Janeway Dev Team

4 Followers About

More on detecting anonymization of documents in Janeway

Janeway Dev Team Jan 30 · 2 min read

Last weekend I wrote about how we were detecting anonymization of metadata in Janeway.

This week, this process has been improved further. We now have the ability to run documents through pandoc in order to detect whether there are specific bits of text *inside* the file itself!

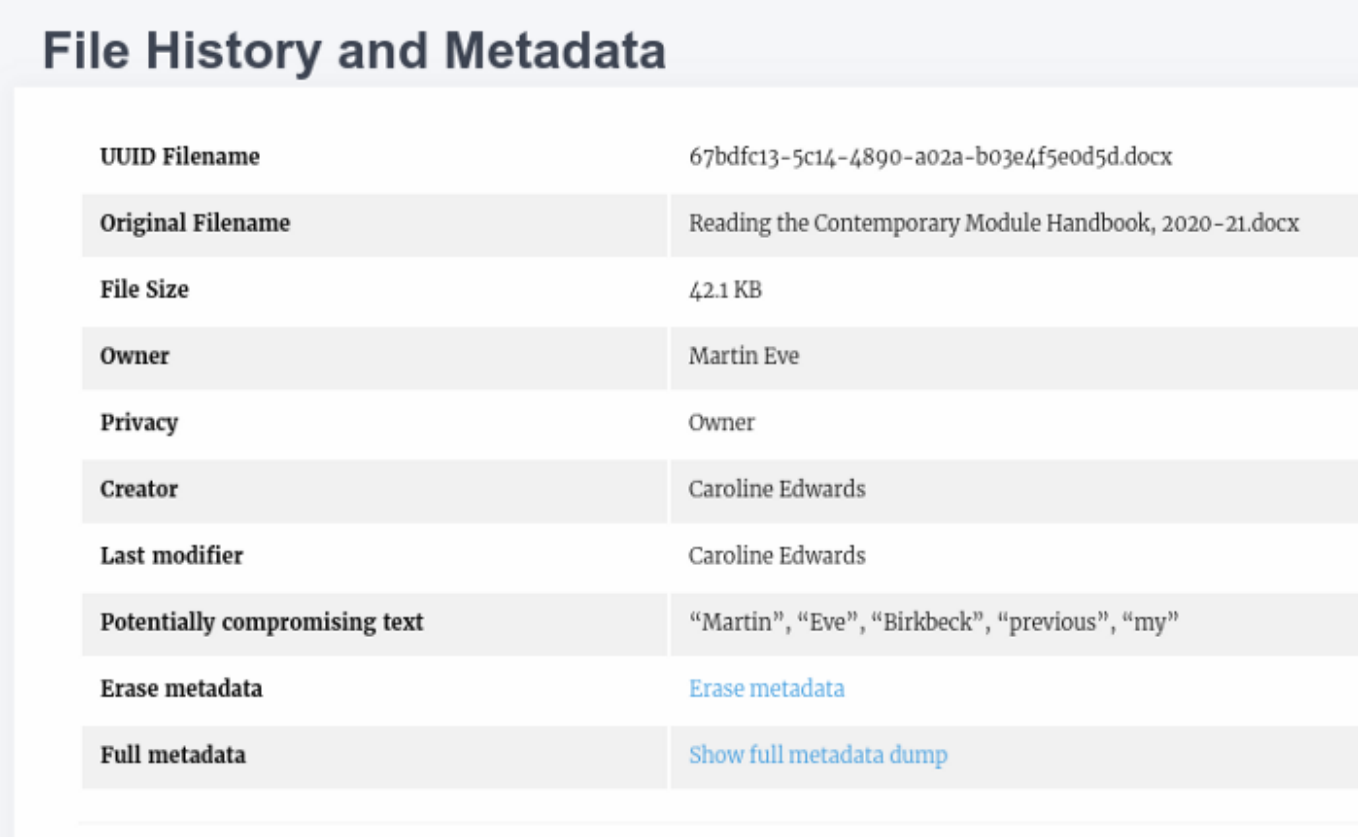
So, for instance, we look specifically for certain terms: the authors' names and institutions; the words "previous" ("my previous work"); words pertaining to "funding". This allows us to set a flag on each of the documents.

FILES

Label	Filename	Type	Uploaded	Download	Replace	File History	Anon?
Manuscript File	Reading the Contemporary Module Handbook, 2020-21.docx	Manuscript	2021-01-24 16:46				
Clean MS	160bfcbe-6252-4cd6-822e-df7c229b4e90.cleaned.docx	Manuscript	2021-01-24 16:46				
Clean MS	e0447862-7574-4cc7-9e9c-cdf4c182746b.cleaned.docx	Manuscript	2021-01-24 16:47				✔
MS File	2021-01-24 09.21.49.jpg	Manuscript	2021-01-30 9:28				?

Showing the document anonymity flags in Janeway.

We can then allow the editor to go in and have a look to see what's causing the flags.



File History and Metadata

UUID Filename	67bdfc13-5c14-4890-a02a-b03e4f5e0d5d.docx
Original Filename	Reading the Contemporary Module Handbook, 2020-21.docx
File Size	42.1 KB
Owner	Martin Eve
Privacy	Owner
Creator	Caroline Edwards
Last modifier	Caroline Edwards
Potentially compromising text	"Martin", "Eve", "Birkbeck", "previous", "my"
Erase metadata	Erase metadata
Full metadata	Show full metadata dump

The document inspector shows us what terms came up that shouldn't be in the document.

Tada! The danger here, of course, is that we might give a false sense of security. This is good for flagging cases where something looks dodgy. But just because it has said it is clean does not 100% guarantee that it is.

Finally, the last step that we need to take is to allow editors to specify *which* words they want to search for, so that we can abstract this out across languages.

— Martin Paul Eve

Publishing Open Source Pandoc Documents Anonymization