

ABBYY FlexiCapture

动态数据捕获系统

机器可读表单 创建指南

目录

什么是表单? 3

机器可读表单..... 4

 表单填写方法.....4

 机器可读表单的元素.....4

 文本..... 4

 输入字段.....4

 复选标记字段..... 5

 复选标记组..... 5

 引用标记..... 5

 行分隔符..... 5

 表单标识符..... 5

 图片..... 5

 表格..... 6

 元素组..... 6

机器可读表单的类型..... 7

 滴定表单.....7

 选择适当的颜色..... 8

 带光栅背景的黑白表单.....8

 带光栅边框的黑白表单.....8

 黑白线性表单.....8

 选择适当的表单类型.....8

机器可读表单的一般要求..... 10

 表单背景.....10

 引用标记.....10

 复选标记字段.....10

 文本标记.....11

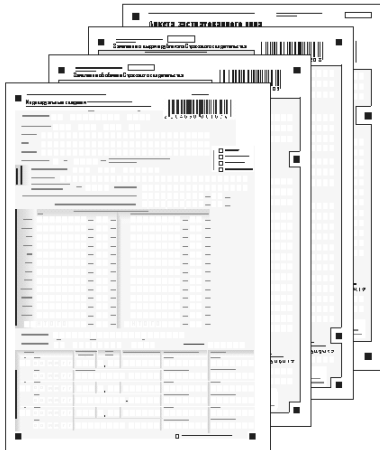
 元素定位.....11

 打印质量.....11

 表单填写.....12

 滴定表单的建议颜色.....13

什么是表单？



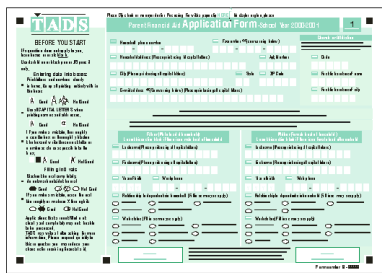
调查问卷、社会保障单、投票单、保单等都是不同类型的**表单**，这些表单用于收集不同类型的信息。

表单和其他类型的文档有何区别？

1. 表单总是有一组字段。
2. 每个字段可能只包含一种特定类型的信息，如“姓氏”字段只包含姓氏（如果填写正确），而“日期”字段只包含日期。

在要向大量的响应者收集信息时，就要用到表单。手工收集信息是个冗长而繁琐的过程，而且打字错误和其他错误等几乎不可避免，而使用机器可读表单可以使这一过程自动化。

自动处理表单的过程由以下阶段组成：



1. 安装表单处理应用程序（创建模板和指定要识别的字段）。
2. 获取表单图像（扫描）
3. 处理表单图像（识别图像和校验提取的数据）。
4. 将提取的数据导出至外部信息系统。

自动表单处理对于满足本章中所讨论的特定要求的表单最为有效。

机器可读表单

要读取表单上的信息，表单处理应用程序必须执行以下任务：

1. 确定表单元素的位置。
2. 从字段边框、文本标记、背景、说明性文本等中间区分出字段内容。

而机器可读表单可以让程序执行上述任务。

要成功执行第一项任务，表单必须**对应于某种表单模式或模板**，即所有同类表单上的所有表单元素的位置必须完全相同。

要成功执行第二项任务，**表单设计必须考虑自动化输入**，即要让程序容易区分要捕获的数据和字段边框、文本标记、背景或说明性文本等此类不可识别的表单元素。

表单填写方法

可以采用以下方法填写表单：

- 手工填写
- 使用点阵打印机
- 使用打字机
- 在印刷厂填写（以 300 dpi 以下的分辨率使用喷墨打印机或激光打印机填写表单）
- 采用上述方法的组合

机器可读表单的元素

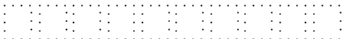
表单上可以显示的元素有：

文本

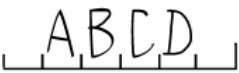
文本是机器可读表单的一种元素，这种元素包含表单标题、字段名称、说明等描述性文本。

