Janeway Dev Team

4 Followers About

Anonymizing submissions to Janeway

Janeway Dev Team Jan 23 · 3 min read

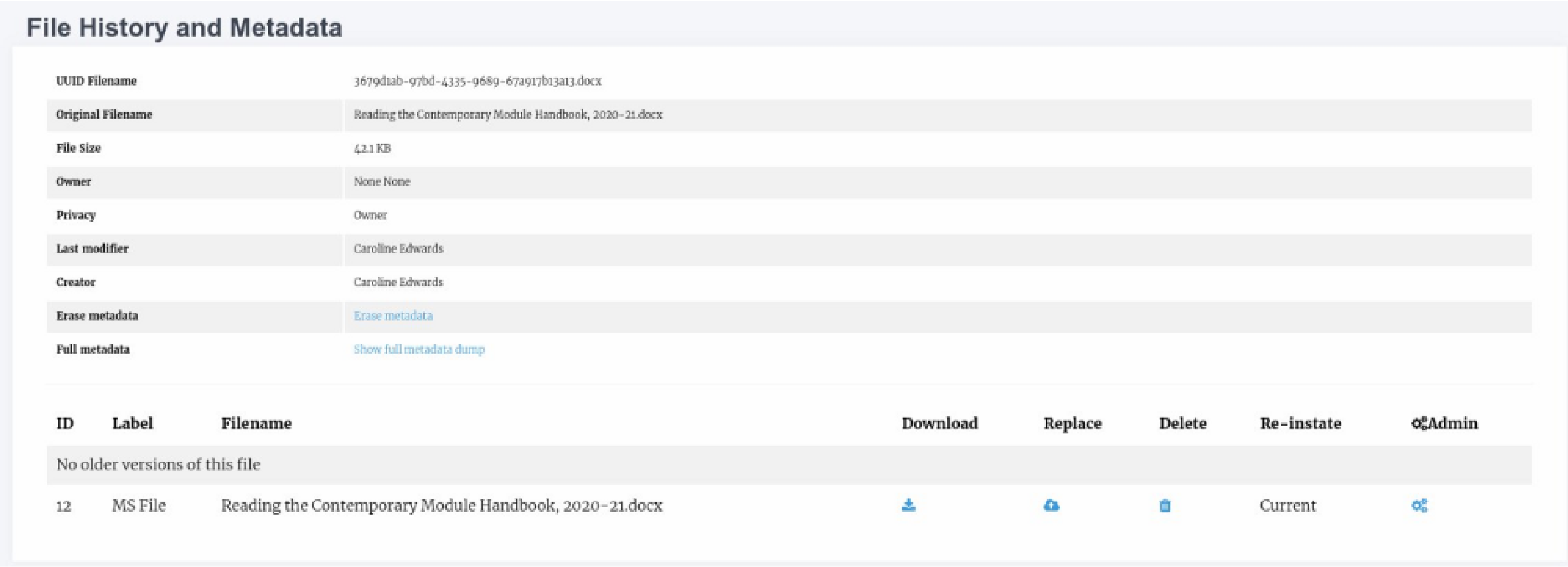
Lots of the journals that run on Janeway, unsurprisingly, use a double-blind peer review process, in which it must not be clear, to various parties, who authored various documents.

People are rubbish, though, at anonymizing their documents. Sometimes they write their details in the file (their name, for instance), other times they cite their own work, sometimes they cite grant numbers that can be traced back to them. It's all very difficult.

Hence, it was of interest, recently, when SAGE announced a $5,000 grant to "build an open source tool that will allow authors to check if a manuscript has been properly anonymized before double-anonymous peer review". The problem is that this was massively over-specced, in my view, for the money. For instance, they wanted to run checks that "author names, emails, author biographies and affiliations do not appear within the manuscript or supplementary information". They wanted to "remove references to funding sources or anonymise the funding source(s), e.g.: The author(s) disclosed receipt of the following financial support for the research, authorship and/or publication of this article: This work was funded by [details omitted for double-anonymised peer review]". This type of full-text search work is incredibly difficult.

We decided, in Janeway, to pursue a more limited anonymization routine that will strip identifying metadata from files and have wanted to do this since September. This works on .docx, .odt, and .pdf files (as well as images). It uses the excellent MAT2 to do its work. We can't, sadly, handle old-school .doc files at the moment. Nonetheless, the basic version of this is now live and kicking!

When these fixes are merged into the main release of Janeway, the File History page now gets a new set of information and options that show if there is identifiable metadata in a file. You can also view a full output of detected metadata by clicking the “show full metadata dump” link. (With apologies to my colleague, Caroline, for using her teaching module handbook as a sample file!)

