李康：

你好！下面是数据的简要说明。如果说的不太清楚，你给我指出来，我再详细解释。

**开始之前**

上这个网站必须要登陆一下，因为部分内容（链接第四层）只有登陆后才能看到。

网 址：http://amma.artron.net/artronindex\_artistall.php

用户名：billybilly10

密 码：shiyang

需要的内容：网页上全部的艺术家2000-2015年的交易信息。（由于这家网站内容变了些，加入了不少新的艺术家和个别新的指标，我想着全面重新记录一下）。

**一、总的来说**

如图0所示，我希望得到这样的最终结果：在一个excel2007文件里面，把**所有艺术家**的**每一次交易的情况**全部列出来；交易情况包括作品名、估计、成交价、交易日期等等；任意一位艺术家的任意一次作品交易占用excel的一行（例如，如果齐白石的作品有133条交易记录，那么齐白石就占excel里面的133行）。

注：这里只是简单说下，具体要求以我发的excel表为准。

网页链接共分成4层，第一层→→第二层→→第三层→→第四层：第一层是艺术家列表；第二层是年份季度；第三层是这个年份季度的所有交易情况；第四层是每次交易的详细情况。

第一层需要的数据：艺术家姓名 (黄色底色)；

第二层需要的数据：无；

第三层需要的数据：序号、名称、估价、成交价、拍卖日期、拍卖机构（蓝色底色）；

第四层需要的数据：名称2、尺寸、作品分类、创作年代、估价、**成交价**、专场、拍卖时间、拍卖公司、拍卖会、**说明1、说明2、说明3**等（绿色底色）。（标**红色**的几个变量，在记录过程中可能会出现些问题，具体可见“第4步”中的“注1”、“注2”）

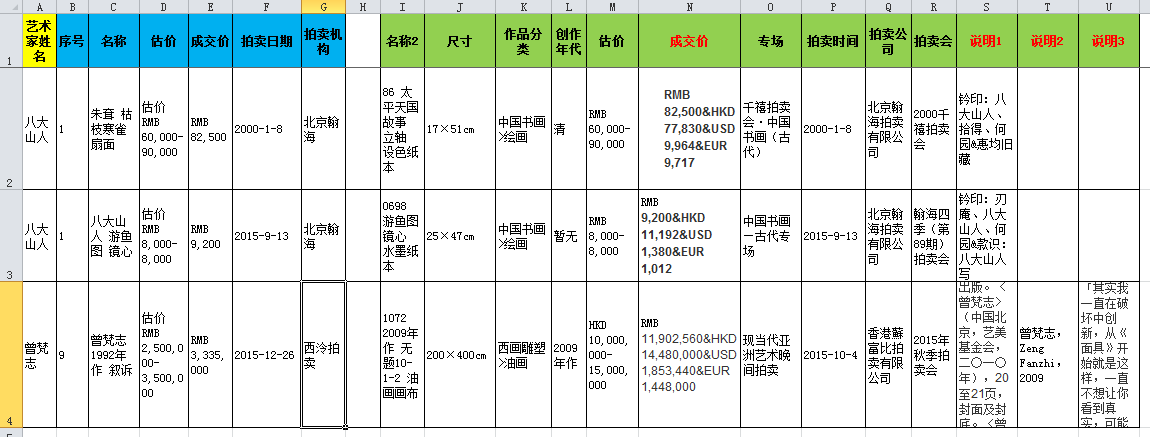


图0 简单示意图

**二、详细步骤**

**第1步**，进入“页面1”（如图1）

网址：http://amma.artron.net/artronindex\_artistall.php



图1

图1是艺术家列表。这里每一位艺术家的名字都是一个链接，我们需要每一个名字后面的数据。

红圈表示我们这里举例点击的艺术家；绿框表示这里要记录的内容，即在进入“八大山人”链接之前，可以先记录下来他的名字“八大山人”。

需要记录内容：艺术家姓名。

点击红圈名字，进入步骤二。

**第2步**，进入“页面2”（图2）

“页面2”中最右边的每一个“查看”都是链接，我们需要每一个“查看”后面的内容。

需要记录内容：无

“页面2”直接显示的数据，不需要记录。我们点击红圈内容为例，进入下一步。



图2

**第3步**，进入“页面3”（图3）

“页面3”是记录数据的主要页面之一。

每一个“名称”都是一个链接。以红圈的链接为例，在点击进入之前，我们需要记录这里的绿框的内容。

需要记录内容：**序号、名称、估价、成交价、拍卖日期、拍卖机构**。

点击红圈，进去最后一层的链接。

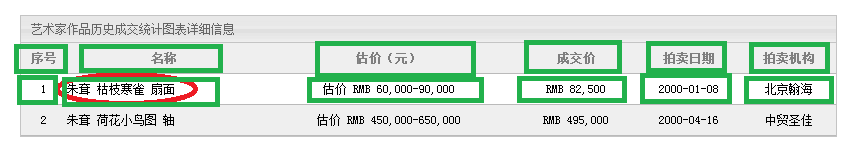


图3

**第4步**，进入“页面4”（图4-1、图4-2、图5）

“页面4”是另一个主要的数据记录页面，也是所有链接的最低一层。

这里要记录的数据如绿框所示。

需要记录的内容：名称2、尺寸、作品分类、创作年代、估价、**成交价**、专场、拍卖时间、拍卖公司、拍卖会、**说明1、说明2、说明3**。

注1：粉色框里面的“成交价”，包含四种货币单位价格。我在手动粘贴的过程中，会出现一次占用四个EXCEL纵向格子的问题。

如果可以的话，麻烦把四种价格都放在“**成交价**”一个格子里，不同价格之间用“&”隔开就行，例如这里为“RMB 　82,500&HKD 　77,830&USD 　9,964&EUR 　9,717”（多余的空格问题，我可以自己处理，如RMB与82500之间的两个空格）。