输入字段

输入字段是机器可读表单的一种元素，填表人可在该字段输入文本。为方便文本输入，输入字段可以包含特殊文本标记。输入字段一般附带有描述或说明输入数据的性质的文本。

Name: 

下表列出了可能的文本标记类型。

文字位于线上	在一条线上输入文本。	
字母位于框中	在组合的框中输入字母。	
字母位于独立框中	在独立的框中输入字母。	
字母位于梳形中	在梳形中输入字母。	
文本位于框中	在框中输入文本。	
文本位于梳形框中	在梳形框中输入文本。	

复选标记字段

复选标记字段是机器可读表单的一种元素（通常为一个称作“复选框”的方形），由表单填写人在该框中进行标记。复选标记字段通常附带有复选标记使用说明的文本。

☐ Subscribe to the news list

下表中显示了可能的复选标记字段的类型。

方块形	圆圈	下划线
		

也可以采用多边形、椭圆形等其他形状。

复选标记组

复选标记组是机器可读表单的一种元素，它合并了多个标记字段字段，但只能从中选择一个标记字段。复选标记组一般附带有描述或说明选择性质的文本

May is the month of
which season: ☐ winter ☐ summer
☐ spring ☐ autumn

引用标记

引用标记是机器可读表单的一种元素，这种元素便于匹配表单图像和模板，以及确定自动处理过程中的表单方向。在程序开始识别表单内容前，必须匹配表单和表单模板。所有同类型的机器可读表单必须符合同一个模板，即所有表单上的所有表单元素的位置必须完全相同。程序将尝试根据表单上的引用标记的编号和位置来匹配表单和表单模板。如果程序能够匹配表单上的引用标记和模板上的引用标记，它就会知道包含有要识别数据的字段的精确位置。

下表列出了标准引用标记。

黑方块形	直角形	十字形	定时标记
			

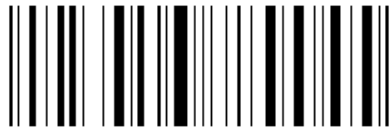
为了取得最佳识别结果，最好在表单上采用标准引用标记，但对于行分隔符等情况，也可以将说明性文本或条形码作为引用标记。但标准引用标记更可靠，它们始终是您的首选。

行分隔符

行分隔符是机器可读表单的一种元素，其形状为一条垂直或水平的黑线。行分隔符用于将表单分割成若干个逻辑块。

表单标识符

表单标识符是机器可读表单的一种元素，这种元素用于在多个模板有相同的引用标记排列时从中选择正确的模板。建议您使用条形码作为表单标识符。条形码是一组有不同宽度的垂直的黑白线条，用于信息编码。



也可以将表单上的一些其他元素用作表单标识符。

图片

图片是机器可读表单的一种元素，这种元素不要进行识别，而要作为图像文件导出。例如，应当将表单中的图章或签名部分放入图片中。

表格

表格包含多列同类元素。

表格列可以包含以下类型的元素：

文本；

输入字段；

日期；

数字；

复选标记。

一个表格可以包含一行列标题。

元素组

元素组是机器可读表单的一种元素，这种元素用于将多个有逻辑关联的元素组合到一个组中。

机器可读表单的类型

机器可读表单采用**自动输入设计**，即程序应当容易区分要捕获的数据和字段边框、文本标记、背景或说明性文本等此类不可识别的表单元素。

可以采用两种方法来达到这个目标：

1. 表单的设计必须保证字段边框、背景和文本标记在扫描以后消失，或者
2. 表单的设计必须保证字段边框、背景和文本标记在扫描以后更突出。

如果采用第一种方法，扫描后将只有引用标记、说明性文本和输入的数据保留在表单上。因此，程序无需从表单元素中分离出数据，这样将提高识别的便利性和可靠性。

对于这种情况，将使用这些类型的表单：颜色滴定表单、带光栅背景的黑白表单、带光栅边框的黑白表单以及黑白线性表单。

如果采用第二种方法，扫描后将只有引用标记、说明性文本、输入的数据以及文本标记和字段边框保留在表单上。程序将查找字段边框包围的信息和文本标记。如果缺少这些表单元素将对识别质量造成负面影响。