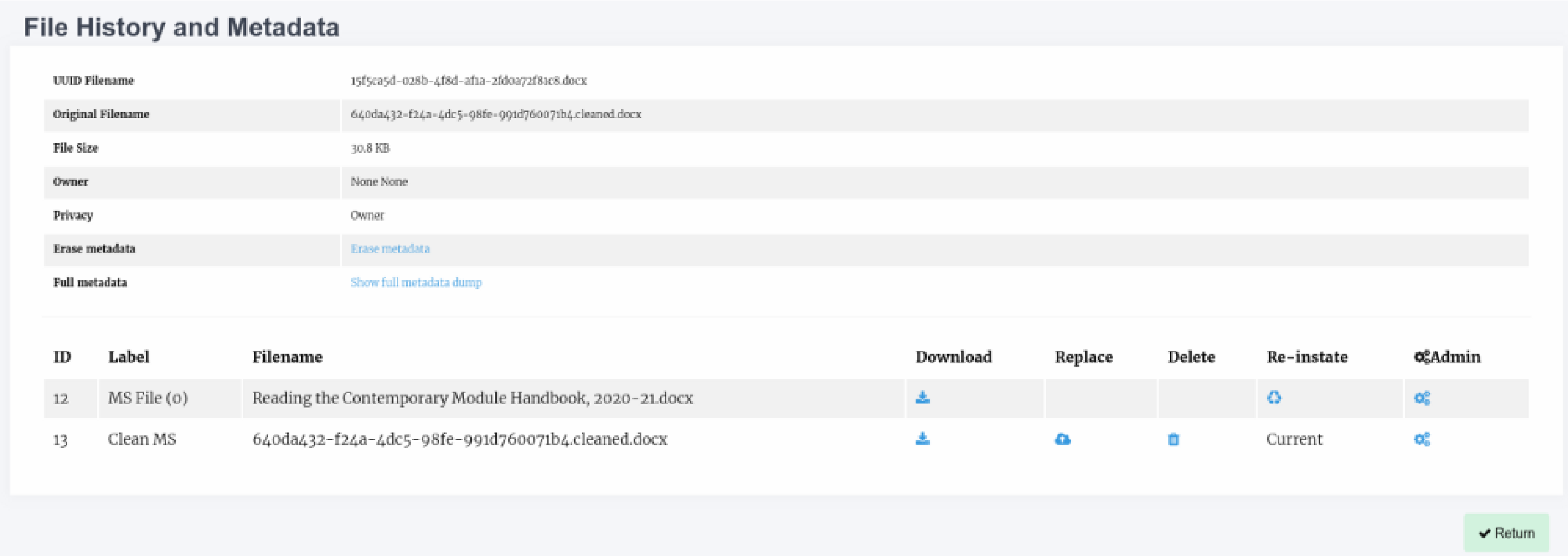


File History and Metadata

UUID Filename	3679d1ab-97bd-4335-9689-67a917b13a13.docx
Original Filename	Reading the Contemporary Module Handbook, 2020-21.docx
File Size	42.1 KB
Owner	None None
Privacy	Owner
Last modifier	Caroline Edwards
Creator	Caroline Edwards
Erase metadata	Erase metadata
Full metadata	Show full metadata dump

ID	Label	Filename	Download	Replace	Delete	Re-instate	⚙Admin
No older versions of this file							
12	MS File	Reading the Contemporary Module Handbook, 2020-21.docx	⤓	☁	🗑	Current	⚙

The “Erase metadata” button here does what you would expect. It attempts to erase metadata in the file, and adds it as the latest revision of that file to the system. As you can see in the below image, the metadata is no longer present and the underlying file has been scrubbed clean, which is much safer for reviewers to see.

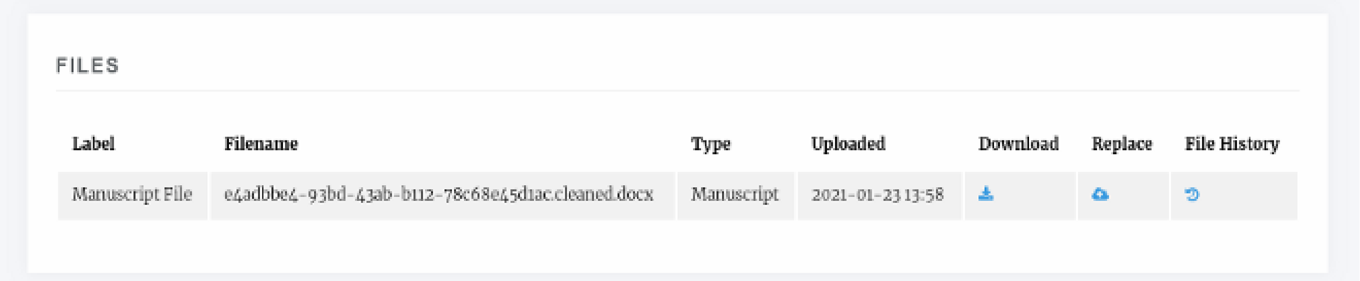


File History and Metadata

UUID Filename	15f5ca5d-028b-4f8d-af1a-2fd0a72f81c8.docx
Original Filename	640da432-f24a-4dc5-98fe-991d760071b4.cleaned.docx
File Size	30.8 KB
Owner	None None
Privacy	Owner
Erase metadata	Erase metadata
Full metadata	Show full metadata dump

ID	Label	Filename	Download	Replace	Delete	Re-instate	⚙Admin
12	MS File (0)	Reading the Contemporary Module Handbook, 2020-21.docx	⤓			↻	⚙
13	Clean MS	640da432-f24a-4dc5-98fe-991d760071b4.cleaned.docx	⤓	☁	🗑	Current	⚙

I’ve then also hooked this directly into the submission system, so that new manuscripts go through automatically. From the author experience side this should be seamless: the manuscript is simply replaced by the anonymized version at upload, with the original file in the history in case of corruption. The below image shows how this appears to an author and an editor, with the cleaned manuscript presented.



FILES

Label	Filename	Type	Uploaded	Download	Replace	File History
Manuscript File	e4adbbe4-93bd-43ab-b112-78c68e45d1ac.cleaned.docx	Manuscript	2021-01-23 13:58	⤓	☁	↺

Overall, this was a fun Saturday hack project. We could do much more. I’d like to upstream some fixes into MAT2 to anonymize comments in LibreOffice. There are also some residual issues with extracting LibreOffice metadata that are odd… I believe it may be a dependency issue in my Docker install, but I’ll investigate that when I can. Nonetheless, all file anonymization is working.

— Martin Paul Eve

Programming Publishing Documents