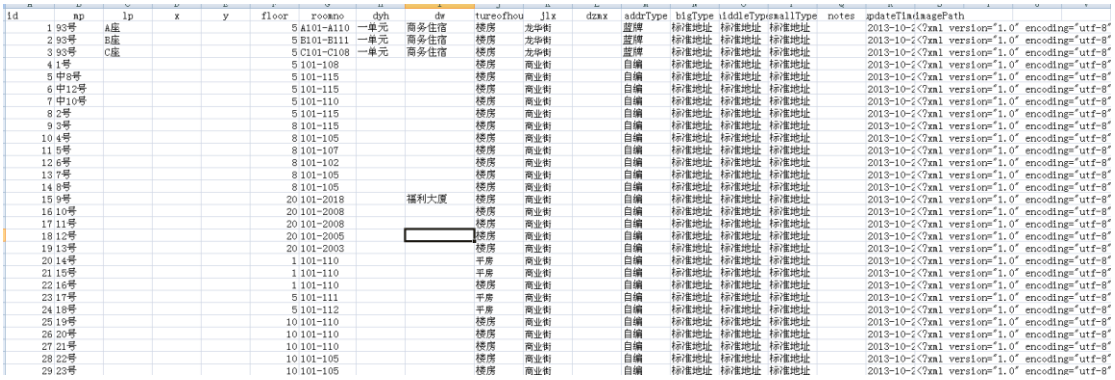
地址库数据表的拆分与提取

2010 年接触 FME 到现在,也有几年时间了.相信不少读者开始接触 FME 纯是为了转换数据格式,从转换格式开始,到后来的做属性传递挂接,以及数据分析统计等,一步步的熟悉,从学习中发现 FME 的强大,慢慢的发现工作中已离不开 FME.真心感谢安图公司对软件的汉化,让更多的从业人员容易理解到软件使用流程.

下面这个案例,是对采集的地址做拆分提取.

源数据是 EXCEL 表格，数据表的格式如截图：



id	mp	lp	x	y	floor	roomno	dyh	dw	tureofhou	jlx	dzmx	addrType	bigType	iddleTyp	smallType	notes	pdateTim	imagePath
1	93号	A座			5	A101-A110	一单元	商务住宿	楼房	龙华街		蓝牌	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
2	93号	B座			5	B101-B111	一单元	商务住宿	楼房	龙华街		蓝牌	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
3	93号	C座			5	C101-C108	一单元	商务住宿	楼房	龙华街		蓝牌	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
4	1号				5	101-108			楼房	商业街		自编	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
5	中8号				5	101-115			楼房	商业街		自编	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
6	中12号				5	101-115			楼房	商业街		自编	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
7	中10号				5	101-110			楼房	商业街		自编	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
8	2号				5	101-115			楼房	商业街		自编	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
9	3号				8	101-115			楼房	商业街		自编	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
10	4号				8	101-105			楼房	商业街		自编	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
11	5号				8	101-107			楼房	商业街		自编	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
12	6号				8	101-102			楼房	商业街		自编	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
13	7号				8	101-105			楼房	商业街		自编	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
14	8号				8	101-105			楼房	商业街		自编	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
15	9号				20	101-2018		福利大厦	楼房	商业街		自编	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
16	10号				20	101-2008			楼房	商业街		自编	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
17	11号				20	101-2008			楼房	商业街		自编	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
18	12号				20	101-2005			楼房	商业街		自编	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
19	13号				20	101-2003			楼房	商业街		自编	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
20	14号				1	101-110			平房	商业街		自编	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
21	15号				1	101-110			平房	商业街		自编	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
22	16号				1	101-110			楼房	商业街		自编	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
23	17号				5	101-111			平房	商业街		自编	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
24	18号				5	101-112			平房	商业街		自编	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
25	19号				10	101-110			楼房	商业街		自编	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
26	20号				10	101-110			楼房	商业街		自编	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
27	21号				10	101-110			楼房	商业街		自编	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
28	22号				10	101-105			楼房	商业街		自编	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"
29	23号				10	101-105			楼房	商业街		自编	标准地址	标准地址	标准地址		2013-10-2	<?xml version="1.0" encoding="utf-8"