对于这种情况，将使用**黑白线性表单**。

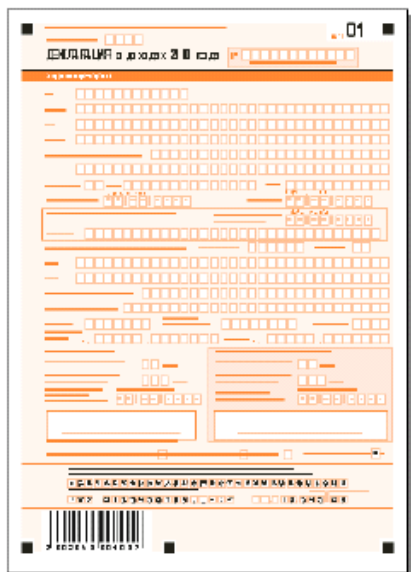
滴定表单

滴定表单有浅色背景，以白色矩形作为字符单元、复选标记字段等等。

以灰色滴定表单为例。它们的背景为灰色。

颜色滴定表单包括只有元素边框被有色的表单。

字符单元是一个小矩形，只能在其中输入一个字符（如一个字母或一位数）。背景色十分重要，因为背景色在扫描期间必须消失。背景色通常为淡红、淡橙或淡绿。有关建议的颜色的列表，请参阅“**滴定表单的建议颜色**”。



理论上，扫描过程中的所有元素（引用标记除外）都应消失，只留下表单图像上要识别的字段内容。

说明性文本可以随背景一起保留（如果以黑色打印）或过滤。要使说明性文本在扫描期间消失，必须采用和表单背景相同的颜色进行打印，但使用的保护度要更大。

在确定是否以黑色或背景色打印说明性文本时，应考虑以下因素：

如果保留说明性文本，图像文件会变大。

以黑色打印的说明性文本和表单其他元素之间的间距至少应保留 1.5–2 mm。说明性文本和相应复选标记字段的间距不小于复选标记字段大小的 2/3。如果以背景色打印说明性文本，文本和其他表单元素的间距就没有这么严格。

以黑色打印带说明性文本的表单可以提高易读性。特别是在表单上没有引用标记时，我们更是建议采用黑色打印说明性文本。

提示 对于以背景色打印说明性文本的表单，在创建这类表单的模板时，务必采用彩色模式或灰度模式扫描表单，以保证说明性文本和元素在表单上清晰可见。

滴定表单必须采用彩色的灯或过滤器进行扫描。灯的颜色要根据背景色进行选择。

有红色或绿色背景的表单必须采用下面的一种方式扫描：

在使用滤色软件的彩色扫描仪上扫描；

在使用红灯或绿灯的非彩色扫描仪上扫描（硬件滤色）；

在使用白灯和红色或滤色过滤器的非彩色扫描仪上扫描（这种情况的过滤质量较低，因为背景可能不会完全消失，或者字段内容被误删）；或者

在使用白灯但没有使用任何过滤器的非彩色扫描仪上扫描。

使用白灯扫描时，有些较浅的背景没有过滤而被删除；采用较弱的饱和度可以确保扫描期间背景消失。对于这种情况，必须对每台特定的扫描仪采用试错法选择背景的颜色和饱和度。

 **注：**很多彩色扫描仪也使用软件过滤来滤除蓝色。但是，不建议在表单上使用蓝色阴影，因为不只是可能用黑墨水填写表单，也有可能用蓝墨水填写表单。对于这种情况，用蓝墨水书写的字段内容将消失。

选择适当的颜色

红色-橙色比绿色更理想，因为这两种颜色与蓝色的对比度最大，因而在使用蓝墨水填写表单时，可以提高扫描和识别的质量。

在“**滴定表单的建议颜色**”中，可以找到表单处理的建议颜色，即使用任何扫描仪在扫描期间都会消失的颜色。也可以从扫描仪制造商/经销商处获取特定扫描仪（以 Pantone 格式或其他任何格式）的“滴定”颜色列表。

