

Just DOI it!

Martin Fenner, Gobbledygook

December 22, 2008

With the December 18 issue Nature started to support XMP markup in article PDFs (reported last week on the Nascent blog by Tony Hammond)¹. XMP stands for Extensible Metadata Platform and is a technology to embed metadata in files, including PDFs². XMP was created by Adobe (with XMP support in PDF files since 2001), but is an open standard with backing by others, including Creative Commons³. The Digital Object Identifier (DOI) is the most important piece of information in the metadata, as the DOI provides a link to the journal publisher website where more metadata can be retrieved. XMP support in scientific PDFs is unfortunately still very uncommon and probably hasn't changed much since Pierre Lindenbaum checked last year⁴.

Adding metadata to PDFs seems to be a no-brainer. We have done the same with music (mp3 ID) and photos (IPTC and EXIF) for years and it has been a tremendous help in organizing these files stored on our computers. Unfortunately there aren't too many tools that can extract the DOI or other metadata from the XMP in article PDFs. But I expect more desktop software to support XMP, once XMP support in scientific articles is more widespread. We will then be able to add a journal PDF to our reference manager of choice and have the relevant metadata (including authors, title, journal and issue) automatically filled in. As well as many other creative uses. Until then we need tools like Papers or Mendeley that can extract metadata from PDF files without this XMP information.

For a more technical discussion of XMP in scientific articles, please read the set of blog posts by Tony Hammond^{5, 6, 7}.

fn1. XMP Labelling for Nature

fn2. Adding intelligence to media

fn3. XMP

fn4. Is there any XMP in scientific pdf? No

fn5. Metadata in PDF: 1. Strategies

fn6. Metadata in PDF: 2. Use Cases

fn7. Metadata in PDF: 3. Deployment