Digital Book Production in PDF for Myanmar Manuscript

1. Introduction

As shown in Fig. 1, the manuscripts in the U Pho Thi library in the Sadhammajotika Monastery, Thaton, Myanmar, are on palm leaves (approximately 6 × 22 cm) between teakwood front and back covers. At this monastery, there are about 780 manuscripts with some 900 texts. In this study, a digital book is created by cropping images of the wooden covers, the palm leaves with the text, the archive data of the manuscript (vertical and horizontal sizes, the overall height of the manuscript, and its package condition if applicable, etc.) using the photographs of the manuscript. The cropped images are put into PDFs as this format is very versatile (see Fig. 2).

Fig. 1

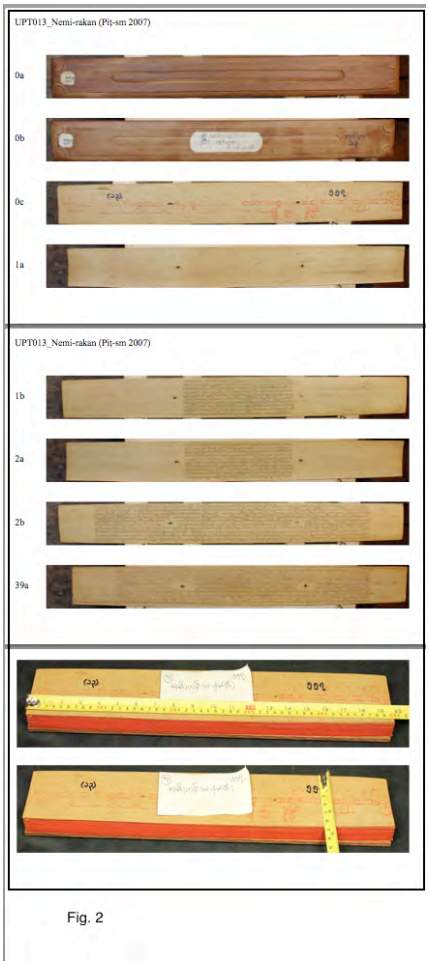


Fig. 2

In order to create good quality PDFs efficiently, it is essential to take high resolution photographs that can be stored and easily accessed. For the first photos taken at the beginning of the project, the front cover and the back cover, or the back (verso) of a palm leaf and the front (recto) of the next leaf were placed on the upper part and the lower part of a wooden easel (see Fig. 3(a)). Later, in order to improve the PDFs, photos were taken using a background of black fabric. Finally, we discovered that using a colored background (green or blue) improved the cropping process. In either case, the photo data of a manuscript is stored in one folder for each manuscript (Fig. 3(c)). We will assign the number 1 to the photo of the front cover and the back cover, followed by the numbers of the leaves beginning with number 2. Next, we created the nested folder "UPTSD" inside the manuscript folder and saved the archive data of the manuscript as in Fig. 3(b) inside this nested folder (see Fig. 3(c)).

In this study, a PDF is created by converting the high resolution raw photo data in JPG format. The data size of a photo is a few megabytes. For example, manuscript number 13, UPT013_Nemi-rakan (Piṭ-sm 2007), was saved using 43 photos, whose size is about 200MB. The manuscript folder name is taken from the name of the manuscript. "UPT" is the abbreviation of the library. This is followed by the manuscript number and the name of the text. When making the PDF, this folder name is used as the PDF header (see Fig. 2 and Section 3).

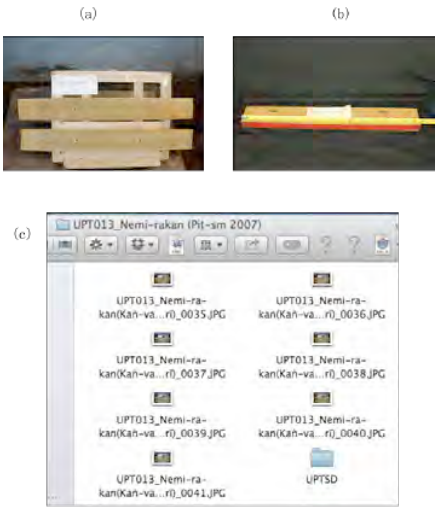


Fig. 3

In this study, we produce the PDF in two steps from a manuscript folder: cropping process of the necessary information only from the photo data, and the PDF production process. In the photo, the palm leaves or the front cover and back covers are on a background. Since the background is unnecessary, of course, cropping the palm leaves,

etc., is the first step in the PDF production program. By executing this process, the same resolution as the raw data is maintained and the size of the trimmed data is reduced to about one fourth the size of the raw data (53MB for the manuscript number 13). By reducing the data size, the memory size of the manuscript database was reduced as well. Furthermore, the data size was kept within the reasonable limits so that researchers in this field can use the PDFs on their personal computer. The resolution of the images makes it possible to view the text at the same size, or an even larger size, as the original manuscript.