带光栅背景的黑白表单

这类表单的字段一般为光栅背景上的空白空间（一般为矩形）。背景由单个的点组成。建议点的大小不大于 0.1 mm，点距约 1 mm。点的大小和点距十分重要，因为会在去斑期间而不是在扫描期间清除光栅背景：程序将把这些点作为噪音处理并进行清除。因此，这些点不会附在扫描图像上。

带光栅边框的黑白表单

这类表单没有背景。所有字段边框（字符单元边框、复选标记字段和其他表单元素）组成了光栅线条，即小黑点序列。光栅点大小一般为 0.39–0.5 磅。建议的大小为 0.39 pt，建议光栅点之间的点距至少比点的大小大 5 倍。点的大小和点距十分重要，因为会在去斑期间而不是在扫描期间清除光栅背景：程序将把这些点作为噪音处理并进行清除。因此，这些点不会附在扫描图像上。

黑白线性表单

这些表单一般没有背景。所有字段边框（字符单元、复选标记字段和其他表单元素）组成了黑色实线。扫描或去斑不会造成字段边框消失。

 **注：**填写这类表单必须保持极其工整。如果填写表单不细心，字母可能会重叠字段边框，从而会降低识别质量。这是因为只要字段内容重叠字段边框，应用程序就很难区分边框和文字。理论上，输入的文字不应当接触到字段边框。

选择适当的表单类型

在选择表单类型时，除了要考虑每种表单的设计和处理要求，同时也要考虑一些其他标准，例如打印成本和打印方法、所需的副本数、得到的图像大小、扫描模式、要达到的识别准确度、表单处理速度、表单的整体外观、填写的工整程度等等。下表概括了每种表单类型的优缺点。

标准	滴定表单		黑白光栅表单		黑白线性表单	
	缺点	优点	缺点	优点	缺点	优点
设计复杂性		方便使用任何图形编辑器进行设计。	文本编辑器为这类表单的设计提供的工具比较有限。光栅点的大小十分关键。			方便使用任何图形编辑器进行设计
打印	较难在内部打印大量高质量的表单。			便于内部打印。		便于内部打印。
打印成本	如果采用专业打印服务，与黑白表单相比打印成本要高一些。			如果采用专业打印服务，与滴定表单相比打印成本要低一些。		如果采用专业打印服务，与滴定表单相比打印成本要低一些。
图像大小		图像文件较小。	图像文件较大。		图像文件较大。	
扫描模式	只能使用一种特殊扫描模式（滤色）。		亮度和对比度的设置范围有限。		亮度和对比度的设置范围有限。	

标准	滴定表单		黑白光栅表单		黑白线性表单	
	缺点	优点	缺点	优点	缺点	优点
处理速度		平均处理速度较高。	由于需要为图像去斑，因此平均处理时间较长。		如果字符粘合在一起或重叠字段边框，很可能会出错。	
识别质量		识别质量很高。		识别质量较高。	识别质量主要取决于填写表单的工整程度。	如果表单填写正确，识别质量较高。
说明性文本的位置。		说明性文本可能插入任何位置，包括插入到字段中间（只要使用与表单背景色相同的颜色打印就可以）。	说明性文本和其他表单元素必须在表单上小心排列。		说明性文本和其他表单元素必须在表单上小心排列。	
表单外观和易读性	背景和说明性文本的对比度设置较低，可能有损表单易读性。	美感	设计吸引眼球的能力较差。易读性较低。		设计吸引眼球的能力较差。	易读性较好。
填写的工整程度		对文本重叠字段边框不太敏感。	填写小字符时必须极其小心，字符要足够大以免被当做噪音清除。	对文本重叠字段边框不太敏感。	对文本重叠字段边框十分敏感。	

