PDF 文档的翻译

由于 PDF 文件的广泛使用性以及它不同于 WORD 及其他文件格式的特性,使得 PDF 文件的翻译在本地化翻译探讨中受到比较大的关注。“怎样以尽可能少的时间有效地本地化 PDF 文件”一直是一个难题。

PDF 文件分为两种，一种是通过文本文件生成，其中也可能会包含图片，另一种是通过图像文件生成。前者可以选取文件中的字符进行编辑，后者只能浏览，进行一些图片性质的操作，不可以修改文字。

生成 PDF 文档的方式有很多种，可以购买专门的制作软件，其中 Abode Acrobat 最专业。

也可以使用 Foxit PDF Creator 等等。最简单的方法是在 Word2007 或者 OpenOffice 中直接生成 PDF。

翻译 PDF 文件涉及以下几方面的内容：

一、安全性限制问题

对 PDF 操作往往涉及到一个问题：PDF 的安全性限制。

一般而言，PDF 文件有其自身的安全证书，如果要对其进行编辑和操作，必须拥有相应的权限。

通常，本地化公司的标准流程是由客户提供源文件（见表 1），通过本地化工程处理工具和 DTP 工具处理，然后再生成 PDF。

表 1

客户可能提供的源文件类型①

DTP 文档类型

.fm.bk .book，.mif

frameMaker

.sgm .sgml .xml

framemaker + SGML Epic Editor Advent 3B2

.htm .html asp .aspx….

DreamWeaver FrontPage

.qxd

QuarkXPress

.pm6 .p65.pmd

PageMaker

.indd .indb

InDesign

.doc rtf

Ms word

.xtg

QuarkXPress 导出的带标签格式的文本

.ttx

Trados Tageditor

.isc

Trados Story Collector for InDesign

图形图像

.ai .eps

Illustrator

.cdr

CorelDraw

.fh8 .fh9

Freehand

.psd

Photoshop

.jpg .gif .png .bmp .svg .wmf .pict .Tiff

Photoshop or other

但在源文件不能获得的情况下，可以采用下列方法处理有安全性限制的 PDF：

在 Adobe Acrobat 中打开该文件，然后通过路径（ “文件” / “属性” / “安全性” ），使用密码去除安全性限制。

或者也可以使用 PDF 解密工具去除限制，例如 PDF Decrypter。

当删除了安全方面的限制之后，或者文件并没有任何限制，可以通过 （“文件”/“另存为”）把 PDF 文件另存为 DOC 或 RTF 格式，

或者从（ “文件” / “导出”），选择 WORD 文档或 RTF 格式。这两种方式得出的 DOC 或 RTF 文件区别不大。

二、PDF 格式转换

现在还不能直接翻译 PDF 文件。通常需要把 PDF 文件转换为其他格式，例如 DOC 或 RTF。重点在于尽可能的保留原文的格式，排版以及图片。下面比较几种工具各自的识别并转换 PDF 格式的优缺点：

1）Adobe Acrobat

使用完整版本的 Adobe Acrobat，把 PDF 格式转成 DOC 或 RTF 格式。非完整版本的 Adobe Acrobat 只能另存为 TXT 格式。由于 PDF 文件一般都包括文本样式及图片，所以如果存为 TXT 格式，将丢失大量样式信息，所以最好不要转换为 TXT 格式。

运用此方法得出的 RTF 或 DOC 的文件，页面的顺序有可能不一致，例如，最后一页被置于第一页。有些文本会被识别为图片;识别出的图片会出现错误，多出很多空白页，布局不紧密; 原文的图片与文本的布局会出现错误，需要大量的后期排版工作。总的来说，对于纯文本的 PDF 文件，这个方法简单方便，而且错误较少，但对于具有图片的 PDF 文件，这个方法得出的 RTF 或 DOC 的文件质量非常低，造成较大的内容调整和页面排版工作。

2）ABBYY FineReader

首先选择识别语言，打开 PDF 文件，选择保存文件的类型为 DOC 或 RTF。

此时还有四个选项供选择：精确复本、可编辑的复本、带格式文本和纯文本。

精确复本得出的文件中的文本是以文本框形式存在。

这种形式会给翻译阶段使用 Trados 带来一点麻烦，即当一个文本框中的内容翻译结束、进入下一个文本框时，会出现错误。

此时最好手动把光标放入到下一个文本框的文本处，再使用 Trados 的“打开/获取”。

同时，精确复本也不能完全保证图像的完好无损。

可编辑的复本去掉了精确复本中的文本框，避免了精确复本的问题，但是有些文档会识别出很多的换行符。可以在 WORD 中采用（“编辑”/“替换”/“高级”/“特殊字符”）里面选择“手动换行符”，查找内容的框里就出现了“^l”，然后在替换内容中不输入（如果文档是中文）或者输入一个空格（如果文档是英文），这样文档中的换行符就可以全部去掉了。对于 Trados 文档可以不去掉换行符，因为 Trados“打开/获取”时，是按句获取，换行符没有影响。