There are some 900 texts, each of which contains the photo data of the leaves numbering between a few dozen to several hundred. Since the processing time of the images is quite long, and the work is very complicated, it is essential to make programs that will automate the process. In addition, there may be problems that are the result of problems associated with the setup used when taking the photos. Since placing two palm leaves on the photo holder has to be done manually, the position of the leaves in each photo will be shifted slightly. This should not be a problem so long as these shifts are located within the acceptable range of the cropping tool, but once it goes beyond this range, it is necessary to make corrections in the photographs prior to the cropping process. We need to make a program to handle this correction process since the routine work must be carried out many times to make the corrections. In order to proceed efficiently, it is necessary to minimize the manual work as much as possible; the tasks should be processed automatically (or at least a semi-automatic processing).

So far, our group has developed several tools for analyzing Pāli literature based on the Macintosh operating system. In this study we need to develop the tools associated with image processing and file processing. Fortunately, since Macintosh computers are suitable for this work, we will continue to use these machines for this study.

The cropping tool and PDF production tool are the most important apps in this study, and they are controlled by AppleScript. Assuming other research groups in the same field will use our programs, we employ a number of standard tools that are commercially available and available as freeware:

❶ An app for cropping the front and back covers as well as the palm leaves is encoded in one tool, and cropping the archive data of the manuscripts is encoded in another program. This process crops images according to a selected area in the photo.

❷ As for the photos with a black, blue, or green background, removing the color background is executed using the cropped data of ❶. This process crops images according to the background color.

❸ A PDF of the cropped images is created by another app that uses the data produced by ❶ or ❷.

To facilitate our explanation, the tools to be used have been numbered.

Procedures for ❶ – ❸: As for commercially available tools, Adobe PhotoShop (Element) is readily available and well known for its image processing, Adobe PDF (the creation tool but not the Reader) and Microsoft PowerPoint are also used. We can process the JPG photographs efficiently by using these tools. In order to use the results on many different platforms, we will produce a digital book in PDF format. For the second procedure ❷, a freeware program written in AppleScript has been adjusted to control PhotoShop.

❹ As for freeware, the PictCatalogX.app is used for displaying a list of the photo data.

❺ As mentioned above, since the leaf position in each photo varies, photo data needs to be checked and adjusted prior to using the cropping tool. For this process we will use CocoaSlideShow.app, which is freeware. This tool is also used to check the leaf images after they are cropped.

In addition to them, several freeware programs available on the internet which are written in AppleScript have been modified so that they will play their assigned roles as follows:

❻ When the entire file has been rotated by the data check prior to using the cropping tool, you should use the rotation correction tool, PhotoSHopMyRot.app.

❼ In cases where extra photos are taken because some leaves were skipped over, or in cases where photos

need to be retaken, the photo file number may be skipped over. In such cases, a program to renumber the file numbers, RenumberingJPEG.app, should be applied. The tool for changing the file name, FileNameChange.app, is usually used in combination with the file renumbering tool.

❼' The cropped data file numbers may deviate because the front and back covers might be different in each manuscript, or because of additional leaves that are not numbered in the manuscript (for example, blank leaves used to identify sections within a manuscript, etc.). In such cases, the correct numbers are re-established by using the two programs above: RenumberingJPEG.app and FileNameChange.app.

When producing a digital book in PDF format from the manuscript photo data, the work required can be very complex — for example, selecting and using the appropriate tools mentioned above. Since we must process nearly 900 texts, production of the digital books would require tedious work over a long period if we were not able to improve each step. In order to avoid many errors which may occur in this work, each tool is combined into one application (.app), and it works by dragging and dropping one folder or collection of files onto this app. As a result, a complex work can be reduced to dramatically improve the work efficiency of an accurate digital book production.

The whole program set with some sample data are summarized in the figure of Appendix (see blow, p. 22).

