Data management

While the importance of recording documentation should be clear by now, the issue of managing it must be addressed as well. Contrary to lessons 1 and 2 where 'documentation' mainly referred to recording events, in this lesson we will have a look at documentation in a broader sense comprising all project-related data, such as outlines, plans, schedules and correspondence. No matter what your project is about, you will accumulate large quantities of such data. But even with the most complete documentation, availability of information may suffer from a lack of structure and organisation.

We will not deal with databases here because maintaining a database requires profound technical knowledge and should not be undertaken unless you have a skilled person acting as a maintainer. We will, however, deal with files, file formats and directory structures.

Keep the balance

In the first lesson to this module, documentation was said to be the collective memory of a project with the accent on how individual notes must be supplemented with background information before submitting them to the collective pool.

Sometimes, the opposite is necessary, and you will have to discard things that are redundant or not worth documenting. To keep this balance, to achieve completeness while being concise and minimising redundancy, is a delicate but necessary task.

Do not pass every preliminary version of your notes and drafts to the shared storage facility of your project. By keeping these in a private place, you will relieve others from plodding through heaps of data that is of no use to them. As you progress, you will be able to decide what deserves to be kept and published and what can be deleted.

If you still wish (or are obliged) to document the entire development process of a certain document, you may comprise preliminary versions in a separate directory or archive so every visitor can tell that these are versions no longer up to date. It is a good idea to give names to these files preceded by a number or date of creation so as to establish a chronologic order right in your file manager. A table of contents in the shape of a plain text file or other simple document commenting the progress made on each version is also a good idea. In open-source software development, this is known as a *change log* and is compulsory when releasing the source files of an application.

Try to avoid having more than seven items on any level of a directory tree. This 'magic' figure, first established in fairy-tales and other folk literature, later confirmed by psychologists, is a good rule of thumb for the number of things we can, as humans, oversee at a glance and keep in short-term memory.

File formats

TEXT DOCUMENTS

For texts that are not supposed to be changed anymore, do not use any other than the **PDF** (portable document) format. PDF is not editable (although it is technically possible to forge a PDF document), so it lends itself to this task more than the built-in file format of your word processor. Also, PDF can be viewed using free-of-charge applications by everybody around the world, regardless of their computer brand, operating system or special software installed. If your word processor does not have a built-in PDF export facility, you can usually create a PDF by printing your document to a file.

You may (and sometimes must) still keep your source files. Consider placing them in a sub-directory or archive as described above.

Sometimes it is important to discern file formats. Icons are often used equally for different formats, but you can configure your file manager so as to display file extensions which is not always preset by default.

IMAGE FILES

With image files, there are several formats in use of which Table 1 gives an overview.

Table 1: Common image file formats

Name	Description	Applications	Compression	Transparency
BMP	Microsoft Bitmap	(outdated)	none	no
GIF	Graphics interchange format	(outdated)	none	yes (fully transparent or fully opaque)
JPEG	Joint photograph experts group	Photographs	lossy	no
PNG	Portable network graphics	Drawings, computer graphics	lossless	yes (semi-transparency is possible)
TIFF	Tagged image file format	Maps, also photographs	lossless	no

PHOTOGRAPHS

JPEG is the choice format for photographs; it has been developed for that purpose. JPEG uses a compression that is able to reduce the file size to less than 10 percent of what non-compressed image data would require. This advantage, however, comes at a loss of quality. Compression is done by combining neighbouring pixels of approximately the same colour, and if you view a strongly compressed JPEG, you will see the effect in the shape of a pattern resembling contour lines on a map.

Your camera will usually choose an appropriate compression grade. But if you post-process your photographs, be aware that the compression mechanism is applied every time you save the file, thus reducing image quality over and over again. For intermediate versions, choose a lossless storage format, either the built-in format of your image processor, or TIFF, or PNG.

COMPUTER GRAPHICS

On the other hand, if you wish to create icons or similar computer graphics, or capture drawings, there is nothing worse than the JPEG format you could use because the compressing mechanism will blur contours and rasterise colour gradients. To preserve quality, you need an image format with a lossless compression, and you will probably want to create transparent and semi-transparent pixels in order to use your graphic against different backgrounds. There is no other popular format than **PNG** that fulfills all these requirements today.

Archiving

If you have large structures of data you will rarely touch, consider creating an **archive**. An archive, the most common type represented by the **ZIP** format, is a collection of files and even entire directory structures which can be handled by your computer as a single file. Optionally, on archiving, a mathematical (and lossless) compression mechanism can be applied to reduce size where data is redundant. Archives are also very handy to attach collections of data to an email. Tools for packing and unpacking ZIP archives are available free of charge for all major operating systems.

File naming

Many people assume that they cannot use special characters in a file name, but most of the restrictions of the past are now obsolete. There are some forbidden characters, and language-specific letters can still be a problem when switching operating systems. Apart from these restrictions, there still remains a fair variety.

Here is a (non-comprehensive) list of characters you can safely use with all major operating systems, including those usually running on internet servers:

Table 2: Characters that can be safely used in file names

Character	Sample	Character	Sample
Latin letters	`A-z, a-z`	Braces	`{[()]}`
Numbers	`0-9`	Hyphen	`-`
Period	`.`	Plus sign	`+`
Comma	`,`	Equality sign	`=`
Semicolon	`;`	Ampersand	`&`
Underscore	`_`	Hash	`#`

With these devices at your command, it is easy to create descriptive file names. You may use the following scheme: **`yyyy-mm-dd.title.author.extension`**

The date of creation (first item) is given in reverse order, so the alphabetic sorting method of your file manager will also provide a chronologic order. Always insert a separator between date components for better reading and to avoid ambiguities.

File metadata

Sometimes it can be necessary to refer to file metadata, that is, superordinate information on a file which is independent of its contents. File managers can usually pop up a 'Properties' window from a context menu which will display file metadata. Some general properties are independent of the file type, although their meaning may depend on the operating system, or rather, the underlying file system used:

Table 3: What file metadata can tell

	Windows	UNIX/Linux	Mac OS
Date of creation	When was the file last moved or renamed?	(not displayed)	When was the file saved to this storage volume?
Date of last modification	When were the contents of the file last modified?		
Date of last access	When was the file last opened?		

The 'date of creation' field, if it exists, is problematic as the question when a file was created is difficult to decide: Under Microsoft Windows, when you move or rename a file, the result is considered a new file and the date of creation will be reset to the current date and time. Mac OS preserves the original date of creation as long as you do not move the file do a different storage volume. Most UNIX and Linux systems work around the problem by not displaying the information at all, although they keep it internally.

Nevertheless, you can always evaluate the **date of last access** and **date of last modification** to get information when a file was last viewed or edited, respectively.

Individual file types can also carry a lot more meta information:

- JPEG photographs contain various metadata recorded by the camera. You can usually view this on a special page of your file manager's property window.
- Audio and video files have their own system of 'tagging' to relate them to artists, genres, albums etc.
- Text documents of all formats, as well as internet pages, hold information as to their author and title. To exploit these features (or not to tag a document with your name inadvertently), you should learn how to edit document metadata in your word processor.