Information behind the scenes

Metadata liberates us, liberates knowledge
—David Weinberger[1]

Each time more and more, documents we create and handle are in a purely digital format, and we distribute many of them openly. This leads to proliferation and exchange of documents, the origin and authorship of which is often lost. On the other hand, and most importantly, the management of all information in these documents is done in an automated way: indexing, interconnections among documents, analysis, extraction of keywords... In all these processes, the presence of an identification of the author within the text, which is enough for the human reader, is hardly recognisable by the computerised systems and, hence, such information will be lost. On top of that, not all instructors care to include that authorship information in their educational documents. Along the same line we may include declaration of the conditions of use (licence), a topic which was recently covered in this section.

This month I am aiming, therefore, to devote this space to share with you a few suggestions about the chances we have of including this kind of information in our digital documents, so that it will be available to the trained eye as well as, particularly, to the automated document management systems –and I am not talking here just of software suitable to librarians, but of the mere internet search engine that will be indexing the documents we publish on the web.

The answer to these issues is what has been called **metadata**: information about information. In practical terms for us, it's a question of data that are not displayed in the document directly, but which are stored in a hidden way, even though accessible and, most importantly, are organised in standardised fields that allow their automated retrieval and interpretation. By way of example, names of authors and editors, publisher, date of publication, keywords, licence of use.

It is, then, feasible that we learn two things: how we can read such information and how we can generate or modify it. There is obviously a strong dependence on the software we use to produce each format of document, so I will provide some examples with those more common formats I have access to; you will be able to extrapolate the idea and methods to other applications you might regularly use.

Information susceptible to be part of metadata

Although the list could be expanded in particular cases, we will briefly identify the fundamental items.

Title

Declaration of a specific title, apart from providing basic information, helps the citation, among other things.

Authorship

Is it not only convenient to "sign", or declare the authors of the document, but equally the organisation they belong to –something particularly relevant for educational materials. Depending on the case, editors, reviewers, publisher... may be added.

Date

Sometimes we judge the validity of a document depending on when it was published, and this is one of the piece of data more often absent from electronic documents, maybe for lack of attention. Metadata may include date of publication and, optionally, some extra information like the date of latest update or review.

Licence of use

As we described in the March issue[2], it is highly recommended that our works specify in a clear and, if possible, standard way a declaration of licence. This may basically fall within one out of three categories: *copyright* (all rights reserved), Creative Commons or similar (some rights reserved), public domain (no limitations)

This article was first published, in English and Spanish, in SEBBM Journal, issue 184, June 2015 (the Journal of the Spanish Society of Biochemistry and Molecular Biology). It is republished here after the journal's original website was taken down.

doi:10.5281/zenodo.1841587

Subject or topic

You can include a reference to the course, subject, collection or any kind of organisational entity the document belongs to, as well as a description with more details than the title.

Word processor documents, presentations, spreadsheets

In a very direct way it is possible to include in these document formats the items of information we have listed in the former subsection. You just need to locate the part of the software that allows to access metadata information. On table 1 we collect some pointers, by way of abbreviated recipes, to locate that in several common applications.

Another, additional, possibility is to use a tool that simplifies the addition into the document contents (not in its metadata) of standard licence information; for instance, there are some plugins for office suites that will insert the Creative Commons licence we choose, with text, icon and hyperlink.[3]

Document in pdf format

I personally find it better, barring exceptional cases, to always publish my documents in the pdf format, so avoiding formats specific to a word processor, spreadsheet or presentation software. By my understanding, accessibility is much better with this format, considering the users will employ diverse operating systems, may not have the same software I have, and in addition opening those formats from links in a webpage always comes with a delay, multiple windows..., in general hindering a swift and efficient navigation. On the other hand, the pdf format ensures the aspect of the document will not be altered when opened in other computers, tablets, or even telephones, something not infrequent with other formats and software.

Even though the best known application is the viewer from Adobe (which name has been oscillating between Adobe Acrobat Reader and Adobe Reader), the *Portable Document Format* (pdf) is nowadays an open and free format, standardised by ISO. This adds a vote in favour of my proposal of using pdf for all documents, as a *universal* and *open* format. On the other hand, as a consequence of such standardisation, pdf files may be accessed using software from other companies, and hence there are several viewer and even editor applications; some are free, and some allow to edit and change metadata in a previously created document. This last issue is especially convenient and suited to the matter in question here.

It must first be mentioned that when metadata have been included in the original document (for instance, in the word processor), it is common that the generated pdf version will keep those metadata. This will of course depend on which tool you use for producing pdf documents. As you surely know, the most popular office suites already include the option to save directly as pdf.

Table 1: Access to metadata information in software for document edition.