In this paper, we will first discuss the steps to follow for cropping the images, and then we will explain the steps to follow in order to produce a digital book as a PDF.

2. The cropping process

Since cropping leaves (tool ❶) involves image processing, it can take a very long time. For manuscript 13 which has 43 leaves, cropping the images can take about 4 minutes. The time needed for processing the data depends on the number of leaves. As mentioned in the introduction, since the images in the photos of the manuscripts are not homogenous due to the leaves being arranged by hand when the photos were taken, it is impossible to crop the photos with high precision in a single process. On the other hand, performing many cropping processes for a single manuscript by trial and error would be a major waste of time. In order to avoid this, we should use the following preprocessing procedures: we first check the photo data with CocoaSlideShow.app prior to using the cropping tool. Next, we use tool ❻ when rotation correction is required. If there are many types of non-homogenous images, one folder should first be divided into sub-folders with each sub-folders containing photos that use the same arrangement. Secondly, we apply the cropping process for each sub-folder. And finally, we group these cropped photos in one folder. By using this final result, production of the digital book can be accomplished in one single process.

First, we will discuss how to do the preprocessing procedures, followed by the method of using the cropping process. Last, the transparency processing of the black or green color background will be explained. By using these cropped results, one digital book in PDF format is going to be created (book production).

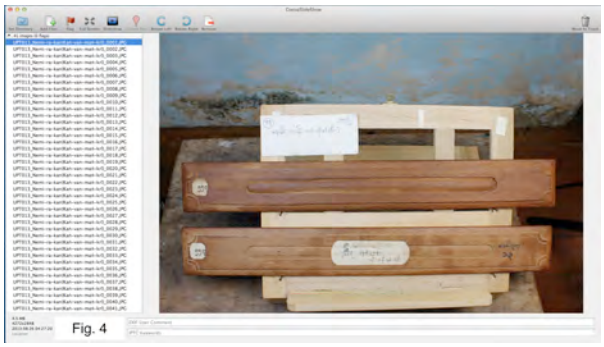


Fig. 4

2.1 Preprocessing: CocoaSlideShow.app

After dragging and dropping manuscript folder Fig. 3(c) into application program, CocoaSlideShow.app, a window will open as in Fig. 4. The total number of the photo files in the folder and individual file names are shown in the left-hand bar, and an image of the selected file is shown in the window on the right-hand side. We can move through the images by scrolling up and down the file names in the side bar. We can see information about the leaves in this step (rotation status; file

names that belong to the group that are in the same location). Based on this data, the files should be divided in advance as required.

2.2 Rotation Correction: Tool for ❺ (PhotoSHopMyRot.app)

If you find that the rotation of the leaves need to be corrected after following the steps in subsection 2.1 (that is to say, by dragging and dropping a folder of the manuscript onto the program PhotoSHopMyRot.app) we can obtain the window for correcting the rotation as shown in Fig. 5. After designating the rotation angle (top left in Fig. 5), a dialog box will enable you to confirm the results (top right). If you click OK, the program will run. Finally, you must click OK in the third dialog box (the lower box). The figure at the bottom shows the result where it was rotated one degree clockwise: we can see that the image of the canvas has been tilted slightly from the vertical alignment. A rotation correction with about 1 degree to the right was sufficient for the manuscript data to be used in this study.