带格式文本中完全去掉了图像。但能够识别出可编辑的复本中不能识别的一些带背景的文字。

一般不选用纯文本格式。

对于图像，ABBYYFineReader

允许用户手动调整识别图像的大小，因为自动识别出的图像有些不完整。用户可以根据需要，删除图像。

3）OCR 软件

OCR 软件可以把纸质文件识别为电子文档。如果客户提供的 PDF 文件为纸质，就必须使用 OCR 软件。

国内的 OCR 软件有尚书、汉王和紫光等。国外的 OCR 软件有 Cuneiform、OmniPage、ScansoftPaperPort 等。通常，OCR 软件支持的文件格式为图像格式，所以如果电子文档为非图像格式，必须先获取图像。再经过识别后，把以尚书七号为例，支持的格式为 bmp、tif、jpg。对于非图像格式文件，必须先转为 bmp、tif、jpg 这三种格式之一。可以采用的方法有：
a）使用屏幕捕捉软件获取图像，

例如红蜻蜓抓图精灵。b）在 PDF 中，打开“文件”→“打印…”，选择

“Microsoft Office Document Image Printer”

打印机，打开“属性”→“高级”，输出格式选择 tif。

总体说来，OCR 软件用处理纸质 PDF 文件提供了方便。但 OCR 软件的识别效果不是很好，容易出现错误。电子文档的获取图像环节增加了工作量，而且图像获取的质量直接关系到识别的效果。

4）SolidConverter PDF

Solid Converter PDF 支持创建 PDF，能够把 PDF 转成其他的格式，包括：DOC、RTF、xml、XLS、TXT。SolidConverter PDF 识别出的效果比 ABBYY FineReader

还要好。不仅完整的保留了文本的所有格式，且页面的排版也与原来 PDF 的排版一致。很少有手动换行符。有些不好的地方是，原文页眉部分的图片被识别成了页眉，颜色受到了影响，得通过译后处理进行调整。但是，比起其他软件，Solid Converter PDF 的效果可以被称为“优秀”了，大大节省了翻译前处理和后期排版所需要的工作量和时间。

图像识别上也比 ABBYY FineReader 好很多，几乎没有错误！最重要的是，它能把图像识别为可编辑的模式，也就是当一个图片可以分割为几个部分时，它识别出的图片的这几个部分是组合在一起的，可以根据需要去掉不需要的部分。

但是因为这个功能，Solid Converter PDF 的另一个功能“提取 PDF 中的所有图片”在提取时，得出的图片是原来图片的组成部分，当然不能分割的图片得出的是完整的图片。

Solid Converter PDF 还有一个特别之处在于，图片中的文字可被识别为可编辑，即直接修改图片中的文字，

而不必使用 Photoshop 等软件。

但是有些不方便的地方是，当 PDF 文件没有被识别为 DOC 导出，还在 Solid Converter PDF 程序中时，是不能以其他语言更改图片文字的。例如，语言的的“文本更正”功能在原文为英文的 PDF 中只能写入英文,而不能直接写入对应的中文,也说是只能检察原文有没有错误，而不能实现本地化的目的。不知道这是 Solid Converter PDF 不持这个功能，还是 PDF 编码格式的问题。但是，这个问题的一个解决方案是，先在 Solid Converter PDF 中去掉 PDF 图片中的文字，在导出为 DOC 格式之后，再添加本地化文字。

5）其他的 PDF 文件转换工具

PDF 转 WORD 的小工具有很多，但大多效果不太好。例如，PDF2Word，它识别出的文本位于文

本框中，并且是每个段落一个文本框。

图片完全丢失。甚至有的文本框重叠在一起。

三、翻译

经过前面的处理,PDF 文件已经转成 DOC 或 RTF 格式,可以使用常规 CAT 翻译工具进行翻译，例如 Trados、雅信、Dejavu 等，由于翻译过程与翻译 DOC 或 RTF 文件的过程类似，这里只对 Trados 进行的几个插件简要叙述。然后对 Google 新推出的 GoogleTranslator Toolkit 进行详细介绍。

1）Trados

Trados 翻译转成 DOC 或 RTF 格式的 PDF 文件的一个问题是容易引起格式差错。因为转成 DOC 或 RTF 格式的 PDF 文件多多少少带有一些复杂格式。

对于 InDesignQuarkXPress 和 PageMaker 的各种格式的源文件，Trados 提供了专业的插件 Story Collector for InDesign Story Collector for QuarkXPress 和 Story Collector for PageMaker 来提取文件。对于 framemaker， Trados 提供了 S-Tagger for framemaker，它能够将 mif 格式文件中需要翻译的文本提取出来，并将格式信息标记化，保存为称为 STF 的 RTF 文件。在 Trados 中翻译后，再使用 S-Taggerfor frameMaker 转回 MIF 格式文件。这些为翻译 PDF 文件的源文件提供了便利。

