**ppt格式文件解析**

总的解析规划： ppt/pptx两种文件格式的了解并解析，由C++语言编写，以动态库的方式提供给外部程序调用。PS: 此动态库只做数据的解析，不提供渲染，算法转换等操作，若要额外实现功能则在上层再封装调用。

**2019-08-26**

相较于pptx，ppt格式（复合文件二进制格式CFBF）更难理解解析，先着重理解ppt格式

1. 了解DocToText项目，此项目初步实现解析ppt/pptx格式，并简单的能获取文本。（缺点：依赖第三库较多，只支持linux环境交叉编译生成windows可执行程序）
2. 参照b2translator项目，初步搭建ppt文件格式的解析。
3. 文件二进制方式读取，字节流读取等基础类实现。
4. 初步实现获取ppt文档的用户信息。
5. 阅读[MS-PPT].pdf，[MS-CFB].pdf文档，了解文件结构

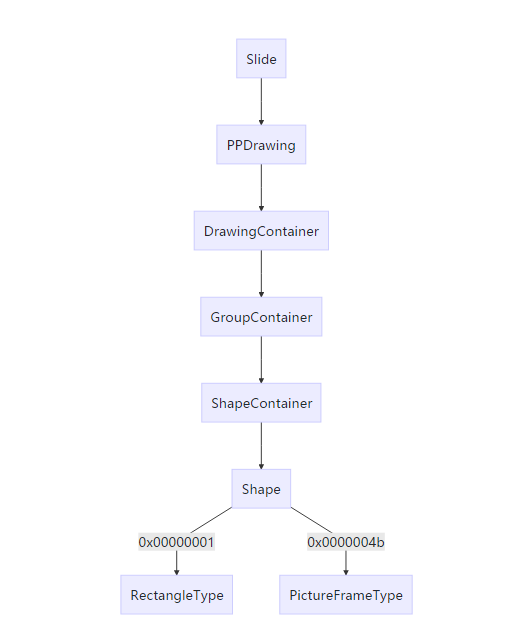
**2019-09-02**

1. ppt格式的顶层文件结构格式的了解，创建对象模型工厂类，根据具体的TypeCode将解析到的ppt文件某块字节流反序列化成对象模型。
2. 顶层的存储和文件流的解析，已实现获取字节流以及流的大小
   1. Current User Stream
   2. PowerPoint Document Stream
   3. Pictures Stream
   4. Summary Information Stream
   5. Document Summary Information Stream
   6. Encrypted Summary Information Stream
3. Current User Stream 字节流的解析，已实现获取文档最后用户修改的信息，版本号等
4. PowerPoint Document Stream 此流包含了一些列ppt文档顶层结构，主要的相关信息都是从此流中解析获取。包含的record容器：
   1. DocumentContainer
   2. MainMasterContainer
   3. HandoutContainer
   4. SlideContainer
   5. NotesContainer
5. Pictures Stream 字节流的解析，目前只实现获取到顶层的字节流，还未剥离须参考[MS-ODRAW].pdf文档说明
6. Summary Information Stream 和 Document Summary Information Stream 与Su 包含一些OLE属性集，还未解析剥离
7. Encrypted Summary Information Stream 文档解密信息字节流，还未解析剥离须参考[MS-OFFCRYPTO].pdf

现阶段正在查阅文档并了解PowerPoint Document Stream 流中的DocumentContainer相关的数据结构，编码实现相关数据的解析。

**2019-09-09**

1. 从ppt文档二进制流中解析并获取ppt文档的幻灯片母版（MainMaster）、讲义母版、备注母版相关信息。
2. 从ppt文档二进制流中解析并获取文档中的slideLayout\*.xml并与pptx格式对应。
3. ppt解析shape数据模型,目前支持解析到Rectangle与PictureFrame，具体对象属性的读取须查看[MS-PPT].pdf文档



**2019-09-16**

1. 以Slide对象为父节点，解析并获取各子节点对象，元素ShapeContainer表示ppt文档的原子对象，可用于标识文本、图片、形状等。
2. ShapeContainer对象的子节点 ShapeOptions包含了ppt文档原子对象的属性集合。
3. 文本的解析：ShapeContainer对象子节点ClientTextbox包含shape中文本相关信息,如：TextAtom对象可解析到文本的相关信息（PS：段落的解析，目前来看ppt是通过\n分割不符合面向对象，具体的信息还需继续了解）。
4. 图片对象的解析：可根据ShapeOptions对象集合是否包含Pib、pibName、pibFlags（填充标识）来判断原子对象是否是图片对象。pid属性值表示图片文件流的唯一标识
5. 通过pid属性值的唯一标识，从Pictures Stream文件流获取相对应的图片文件流。

对象模型层次图：