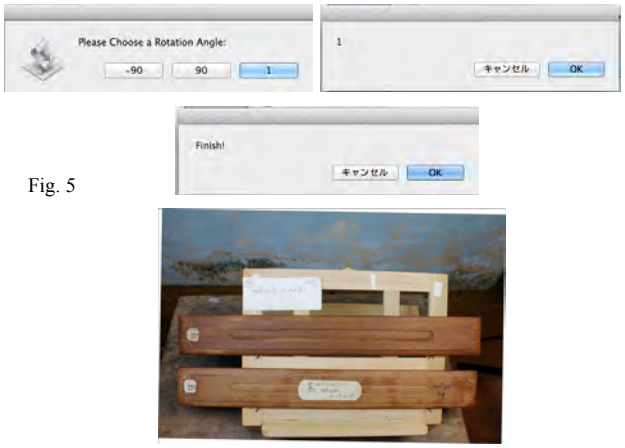


Fig. 5

2.3 The File Renumbering Process: The tool for ❼ (RenumberingJPEG.app)

In case of irregular file numbering, the renumbering process is required: drag and drop the manuscript folder onto the program called “RenumberingJPEG.app”. When the program runs, the left top dialog in Fig. 6 will be shown. Enter an appropriate identification name for the file, and finally click the OK button (“UPT_” will appear in the option). Next, the right top dialog box appears, so enter an appropriate number corresponding to the first file (number “1” is used in the option). When we apply the file renumbering program to the folder of manuscript number 13, we can obtain the result shown at the bottom of Fig. 6; this is displayed by using CocoaSlideShow. (If we use “UPT_” and number “1”, the files in the folder are numbered in order as UPT_0001, UPT_0002, etc.). The execution time for this folder is less than 1 second. The revised file name is displayed in the side bar. The file renumbering process can be confirmed by comparing Fig. 6 with Fig. 4.

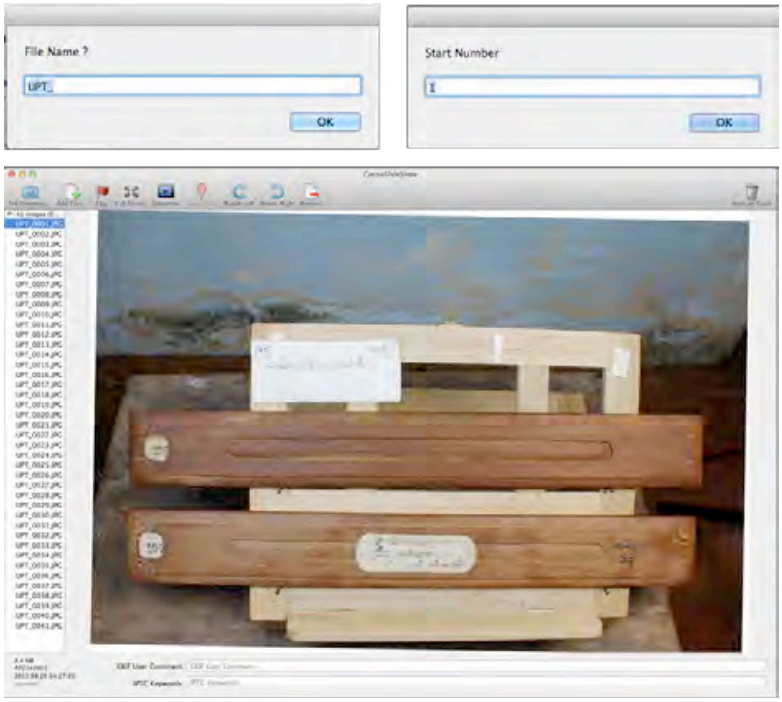


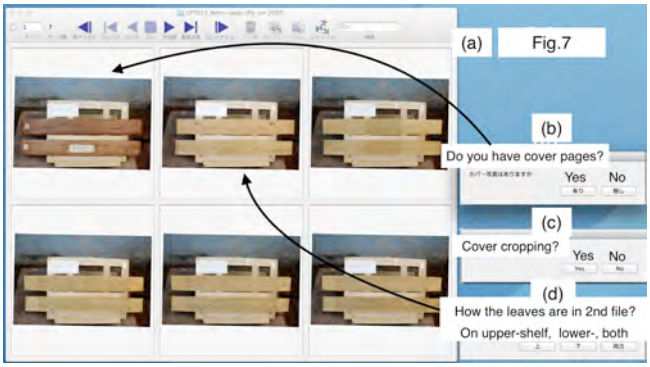
Fig. 6

2.4 The Cropping Process for the Front and Back Covers, and for the Leaves: ❶’s tool (Crop.app)

Since there are many leaves between the front cover and the back cover, the two covers are slightly larger than the leaves. So we should make two steps for cropping them: first, we crop the covers; then we crop the leaves.

2.4.1 The Cropping Process for the Front and Back Covers

When dragging and dropping the manuscript folder onto the processing tool "Crop.app" (see the figure in Appendix), we can get the cropping process for the front and back covers started as follows:



① By using PictCatalogX.app (see❹) a list is displayed (see Fig. 7(a))

② The dialog box (Fig. 7(b)) asks if you have a cover file. In this case click “Yes”.