Program	Access to metadata	Customisable options *
MS Office 2003	File menu > Properties > "Summary" and "Custom" tabs	Tools menu > Options > "Save" > "Prompt for document properties" checkbox
MS Office 2007	Main menu > Prepare > Properties > the "information panel" opens	
MS Office 2010-2013	File menu > Information > Properties (a panel to the right) You may choose "Show document panel" and "Advanced properties"	
LibreOffice (v.4.2)	File menu > Properties > "Description" and "Custom properties" tabs	Tools menu > Options > "Save" > General > "Edit document properties before saving" checkbox
PDF Creator (pdfforge)[11]	Several fields in the dialog displayed before saving	
Adobe Reader	File menu > Properties (you can view them but not modify them)	
PDF-XChange Editor[12]	File menu > Document properties > Description > "Information" and "Additional metadata"	

* If you are prone to forget adding such details, you may activate an option to be prompted every time you save the document

doi:10.5281/zenodo.1841587

To cater for programs that do not have this feature, it is easy to install a "virtual printer driver" that, rather than sending to a printer, will send to a pdf file that is saved to our computer; there are several available out there in internet, both for free and for a fee.[4] During the process of saving the pdf file there may be a second chance to add or edit the metadata.

In any case, after the pdf document has been generated we still have a third chance, since it is possible to edit its metadata in order to check, amend or expand them and then save the pdf file again with the extra information.

Web pages

From its conception, the HTML format used to write web pages includes the META tag,[5] exactly targeted at holding metadata, information about the document which is not displayed by the browser. This is an essential item for the indexing of web pages, although nowadays internet search engines do not rely just on this, but they "read" the full content of the page –even that of pdf documents.

Among the most useful fields in meta tags we may cite author, language of the document, description, keywords, date of creation, estimated lifespan during which the page information is not expected to change.

We may check that hidden information –initially targeted to browsers, search engines, web bots and other automated systems that scan the web– by using some of the features in our browser. For example, in Firefox a right click on the page background will offer "View page information". For lack of a help like this, we may reveal the source code from the browser and locate the meta tags in it.

The way to add this information depends on the way you edit your pages. If you use a webpage editor, try looking in the software for an option like "page properties" or "html tags". If you know some html code, you may insert them manually into the page source code (`<meta>` tags within the `<head>` section). You can also use an assistant or *wizard*, with a more friendly interface, that will generate those tags for you and then you can paste them into your page.[6]

Another kind of metadata are those attached, not to the page as a whole, but to images or other elements embedded in the page. For instance, images inserted in an html document may have added an `alt` parameter that provides an alternative text, a description of what the image contains or represents; this is a feature designed for accessibility, for people with limited eyesight to be able to perceive the information the image conveys –page-reading software, a basic tool for them, reads aloud the alternative text when an image is reached. Another metadata element is the `longdesc` parameter, that directs to a new webpage which provides detailed, extended, information about the image; unlike `alt`, the `longdesc` can also be applied to other kinds of multimedia item. A third container for metadata is the `title` parameter, which is applicable to any kind of item in the page –let it be an image, video, sound, a word, a button...– Its contents are displayed in the browser as a small balloon-like vignette, a *tooltip*, when we hover the pointer over the item in question. Apart from the primary purpose of these three utilities to advise the reader of the page, the search engines also access these fields as part of the page contents.

For a correct use of these parameters in the html tags one needs a knowledge of the html source code, or else to locate these features (once we know they exist and what is their meaning) in our editor of choice.

Images

The most trivial solution is to include information as part of the image, typically using a small icon or logo and some text, in a corner or as a watermark. We might not consider this metadata proper, since it will not be recognised by a system other than human brain, but it does fulfil the purpose of conveying information to users. A noteworthy objection is it may aesthetically harm the image or affect the *cleanliness* of the graphic material.

In regard to images that are part of a webpage, in the previous section we have already commented on several solutions, related to using either alternative or supplemental text.

Finally, depending on the file format of the images and on the software used to edit them, there are ways to insert true metadata within the file. The question is somewhat technical and tied to the specific editing software, so we will just provide a generic description as guidance.

The **jpeg (or jpg)** format permits insertion of additional information apart from that making the

doi:10.5281/zenodo.1841587

image itself; you may not have come to think of it, but you likely have some experience: photographs taken with a digital camera (hey, are there any other kind of cameras?) include, at least, date and time when the picture was taken, camera brand and model, some optical parameters like focal distance, sensitivity of the "film"... If you happen to take the picture using a tablet or smartphone, maybe even your GPS location! All this was made possible by a specification called EXIF data. Photo editing software can read that information, and even the file manager in your computer can (try "file properties"). Taking advantage of this principle, some programs may include information in that section of the file.

The **png** format also allows to include metadata.[7] In a most extreme case of this feature, the Jmol software –a viewer for molecular structures[8]– is able to write png files that, in addition to the snapshot, hold the atomic coordinates data for the molecule, its rendering style... the whole *scene* of the molecular model. Most programs will just see an image file, but Jmol may read the full contents of the file and regenerate the 3D model of the molecule as we had it when we saved the png file. The secret is all the molecular model information has been included in that area reserved for data. We could say that, even more than metadata, that is ultradata! In a more conventional approach, the idea is we can include the information on authorship, licence, description, etc. within the png image file. To do so, we need some image editing software that supports this feature.[9] However, it seems this is still a poorly supported feature.