注2：粉色框里面的“说明”，在不同的艺术家网页链接中可能不一样，可能由多个绿色框格构成（目前看到最多为3个），如图5。每一个框格里面，又可能包含多行文字。我在手动粘贴的过程中，也会出现一次占用多个EXCEL纵向格子的问题。

如果可以的话，麻烦：（1）把这些内容分别放入EXCEL“**说明1”、“说明2”、“说明3**”；（2）每一个网页链接中的框格，对应一个EXCEL“说明”（如果网页中有超过三个的框格，最好也放在“说明3”中）；（3）每一个框格内的多行文字，用“&”隔开就行，例如，图4-2对应一个说明，“说明1”为“钤印：八大山人、拾得、何园&惠均旧藏”。图5对应三个说明，“说明1”、“说明2”、“说明3”。



图4-1 八大山人作品

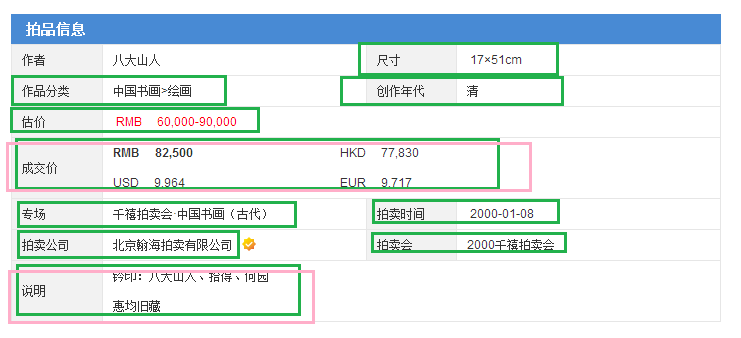


图4-2 八大山人作品



图5 曾梵志作品（三个“说明”框格）

详细步骤就是上面的4步。我把上面举的这个例子的结果附上了（详见“数据样板”文件），其中共有两张表格：“第一版”（图6-1）与“‘说明’简化版”（图6-2）。

如果“说明”部分实在难以处理，就按“‘说明’简化版”，把说有的那些“说明信息”，都放在一个EXCEL格子里。



图6-1 第一版

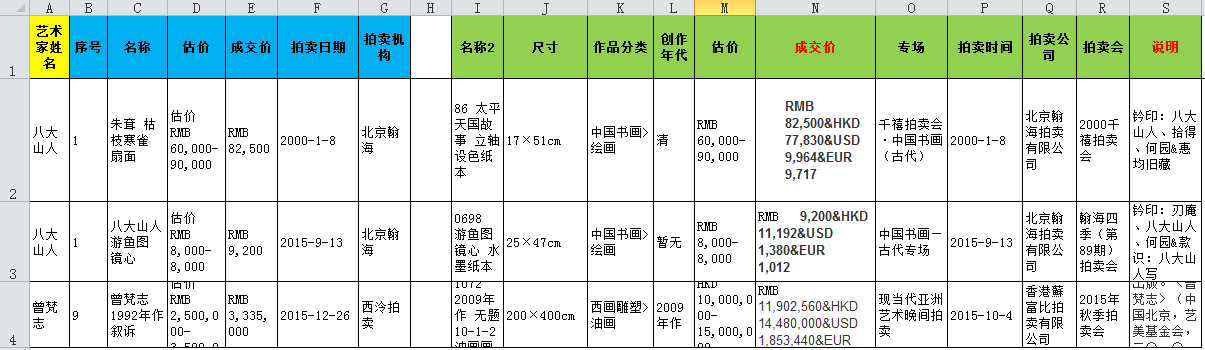


图6-2 “说明”简化版

**三、总结**

总的来说，网页链接共分成4层，第一层→→第二层→→第三层→→第四层：第一层是艺术家列表；第二层是年份季度；第三层是这个年份季度的所有交易情况；第四层是每次交易的详细情况。

第一层需要的数据：艺术家姓名 (黄色底色)；

第二层需要的数据：无；

第三层需要的数据：序号、名称、估价、成交价、拍卖日期、拍卖机构（蓝色底色）；

第四层需要的数据：名称2、尺寸、作品分类、创作年代、估价、成交价、专场、拍卖时间、拍卖公司、说明1、说明2、说明3等（绿色底色）。

每一个艺术家的每次交易占用excel的一行，预计总共有30万以上吧。

因为excel2007文件最大行数好像没有限制，所以就用这个吧。Excel2003文件最大行数为6万多行。估计电脑打开，数据要下个几天几夜。最好中间能够有所储存，以防万一出些意外。

另外，上次的数据，有极少数出现了一点点小问题：可能由于网页框格“说明”部分太长或有多个换行等原因，个别交易记录占用了两行，不过没有影响其余数据的准确性。所以，最好每一条数据都明确的在EXCEL中另起一行，这样即使有些小问题，不至于影响后续数据的记录。

大概就是上面这些内容，可能写的有点啰嗦，没明白就跟我说哈：）

再次表示感谢！