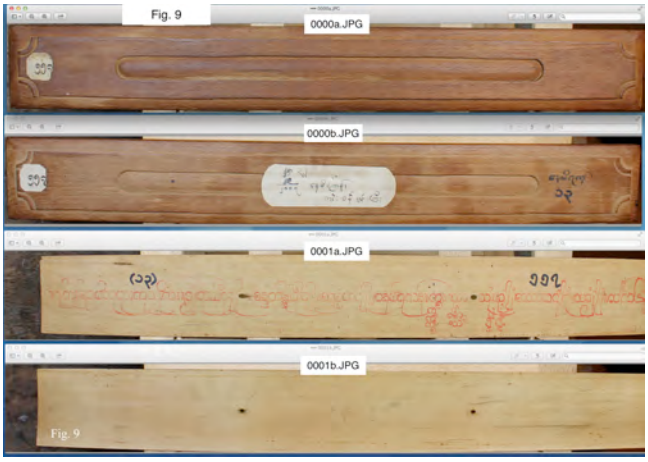
③ Next, the dialog box asks about cover cropping (Fig. 7 (c)). Click “Yes”.

④ Then the dialog box asks about the arrangement of the leaves in the second file (Fig. 7(d)), so click “Both” in this case.

⑤ Then PhotoShop will start automatically and the front and back covers will be shown as in Fig. 8(a). This page is used for the designation of the cropping areas.

⑥ First, a dialog box (Fig. 8 (b)) for setting the cropping area of the top (the front cover) is displayed. After drawing a range on the screen, the selection range is indicated by dotted lines, so click “OK”.

⑦ Next, crop the lower part (the back cover) in the same way.



⑧ Then the cropping process starts. The cropped result is saved in the (automatically generated) “TrimedFile” folder inside the photo folder. Those cropped images of the front cover and the back covers are named “0000a.JPG” and “0000b.JPG”, respectively: These are created by deriving “0000” from the photo file “UPT_0001, and by adding “a” and “b”, respectively. As you can see, each file name can be confirmed from the cropped result, Fig. 9 (the top two figures), and the background is almost completely removed from the photo.

⑨ Then, the same cropping ranges are applied to the second file and saved in “TrimedFile”: the results are shown as two figures at the bottom of Fig. 9. In order to save time, the image cropping process will stop at this stage.

⑩ Move this “TrimedFile” temporarily out of the manuscript folder, UPT_0013. (The purpose of this process is to avoid deleting the cropped cover files: the subsequent cropping process of the leaves overwrites the result in the same folder, “TrimedFile”).

2.4.2 The Leaf Cropping Process

Cropping a manuscript leaf uses almost the same process as the one for cropping the font and back covers.

You should make the difference between these two cropping processes clear by underlining one and adding a description in the leaf cropping process..

Drag and drop the manuscript folder onto the processing tool called "Crop.app"), then the following process will start:

① By using PictCatalogX.app (see❹) a list is displayed (see Fig. 7(a))

② The dialog box (Fig. 7(b)) asks if you have a cover file. In this case click “Yes”.

③ Next, when the same question (Fig. 7) is asked about cropping the cover, click “No” since this was already done (see 2.4.1).

④ Then the dialog box asks how the leaves are arranged (see the second file which is like Fig. 7(d) above). Click “Both” in this case.

⑤ Then PhotoShop will start automatically, and the fourth photo file page is shown as in Fig. 10(a). This page is used for designating the areas to crop.

⑥ First, a dialog box is displayed (Fig. 10 (b)) for setting the area to crop in the upper part (the verso side of the leaf). After you select the area on the screen, it is indicated by dotted lines. Click “OK”.

⑦ Next, crop the lower side (the recto side of the leaf) in the same way.

⑧ Then the cropping process starts. The cropped result is saved in the (automatically generated) “TrimedFile” folder inside the folder of photos. The numbering process is explained below. It depends on both the leaf number of the second photo file (the beginning page of the manuscript itself) and its location arrangement.

⑨ After using of the same process for all the leaf photos in the photo folder, the cropped results are saved in “TrimedFile”.

⑩ Put back the files “0000a.JPG” and “0000b.JPG” into this “TrimedFile” (these are the files that were created in the cropping process for the front and back covers).

Next we will discuss the numbering process as mentioned in ⑧. There are three different possible arrangements which depend on the number of the leaves in the second photo: (1) two leaves are located in the photo, one each in the upper and lower positions; (2) there is only one leaf in the upper position and the lower position is vacant; (3) there is only one leaf in the lower position and the upper holder is vacant. This situation does not depend on the presence or absence of covers.

