Schmitto, J. Preparing a scientific manuscript in Linux: Today's possibilities and limitations. *BMC Res. Notes* **2011**, *4*, 434.
2. International Journal of Clinical Practice—Journal Information. Available online: https://www.blackwellpublishing.com/ijcp_enhanced/submit.asp (accessed on 4 November 2012).
3. Kesan, J.; Shah, R. Open Standards in Electronic Governance: Promises and Pitfalls. In Proceedings of the 2nd International Conference on Theory and Practice of Electronic Governance, ICEGOV '08, Cairo, Egypt, 1–4 December 2008; pp. 179–182.
4. Stallman, R. We Can Put an End to Word Attachments. Available online: <https://www.gnu.org/philosophy/no-word-attachments.html> (accessed on 4 November 2012).
5. Griffin, B. An Introduction to Viruses and Malicious Code, Part One: Overview | Symantec Connect Community. Available online: <http://www.symantec.com/connect/articles/introduction-viruses-and-malicious-code-part-one-overview> (accessed on 4 November 2012).

6. Goldberg, J. MS-Word is {Not} a document exchange format. Available online: <http://www.goldmark.org/netrants/no-word/attach.html> (accessed on 4 November 2012).
7. Steingold, S. No to Proprietary Binary Data Formats. Available online: <http://www.podval.org/~sds/data.html> (accessed on 4 November 2012).
8. Free Software Foundation What is free software? Available online: <https://www.gnu.org/philosophy/free-sw.html> (accessed on 20 May 2013).
9. Li, D.; Parkhurst, D.J. openEyes: An open-hardware open-source system for low-cost eye tracking. *J. Modern Opt.* **2006**, *53*, 1295–1311.
10. Liu, W.; Winfield, A.F.T. Open-hardware e-puck Linux extension board for experimental swarm robotics research. *Microprocess. Microsyst.* **2011**, *35*, 60–67.
11. Hoffman, F.; Hargrove, W. Parallel computing with Linux. *Crossroads* **1999**, *6*, 23–27.
12. Valiev, M.; Bylaska, E.J.; Govind, N.; Kowalski, K.; Straatsma, T.P.; van Dam, H.J.J.; Wang, D.; Nieplocha, J.; Apra, E.; Windus, T.L.; *et al.* NWChem: A comprehensive and scalable open-source solution for large scale molecular simulations. *Comput. Phys. Commun.* **2010**, *181*, 1477–1489.
13. Boyd-Wickizer, S.; Clements, A.T.; Mao, Y.; Pesterev, A.; Kaashoek, M.F.; Morris, R.T.; Zeldovich, N. An Analysis of Linux Scalability to Many Cores. MIT web domain, 2010.
14. Kshetri, N. *Increasing Returns and the Diffusion of Linux in China*; Social Science Research Network: Rochester, NY, USA, 2007.
15. Bahyl, V.; Chardi, B.; van Eldik, J.; Fuchs, U.; Kleinwort, T.; Murth, M.; Smith, T. Installing, Running and Maintaining Large Linux Clusters at CERN. Available online: <http://arxiv.org/abs/cs/0306058> (accessed on 21 June 2013).
16. CERN To all users of mathematics tools at CERN. Available online: <http://mathtools.web.cern.ch/node/13> (accessed on 20 May 2013).
17. Nakakoji, K.; Yamamoto, Y.; Nishinaka, Y.; Kishida, K.; Ye, Y. Evolution Patterns of Open-Source Software Systems and Communities. In Proceedings of the International Workshop on Principles of Software Evolution; IWPSE '02; Orlando, FL, USA, 19–20 May 2002; pp. 76–85.
18. Weir, R. OpenDocument format: The standard for office documents. *IEEE Internet Comput.* **2009**, *13*, 83–87.