Exploring file metadata

The purpose of this exercise is to practise the use of metadata in order to classify unsorted files. Students are presented a directory containing 27 unsorted files (PDF documents and photographs) from a project documentation whose names are partly obscure. Their task is to relate every single file to one of four events.

The students may first be discouraged by the lack of order in the start-up directory. This is exactly the experience one makes when coming across ill-organised data collections. To overcome that feeling and succeed in establishing a classification is the purpose of the exercise.

Requirements

- Duration: approx. 15 min,
- Computer workstations with free hard disk space of 1 Mbyte for every group.

Preparation

Copy the contents of the **`exercise_data_4-2`** package to each of the computers. You may also copy the exercise sheets to hand them to the students or use the **`exercise_slides_4-3`** presentation.

Schedule

1. Divide the students into groups or assign the task to everybody individually.
2. Hand out the exercise sheets or explain the situation.
3. Students are to solve the task within 15 minutes. If a student or group is by no means able to classify the images, you may give some hints based on the 'Solution' section below.
4. Whoever thinks to have found the correct solution is invited to explain it. The focus should be on how the sorting was done.

Solution

A complete solution is not to be achieved without referring to metadata, although evaluating file contents can corroborate your findings. There are three vital rules concerning metadata, partly also set out in Handout 4.3:

- With Microsoft Windows, the 'date of creation' indicates the last time a file was copied and is of no value; neither is the 'date of last access'. The 'date of last modification' refers to the last change in file content and preserves the creation date of a file if no further processing has been applied.

- The only reliable method of dating photographs is by their EXIF capture date. The date of last modification of the file may refer to an image-processing operation. However, if the date of last modification coincides with that of one of the events, it is very likely to be correct.
- With PDF documents, the date of last modification is to be evaluated with caution because it refers to PDF export from a text document which may have been written earlier.

It is obvious that the first step must be to have a look through the text documents in order to get an overview of events. There are four events documented:

- **First meeting**, 14th December 2011,
- **First excursion**, 4th April 2012,
- **Second excursion**, 16th April 2012,
- **Final meeting**, 8th June 2012.

Corresponding folders should be created to move individual files to in the process of sorting. The classification itself is best done the following way:

1. The two images named **`IMGxxxxx.JPG`** can be related to the **first excursion** on 4th April 2012 by their EXIF capture dates. The **`PICTxxxx.JPG`** files also have EXIF metadata, pointing to the **first meeting** on 14th December 2011.
2. The images named **`excursion_x.JPG`** can be classified by their 'last modification' dates as belonging to the **second excursion** of 16th April 2012. The same goes for those named **`meeting_x.jpg`** whose modification dates hint to the **final meeting** of 8th June 2012. It is fortunate that none of these images have been post-processed because they contain no EXIF data, except for the camera make and model with the **`excursion_x.JPG`** set.
3. The remaining three pictures have obviously been post-processed (note the unusual aspect ratio), so their date of last modification is of no value. However, there are other methods to classify them:
   - The **`proposals.JPG`** picture can be related to the **first meeting** on 14th December 2011 by its EXIF capture date.
   - **`In the woods.JPG`** does not seem to offer any clue. But, fortunately, if you browse through image metadata, you will find that at least the camera make and model have been recorded, namely CANON and SUPER 105. You will find the same specifications with the other images from the **second excursion** of 16th April 2012, and that will do for a secure classification, as this camera was not used at any of the other events.
   - Finally, **`By the lake.JPG`** does not even give a camera model and it is exactly that fact which tells us that it originally belonged to the meeting_x.JPG collection documenting the **final meeting** on 8th June 2012.
4. There remains a short text document named **`addendum.pdf`**. This time, dates are of no value: the PDF was obviously created later. But looking at PDF metadata (click File > Properties in your PDF viewer), will reveal that the author is Susan Sauer who has written the minutes of the **first meeting** only, so her addendum will most likely refer to this event.

This all provided, the correct solution is as follows:

First meeting, 14th december 2011	First excursion, 4th April 2012	Second excursion, 16th April 2012	Final meeting, 8th June 2012
meeting.pdf addendum.pdf	excursion report.pdf	Second excursion.pdf	Meeting minutes 8 Jun 2012.pdf
PICT0009.JPG PICT0011.JPG PICT0021.JPG PICT0025.JPG proposals.JPG	IMG00012.JPG IMG00015.JPG	excursion_1.JPG excursion_2.JPG In the woods.JPG excursion_4.JPG excursion_5.JPG	meeting_04.JPG meeting_08.JPG meeting_09.JPG meeting_10.JPG meeting_11.JPG meeting_12.JPG meeting_15.JPG meeting_16.JPG By the lake.JPG meeting_22.JPG

Exercise sheet

THE SITUATION

You have joined a planning project in its final stage. Your predecessors, **Susan Sauer** who left the project by the end of 2011 and **Ted deBaer** who replaced her at the time, have left a directory with unsorted pictures and PDF documents documenting **four different events** during the planning process.

The photographs were taken using a different camera in each of the events, but always the same camera was used throughout one event.

YOUR TASK

Your task is to **relate each of the files to one of the events** by moving them to different folders you will have to create. In order to do so, you can evaluate both contents and metadata of the files.

Some of the images may have been post-processed so their date of last modification will not tell when the photograph was taken.

However, the bare fact that all images are of very small dimensions (less than a digital camera would supply) should not be taken as a hint to post-processing. The scaling down has been done for the purpose of this exercise only, and the resulting traces in file metadata have been carefully wiped out.