As mentioned in Fig. 3(a) of §1, in the early stage of photo shoots, the front and back covers, or the rear side of the palm (verso) and the front side (recto) of the next palm are placed on the upper part and the lower part of the photo holder made with a wooden frame, respectively. We add the identification marks, “0a” and “0b” for the front cover file and the back cover file, respectively. We also add “b” and “a” for the upper photo (verso of leaf) and the bottom photo (recto of leaf) of the leaf photo holder, respectively.

If we follow these instructions, the numbering the leaves (regardless of whether there is a set of covers) will be as follows:

(1) if two leaves are located in the two holders: “1b, 2a, 2b, 3a, 3b …”

(2) if one leaf is in the upper holder only: “1a, 2b, 3a, 3b …”

(3) if one leaf is in the lower holder only: “2a, 2b, 3a, 3b …”

This is the basic numbering for the usual manuscript where the first leaf is 1 (i.e., “ka” in the Burmese numbering system). Numbers are written on the back (verso) of the manuscript leaves except for the last leaf of a text or a section where the text usually ends on the front (recto) side of the leaf, so the number is put on that side.

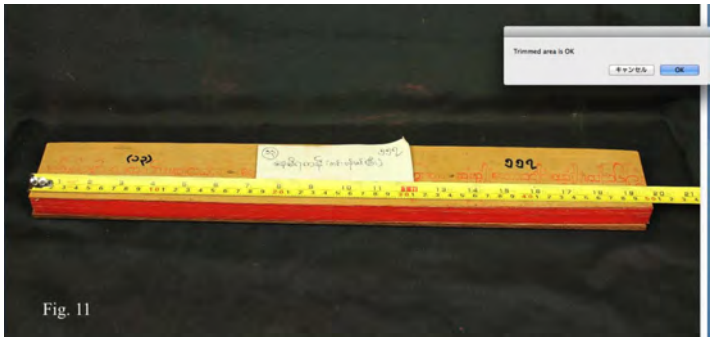


Fig. 11

There can be a number of exceptions in the numbering. The manuscript may be only part of the original with the first part missing. The scribes can make mistakes in the numbering; for example, two leaves may have the same number, in which case regular numbers “1”, “2”, etc., may be added to clarify the situation. Or a number might be skipped, in which case it is not always clear if it is a mistake in numbering or if a leaf is missing. More than one text may be in one manuscript, with later texts starting the numbering at 1 again. There may also be unnumbered leaves between sections, and these may include titles of the sections. These are not always correctly positioned. In order to find an array of irregular leaf numbers, use CocaSlideShow.app ❺ for checking the photo data and the leaf data. (It is very difficult to handle these exceptional cases using the cropping program.) After finding some errors, we must correct those file name/number/identification mark of the cropped results using several program tools: we can use the file renumbering processing tool ❼, or the file name change program ❽.

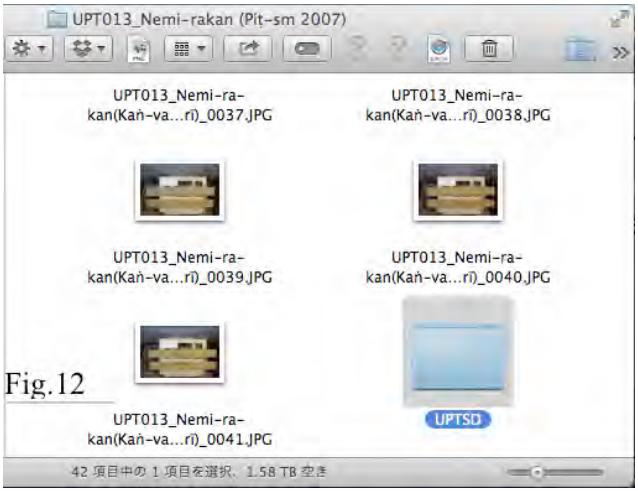
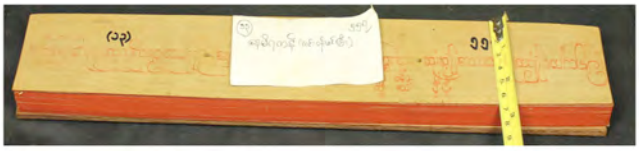


Fig.12


Fig. 13



2.5 The Cropping Process for the Archive Data of the Manuscript: The tool for ❶ (SizeDataPDF.app)

For this cropping process, inside the manuscript folder prepare the “UPTSD” folder for the archive data photos in JPG format (see Figs. 11 and 12). After dragging and dropping this UPTSD folder onto the processing tool “SizeDataPDF.app”, the following process will start:

① First of all, a blank page in PowerPoint will appear. The cropped data will be treated here later, but since this page is for an intermediate process, leave it as it is.