2）Google Translator Toolkit

Google Translator Toolkit 支持上传与下载文件。用户把需要翻译的源文件上传到 Google Translator Toolkit，翻译之后可以把译文下载到自己的计算机上。Google Translator Toolkit 支持上传文件的类型包括：

HTML (.html)

Microsoft Word (.doc)

OpenDocument Text (.odt)

纯文本 (.txt)

富文本 (.rtf)

Google Translator Toolkit 还支持指定 URL，对网页进行翻译。

如果只有以上功能，那么这款新工具与“Google 翻译”，并无大的区别。

Google Translator Toolkit 的特别之处在于，它用到了 CAT 软件才用到的 Translation memories 和 Glossaries。目前，Translationmemories 支持上传的格式是 TMX;

Glossaries 支持上传的格式是 CSV。这样它结合了翻译记忆与 Google 原有的自动翻译，为译员节省了更多的时间。

笔者认为 GoogleTranslator Toolkit 有几个缺陷：一是用户不能更新已上传的翻译记忆与术语库。如要更新，需要把更新过的翻译记忆与术语库重新上传。

二是在翻译过程中，不能检索翻译记忆库以寻找相似译文，文档在上传过程中被自动翻译，也就是当文档上传完成，用户不能使用翻译记忆。

三是翻译记忆不支持模糊匹配，这将对翻译记忆造成浪费。

总的来说，这三方面都是 CAT 工具所具有的功能，它们对翻译记忆管理、方便译员、以及提高翻译记忆利用效率方面具有重要作用，是高效地形成优质译文的必备元素。

Google Translator Toolkit 的界面为所见即所得。原文与译文按句子一一对应。原文以黄色突出显示，译文以编辑框显示。点击原文或译文，会自动跳到对应的译文或原文位置。而且相对 Trados 来说，速度较快。这种编辑环境能明显帮助译员提高效率。Google Translator Toolkit 虽然支持所见即所得，但是对一些样式的支持不是很好。例如，项目符号，下载的译文中包括项目符号的文本部分不能正常显示。这可以通过复制界面中的译文来解决。

四、译后处理

文件翻译好后，还必须经过译后处理，包括文本格式调整、图片处理、表格处理、排版，可能还包括重新转成 PDF 文件。

1）图片处理

对于纯图片，即没有文本需要本地化的图片，只需从原 PDF 文件中复制或者截图即可。也可以采用格式转换时，软件识别出的图片。

对于即有文本需要本地化的图片，通常使用 Photoshop 等图片处理软件。方法是：

a. 在 Photoshop 中打开该图片（先从文件中取出进行保存）

b. 使用“仿制图章工具”用文字周围的背景涂抹掉原来的文字

c. 利用“文字工具”在此处添加对应的本地化文字，调整文字大小与位置，使之美观

d. 保存图片，格式依后面的排版需求而定，一般为 TIFF 格式。

除了图片处理软件，还可以使用格式转换时用到的软件，例如前面提到的 Solid Converter PDF。

类似的软件还有 Foxit PDF Editor。Foxit PDFEditor 可以编辑 PDF，但不能转为其他文件格式。而且相对来说，Foxit PDF Editor 的文本编辑方式不是很灵活，不太方便操作。

2）表格处理

可以重新制作表格，如果原表格不复杂，这样做比较节省时间。

也可以采取与图片处理同样的方式，先去掉原来的文字，再添加相应的本地化文字。

这适用于比较复杂的表格，例如有背景色的表格。

3）排版、格式调整及转换成 PDF 文件

常用的排版软件有 Adobe InDesign，QuarkXPress 和 Page Maker 等。这三者都支持导入 PDF 的背景。但是获得背景有一些复杂。可以通过 Foxit PDF Editor，去除 PDF 中的文本和图片，从而获得背景。

但是 Foxit PDF Editor 只能导出为 PDF 格式，而 QuarkXPress 导入 PDF 格式的背景时，不能保证完好。对于 Adobe InDesign 和 PageMaker，PDF 格式的背景是可以的。

使用 DOC 或者 RTF 格式导入译文时，可能会现乱码，这时，需要把 DOC 或者 RTF 改为纯文本格式之后再导入，当然也就必须得重新对所有内容进行格式调整。不过，最后的格式调整总是难以避免的，只是工作量的多少而已。可采用格式刷工具刷新格式。

Adobe InDesign 可以直接导出为 PDF 格式文件。

五、结语

PDF 文件的翻译主要涉及到安全性限制、格式转换、翻译和译后处理几个方面。

在前期格式转换时，应该尽量完好地保留原文的样式和布局，以便减少后期的排版工作量，节约项目时间。

在翻译过程中，翻译记忆和术语的利用很重要。但需要大量的积累为前提。Google Translator

Toolkit 在及时与简易性方面具有一定优势，值得借鉴。

通常，如果在翻译阶段不必考虑文本的格式，对于译员来说，是比较理想的。但是如果这将导致后期大量的排版工作，那么只能在翻译阶段尽可能保留原文格式。

PDF 文件的译后处理过程，相对于 WORD 要复杂很多，需要专业排版软件的参与。