机器可读表单的一般要求

现在让我们来看以下达到机器可读表单的一般要求。

表单背景

1. 如果可能，请使用滴定表单或带光栅字段边框的表单。
2. 创建滴定表单时，使用一种在扫描期间会消失的颜色。有关建议的颜色的列表，请参阅“**滴定表单的推荐颜色**”。
3. 说明性文本和字段内容之间务必要保持距离。对于滴定表单，建议采用与表单背景相同的颜色进行打印（说明性文本用作引用标记或与黑方形组合使用的情况除外）。通过采用与背景色相同的颜色打印信息，可以随意地在表单上插入信息，其中包括可以插入表单字段里面。
4. 请不要对要用手工填写的表单使用蓝色。蓝色背景可能会与填写表单使用的蓝墨水融合，因而对识别质量造成不利影响。

引用标记

1. 建议对机器可读表单使用标准引用标记（即黑方形、直角形、十字形）。建议使用 5 个引用标记：表单每个角各一个引用标记（如果在每对之间画假想线，它们应当构成一个矩形），假想矩形的一条边上一个方形。
2. 如果因故不能使用标准引用标记，建议使用以下元素组合作为引用标记：
 - a) 至少两个垂直线分隔符和两个水平线分隔符
 - b) 至少四个**文本**元素和一个**表单标识符**元素
 - c) 也可以使用其他组合
3. 引用标记的边与最近的表单元素之间的距离不得小于 3 mm。
4. 引用标记的边与页边之间的距离不得小于 8 mm。
5. 应使用黑色（建议的颜色是黑色）打印引用标记，这样在扫描期间就不会消失。

除了上述一般要求外，每个引用标记的类型也有其本身的特定要求。

黑方形

相同类型表单上的黑方形的大小应相同，如 4 × 4 mm 和 8 × 8 mm。建议的大小是 5 × 5 mm。不允许使用矩形。

十字形和直角形

1. 绕此类引用标记绘制的假想方块的大小必须在 4x4 至 8x8 mm 的范围内。建议的大小是 5 × 5 mm。
2. 一张表单上的某一类引用标记的大小必须相同。
3. 引用标记的所有边必须完全相等。
4. 在同一张表单上的引用标记使用的线条厚度必须相同，其范围为 0.3 至 1 mm。建议的厚度是 0.5 mm。

文本（若用做引用标记）

字号不小于 7 磅（标题 – 不小于 14 磅）

行分隔符（若用做引用标记）

线条厚度不小于 1 磅。建议的厚度是 1–1.5 磅。

条形码（若用做引用标记）

1. 条形码宽度（最左边的条形和最右边的条形之间的距离）不小于 47–50 mm。
2. 条形码高度不小于 12–15 mm（不包括数字）。
3. 条形码方向（条形的方向）应与页面方向相同（建议以条形码的方向扫描页面）。
4. 建议的条形码格式是 EAN 13。
5. 建议条形码和其他任何表单元素之间的距离不小于 10 mm。

复选标记字段

1. 建议的字段大小为 3.5x3.5 mm、4x4 mm、4.5x4.5 mm 或 5x5 mm（如果使用圆形复选标记字段，则这些字段必须适合建议大小的方形）。

2. 建议的复选标记字段边框的线宽为 0.4 mm（不论字段大小如何）。

文本标记

光栅点大小：

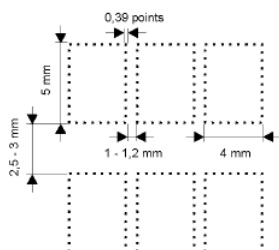
1. 如果字段边框为光栅点，光栅线的厚度（即光栅大小）必须为 0.39 磅。
2. 光栅点之间的最佳距离为该大小的五倍。