② Then “Photoshop” will start automatically. This is used for the cropping process, and the photo file (Fig. 11) for designating the cropping range of the archive data is displayed.

③ Designate the cropping range and click “OK” in the dialog box, then the cropping process begins. The cropped result is assigned to a designated location on the blank page displayed in ①.

④ After repeating the above process as needed, this tool will convert the created file in PowerPoint into PDF format, and it will name the file as “UPT_Scale.pdf”, then save it in the manuscript folder. The cropped result of manuscript number 13 is shown in Fig. 13.

2.6 The Process for Removing Black (or Green) Color in the Background of Cropped Photo Data: The Tool for ❷ (Remove_background.app, see the folder [2] in the figure of Appendix)

This removal process is performed as follows: Drag and drop the folder UPT013, which contains the previously created TrimedFiles folder, over to the tool Remove_background.app. This tool selects the color at the position of the left top background in the cropped photo file of the TrimedFiles, and then makes it transparent. Figure 14 shows a photo after the background has been removed: the background color (black) has been replaced by white (transparent). Note that the processed data has been saved in the the TrimedFiles (one *m*), while the original data is shifted to the new folder TrimmedFiles (two *m*'s) inside folder UPT013 (where this new folder is automatically created by this tool).

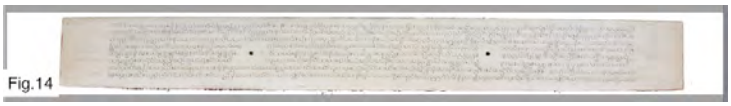


Fig.14

How to set up this tool: this tool is comprised of 3 freeware programs (see Fig. 15) and was created based on Apple Script as follows:

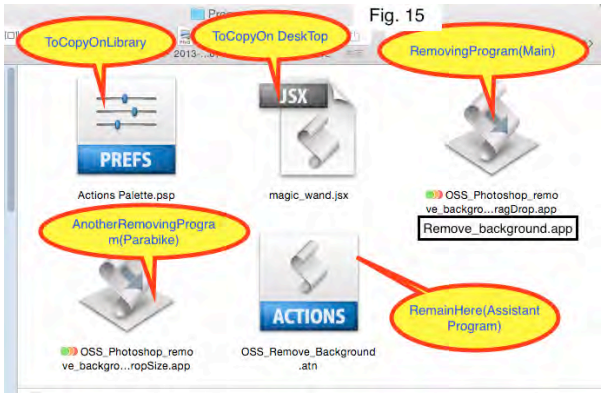


Fig. 15

① Remove_background.app makes Photoshop start up, and it makes the magic wands tool operate.

② "OSS_Remove_Background.atn" and "Actions Palette.psp" are the action files of Photoshop. The action file "Actions Palette.psp" needs to be taken into /library/Preferences/Adobe Photoshop *** Settings/, where "***" means your Photoshop identification (see Fig. 16). To access the hidden folder "Library", hold down the option key and click on "Go" in the menu bar of the desktop finder. "Library" will appear in this sub-menu. Open the Library folder, then drag and drop "Actions Palette.psp" into the "Adobe Photoshop *** Settings" folder. (This action tool becomes available by selecting the window/action in the top menu of Photoshop).

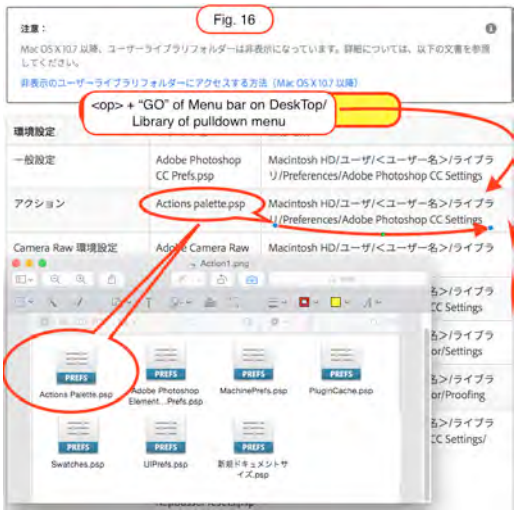


Fig. 16

③ The magic_wand.jsx is a tool for controlling the magic wand tool operation coded in Apple Script. This file should be on the desktop.