成果要求是提取照片号，拆分楼层房间号，最后再组合成标准的地址串。

思路：首先提取表格中的照片编号，再对楼层进行拆分复制。然后再对拆分楼层后的数据进行房间号码拆分处理。最后生成标准地址串

1.照片编号的提取

表格的内容如下：

<?xml version="1.0" encoding="utf-8"?><multimedia><pictures><picture name="2014-05-12ab1019d151.jpg"><desc><![CDATA[门址照片]]></desc></picture></pictures><audios/><videos/><offices/></multimedia>

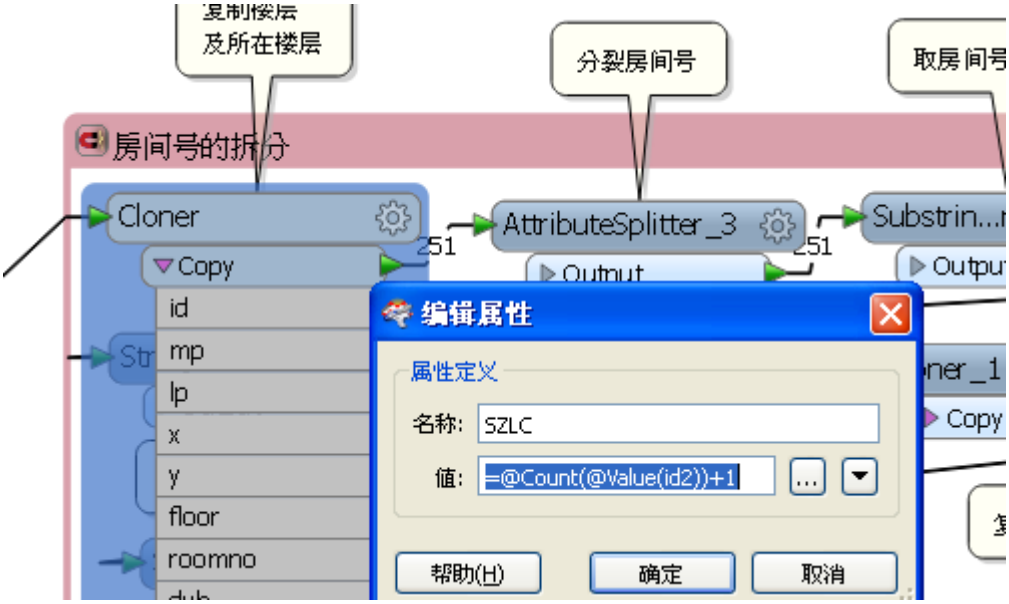
这个比较简单，使用 StringSearcher，正则填写 [^"]+.jpg 提取照片编号.

如下图所示：

对于数据的提取匹配,FME 的正则表达式可以说是非常有用的一个方法,常常能得到一个事半功倍的效果.

2.楼层与房间号的拆分处理

每一条数据的楼层（floor）是正确的，但房间号(roomno)的楼层却不一定标准。比如楼层最高 18 层，每层有 8 个房间。房间号(roomno)可能会写成 101-108，也有写成 101-1808.
因此，先对楼层进行复制，使用 Cloner 复制后再进行编号。得出每条数据所在楼层(SZLC)：



房间号的拆分做了一个取巧的方法。先用 AttributeSplitter 转换器，以“-”对房间号(roomno)进行分裂，再使用 SubstringExtractor 提取分裂后_list{0}及_list{1}的后二位。SubstringExtractor 设置如下：

计算每层楼的房间数：(_R)-(_L)+1，使用 Cloner 进行复制后再编号。得出每个房间号码（FJS_OK）。

最后就是：SZLC+FJS_OK=完整的数字房间号。

处理到这一步还没完，因为有不少房间是带有英文的，比如 A101 之类。
同样可以使用 StringSearcher，正则填写[a-zA-Z]，过滤出英文
_en+SZLC+FJS_OK=完整英文房间号
如果带有中文比如甲 101，丙 105 之类，则可以使用正则 [\u4e00-\u9fa5]，过滤出中文

以上运行完成后,从 29 条数据生成 2157 条数据.
生成的标准地址串为: 广东省广州市天河区商业街 9 号福利大厦 612 房
附上拆分的模板和测试数据。FME 版本为 2014 SP4 14433