重要事项！ 最有可能粘合在一起的点就是字符单元角中的点。

提示

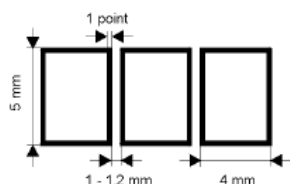
如果软件不允许指定光栅点之间的距离，必须选择一种其上的点保持足够距离的线型，以防止它们在扫描图像上粘合在一起和图像去斑后仍留在图像上，因为 OCR 系统不会将大点解译为噪音，也不会在去斑过程中被清除。

字符单元大小：



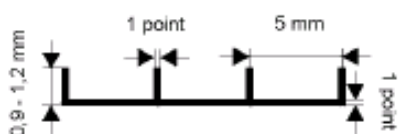
1. 字符单元的建议大小为 4x5 mm；如果使用更小的尺寸，则要求填写表单的准确度更高，而这一点很难达到。同时也不建议使用大的字符单元，因为这会鼓励用异常大的字母。理论上，字符单元大小应反映出较平均的字母大小。
2. 字段行距不小于 2.5 - 3 mm，同一行中两个相邻字符单元的距离不小于 1-1.2 mm。

线条厚度：



如果边框为黑线（“文字位于框中”、“字母位于独立框中”或“字母位于框中”），线条厚度应为 1 磅。

梳形：



1. 建议的梳形中凹口与凹口之间的距离为 5 mm。
2. 两个梳形（一个梳形放在另一个梳形下面）的水平线之间的距离不小于 7.5 - 8 mm。
3. 凹口高度约 0.9 - 1.2 mm。
4. 凹口和水平线的厚度均应为 1 磅。
5. 如果使用带梳形的框，框与框之间的最小距离应为 2.5 - 3 mm。

元素定位

1. 务必保证表单有足够边距，否则，太靠近表单的边的表单元素可能会在扫描过程中消失，从而导致数据丢失。
2. 相邻表单元素之间的距离不小于 2 mm。
3. 复选标记字段和说明文本之间的距离不小于复选标记字段大小的 2/3。
4. 确保说明性文本不重叠输入字段边框。
5. 如果说明性文本以黑墨水打印，确保说明性文本和其他元素之间至少有 1.5-2 mm 的距离。如果采用和背景色同色的墨水打印说明性文本，则该距离无关紧要。

打印质量

打印空白表单时，请记住以下要求：

- 1. 必须以专业方式打印表单或使用打印机打印表单。建议采用专业方式打印彩色表单。如果可能的话，在打印之前请在扫描仪上测试以下表单颜色。
- 2. 必须使用相同的源文档打印所有表单副本，因为每张表单上的字段位置必须完全相同。
- 3. 边距（即页边和最近的表单元素之间的距离）至少要达到 8 mm（建议使用 12 mm 边距）。
- 4. 允许的表单元素的线性偏差不大于 0.15%（对于 § 4 页面而言，不大于 0.5 mm）。
- 5. 如果使用打印机打印表单，请不要使用低于 600 dpi 的分辨率进行打印。
- 6. 建议使用同一台打印机或同型号的不同打印机打印表单。

 **提示。**不推荐使用复印设备复制机器可读表单。使用这类设备可能会稍微改变表单上的元素尺寸或元素排列，例如，表格或输入字段边框可能变粗，或者色深发生改变，这些都有损识别质量。

表单填写

笔和墨水

最好以大写字母、使用黑色圆珠笔、中性笔或毛细管绘图笔工整地填写表单（深蓝和紫色的也可以）。不推荐使用软尖/毡尖笔，因为使用这类笔填写的字符可能较粗，会导致发生识别问题。使用铅笔或浅色墨水笔填写的表单的识别质量最低。

字母大小和样式

为了取得最佳识别结果，应采用正楷大写字母的形式填写表单。必须将字母填满字符单元，但不得重叠单元边框。如果可以很清楚地看到字符单元（即使用以下文本标记类型：字母位于框中、字母位于独立框中），填表时将字母粘合在一起的可能性也低一些。