3. PDF Production Process

We will produce a digital book as a PDF of the manuscript material by using the cropped results obtained in "2.4 Cropping Process of Front /Back Cover, and that of Leaf: ❶'s tool (Crop.app)" and "2.5 Cropping Process of Archive Data of the Manuscript: Tool for ❶ (SizeDataPDF.app)". For the photo data with black (or green) background, we will use the data obtained in 2.6 Process for Removing Black (or Green) Color in Background of Cropped Photo Data: Tool for ❷ (Remove_background.app).

At this stage, all the required data for PDF production have been prepared inside the manuscript folder, UPT_0013. By dragging and dropping this manuscript folder onto the tool "PDFProd.app", the PDF production starts.

In this PDF production process, two types of PDF and their file names are automatically created. One is a full-size digital book for scholars; the file name is set as "UPT013F.PDF", etc. In this context, "UPT" is the abbreviation for the monastery library where manuscripts are kept, and "013F" means the manuscript number, while the "F" means a full set of images. The other file is a simplified version of the PDF to be uploaded on the web database. This file is named "UPT013S.PDF", where the "S" means a simplified version. These 2 types of PDFs are stored in the manuscript folder, UPT_0013.

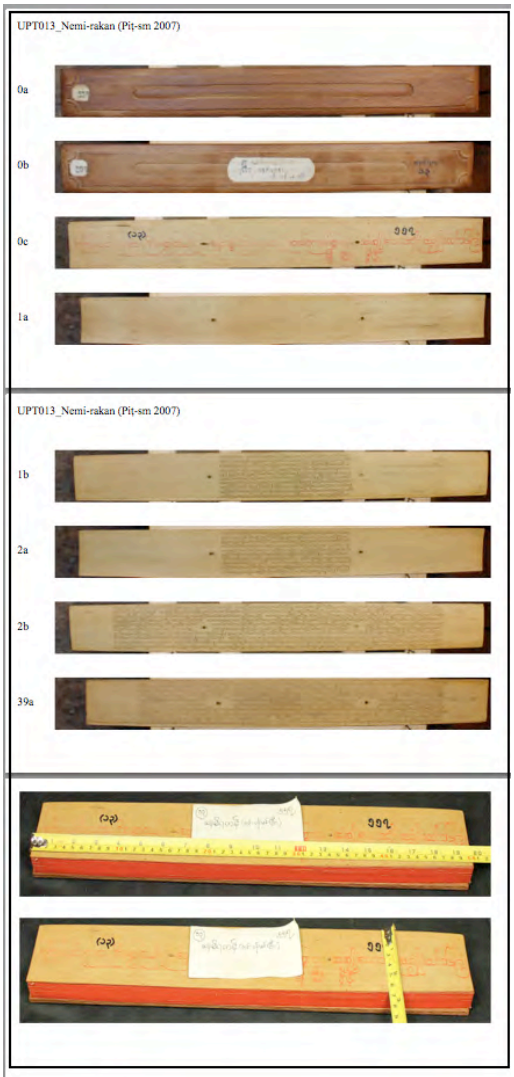


Fig. 2

The two files "UPT013F.PDF" and "UPT013S.PDF" are also saved in separate folders "DataBaseFull" and "DataBaseSimple" for easy access after PDF production. We need to create the folders with these names on the desktop in advance. After the PDFs are produced, they are stored in both the manuscript folder, UPT_0013 and the two folders prepared in advance on the desktop.

As mentioned in the introduction, the folder's name — "UPT013_Nemi-rakan (Piṭ-sm 2007), for example — which corresponds to a manuscript name, is inserted in the header of the PDF (see Fig. 2). Due to the restrictions on the characters that can be used in the Macintosh operating system (where "." and ":" are unavailable), the following process is performed for typing the folder names: To type a folder's name, the standard period "." and the specially prepared period "." and colon ":" for Burmese characters are replaced with the capital letters "Z", "W", and "Q", respectively. In the PDF production process, these capital capitals are changed back to the original characters, and typed in the header.

Appendix.

The whole program set with some sample data are shown in the figure, where [1] in the file name corresponds to the number ❶ appearing in §1, etc.