The **svg** format is officially a format with a future ahead; Wikimedia Commons and Wikipedia have it as the recommended format for images; it has advantages by being an open and vectorial format –therefore scalable without loss of quality. Despite all this, it is not yet extensively established. By being a text-encoded format, structured in fields and tags, and hence very flexible, it allows perfectly well the insertion of all kinds of information.

Video, animation and sound

In these cases it is likely harder to include and check metadata information. This will strongly rely on the software being used for edition of the video or audio. The final file format may also be determinant, since it could or not support that extra information. I cannot get into details, not being an expert and in any case they would be rather specialised. Those among you who are used to prepare this kind of media will surely be able to investigate the documentation and the options of your particular software in search for the way to add metadata.

An added source of trouble is visibility of the metadata to users: even though the video or audio file may include those metadata, it is more unlikely that the player will be able to display them. An additional solution, maybe somewhat *home-made*, but particularly relevant since we are interested on manifesting authorship and licence, is to include such information within the contents of the video or audio recording. Well, I shall retract, it is not so much home-made: Commercial videos most often include the credits both at the beginning and the end of the movie! Following this idea, when we edit the video we can add a brief screen (or better, leave it fixed at the end) with the identity of authors and the licence of use.[10]

As an additional choice, some internet sites dedicated to storage and sharing of audiovisual resources have their own systems for *tagging* the materials and including declaration of licence. This means that even though your video or image may not embed the information –or it does but it is invisible to the user–, when accessing it in the service provider website this will display the information you added while uploading, by filling-in those tags. In the case of licences, it is frequent to find a pre-made list of them; then you don't need to write all the terminology, but just pick your preferred one from the list. As one further consequence of this sorting system, it is often possible for users to search materials in the repository according to the type of licence they have.

I will finish with the wish that this presentation has been interesting, inspiring and enjoyable to you, and with the challenge that, between marking exams, you will find some gaps to go over your materials and add to them that "hidden treasure" of authorship, licence... metadata!

Angel Herráez
Biochemistry and Molecular Biology,
Dept. of Systems Biology,
University of Alcalá
Alcalá de Henares, Spain

  doi:10.5281/zenodo.1841587

References and notes

1. David Weinberger (2008) *Knowledge at the End of the Information Age.* Bertha Bassam lecture, University of Toronto, 7 February 2008. Available at http://bit.ly/1JMwFhK (Accessed 5 May 2015)
2. Angel Herráez (2015) *Mine, theirs... ours?* SEBBM Journal 183, 34-38. Available at ~~http://www.sebbm.com/revista/articulo.asp?id=10819&catgrupo=270~~ doi:10.5281/zenodo.1793992
3. (a) *Creative Commons Add-in for Microsoft Office.* http://bit.ly/1JjKqqh Valid for Word, PowerPoint and Excel, versions 2007 to 2013 for Windows
   (b) *LibreOffice plugin: paste images with credit*. http://bit.ly/1QOv3bz When pasting into a document an image that comes from internet and has metadata, this plugin adds a line with credits and information.
4. To produce pdf files from any installed program, I can recommend *PDF Creator* offered in SourceForge (http://sourceforge.net/projects/pdfcreator/) as well as in the website of the producer, pdfforge (http://pdfforge.org/). It is free, without advertising and provides excellent features. I apologise to those of you that do not use Windows, as I am not aware whether this software is available for other systems; but maybe it's a functionality provided by default in your system.
5. *Meta element* (1 May 2015) In Wikipedia, The Free Encyclopedia. http://en.wikipedia.org/w/index.php?title=Meta_element (Accessed 5 May 2015)
6. Miarroba Networks S.L. (2015) *Generador de Meta Tags.* http://metatags.miarroba.es/ (in Spanish; accessed 5 May 2015)
7. Exiv2 Community (2015) *The Metadata in PNG files*. http://bit.ly/1QON89m (accessed 5 May 2015)
8. *Jmol: an open-source Java viewer for chemical structures in 3D*. http://jmol.org and http://wiki.jmol.org/
9. Some solutions:
   (a) *ImageMagick* (multiplatform) http://imagemagick.org/
   (b) *TweakPNG* (Windows) http://entropymine.com/jason/tweakpng/
   (c) *PNGCommentator* (MacOS) http://echomist.co.uk/software/PNGCommentator.html
10. (a) You may see a simple example at the beginning and the end of the animation at http://bit.ly/sintasaAG
    (b) The Creative Commons Wiki provides some advice on including licences in videos and other media: https://wiki.creativecommons.org/Marking_your_work_with_a_CC_license
11. pdfforge GmbH (2015) *PDFCreator - the free PDF converter tool*. http://pdfforge.org/pdfcreator (There are many other equivalent programs, which let you save a pdf document starting from any program by choosing the Print option)
12. Tracker Software Products (2015) *PDF-XChange Editor*. http://tracker-software.com/product/pdf-xchange-editor