以下文本标记类型会给识别带来大问题：文本位于梳形框中、字母位于梳形中。它们只预定义了字符单元的宽度，但缺少边的边框，容易使字母混在一起。对于没有框的梳形，容易造成填入的字母太小或太大的问题。

如果采用文本位于框中或文本位于一条线上的方式，字符单元的宽度米有明显限制。字母可能会粘合在一起，这会给识别质量带来不利影响。对于文本位于一条线上的情形，没有限制字符高度。因此不推荐对要求达到高识别质量的表单采用这种标记类型。

另外一个重要要求就是字母不能重叠字段边框。这对于滴定表单和黑白光栅表单无关紧要，因为字段边框会在扫描和去斑过程中消失。但如果黑白线性表单上的字母重叠字段边框，就有可能导致识别质量大大降低。

样本

为促进正确填写表单，建议在表单上附上一份类似下面的注释。

Important! The form should be completed in CAPITAL LETTERS using a BLACK or DARK BLUE ballpoint/fountain pen. Characters and marks used should be similar in the style to the following:

ABCDEFGHIJKLMN O PQRSTU VWXYZ I 2 3 4 5 6 7 8 9 0 ☒ ☒

滴定表单的建议颜色

下表包含滴定表单样本及其相应的 Pantone 数值。这些背景色在扫描时会消失：

使用任何扫描仪；

Process Yellow U	100U	101U	388U	3935U	393U
102U	106U	107U	3945U	394U	3955U
108U	109U	113U	395U	3965U	396U
114U	115U	116U	461U	475U	503U
1205U	120U	1215U	586U	587U	600U
121U	1225U	122U	601U	602U	603U
123U	127U	1345U	607U	608U	609U
134U	1355U	135U	614U	691U	705U
1485U	148U	149U	706U	712U	713U
1555U	162U	169U	714U	719U	Cool Gray 1U
372U	386U	387U			

使用带有软件过滤红色的彩色扫描仪，或使用带有红灯/红色过滤的非彩色扫描仪。

100U	101U	102U	1905U	190U	1915U	488U	489U	495U
106U	107U	108U	191U	1925U	192U	496U	5035U	503U
109U	113U	114U	196U	197U	198U	509U	510U	5175U
115U	116U	1205U	199U	203U	204U	517U	5245U	530U
120U	1215U	121U	205U	206U	210U	5315U	531U	585U
1225U	122U	1235U	211U	212U	213U	586U	587U	600U
123U	127U	128U	217U	218U	219U	601U	602U	603U
129U	130U	1345U	223U	224U	225U	604U	607U	608U
134U	1355U	135U	226U	230U	231U	609U	614U	656U
1365U	136U	1375U	232U	2365U	236U	663U	664U	670U
137U	141U	142U	2375U	237U	2385U	671U	672U	673U
143U	144U	1485U	238U	2395U	239U	677U	678U	679U
148U	1495U	149U	243U	244U	245U	684U	685U	691U
1505U	150U	151U	246U	250U	251U	692U	698U	699U
1555U	155U	1565U	252U	2562U	256U	700U	705U	706U
156U	1575U	157U	2572U	263U	2706U	707U	708U	709U
1585U	158U	1625U	379U	386U	3935U	712U	713U	714U
162U	1635U	163U	393U	3945U	394U	715U	719U	720U
1645U	164U	1655U	395U	459U	460U	726U	Cool Gray 1U	Orange 021U
165U	169U	170U	461U	4685U	468U	Process Yellow U	Process Magenta U	Purple U
1787U	1788U	178U	473U	474U	4755U	Red 032U	Rhodamine Red U	Rubine Red U
182U	183U	184U	475U	481U	482U	Warm Gray 1U	Warm Red U	Yellow U
185U	1895U	189U	485U	486U	487U			