多格式文档解析与管理

邱磊　陈家辉　李少飞　介来拉石　肖克　卫青

2021年9月27日

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

目录

CONTENTS

01

文档格式标准与发展

展示人：邱磊

1、常见文件格式

“文件格式” 是指电脑为了存储信息而使用的对信息的特殊编码方式，是用于识别内部储存的资料。

PNG

GIF

文档解析——没有标准的困境

办公软件互不兼容

微软office、SUN的StarSuite、红旗中文贰千的RedOffice、金山WPS、永中Office等

封闭格式 & 开放格式

基于二进制的封闭格式的文档，可能面临数据损失、隐私泄露等问题。
基于XML的纯文本格式的文档，具有开放性和可继承性，不受软件约束

文档格式标准——三足鼎立

- **ODF** (Open Document Format, 开放文档格式)
  - 2002年，SUN、IBM等36家公司创建ODF联盟。
  - **2006年5月**ODF被确立为国际标准。
- **UOF** (Unified Office document Format, 标文通)
  - 2002年，红旗中文贰千、金山、永中等企业及中科院研究所，推出UOF。
  - **2007年5月**被确立为国家标准。
- **OOXML** (Office Open XML)
  - 微软在Office文档格式基础上开发的技术标准。

文档格式的国际标准之争

ODF

国际标准，开放性、可继承性，支持不同程序不同平台交换文件

VS

UOF

中文文档格式规范，兼容ODF，2007年9月正式颁布实施。中文文档开始有了**自主格式标准**。

OOXML

包含大量微软私有标准和技术，文档有超长的6000页，只有微软单个产品能实现全部功能（**垄断**）

Chrome HTML Document	1,252 KB
Chrome HTML Document	120 KB
Microsoft PowerPoint 演示文稿	12,559 KB
Microsoft Word 文档	20 KB
PNG 文件	3,167 KB

文档管理

指文档、电子表格、图形和影象扫描文档的查阅、存储、分类和检索。

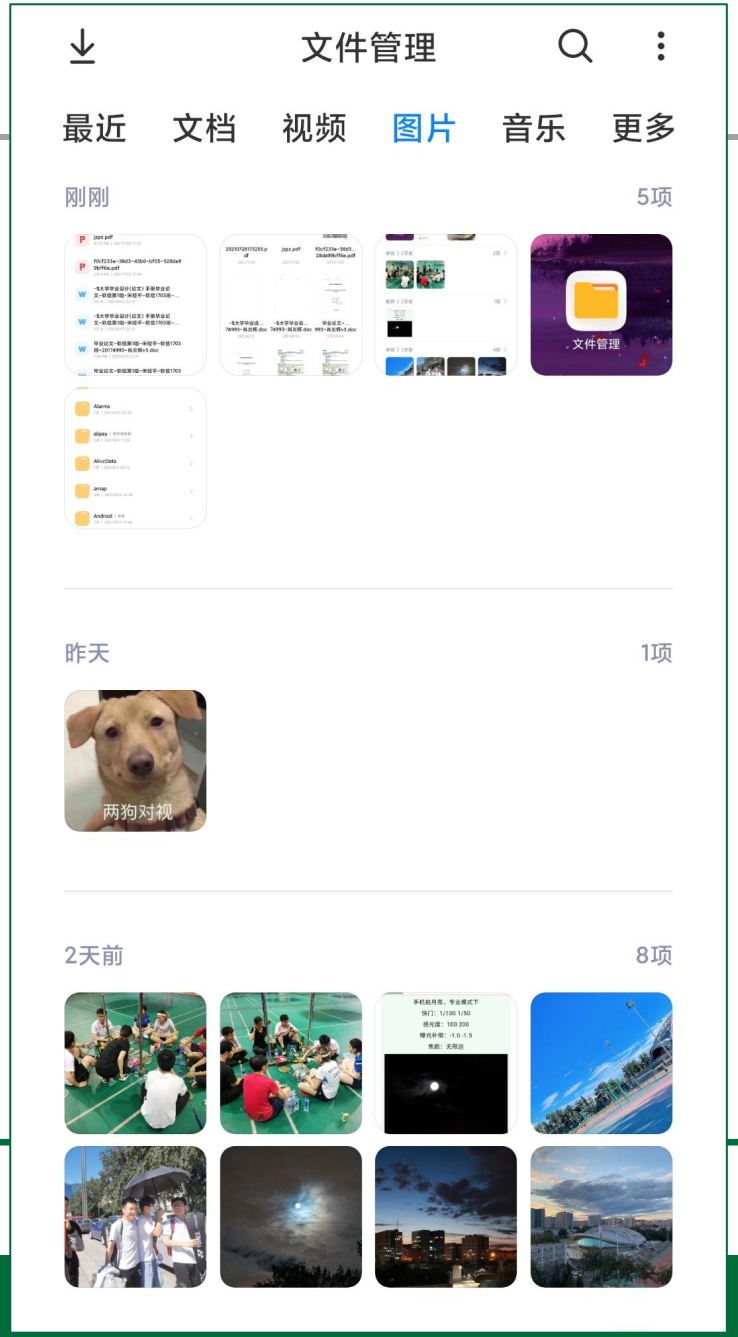
每个文本具有一个类似于索引卡的记录，记录了诸如作者、文档描述、建立日期和使用的应用程序类型之类的信息。

传统的文档管理：管理**存档的资料**，被搜索重用频率不高。

现代的企业场景：管理**过程性文档**，文档生成和修改频率高，对成稿时间要求高，团队多地域分布，需要员工随时能看到**最新版本信息**，随时能从海量文档中**搜索到**需要的信息，比如合同、简历、法律文书、PR稿等等。

1.4 手机文档管理


文件管理


文件管理
最近
文档
视频
图片
音乐
更多
刚刚
5项
文件管理
昨天
1项
两狗对视
2天前
8项


文件管理
手机
最近
文档
视频
图片
音乐
20210728173255.pdf
43.12 KB | 2021/7/28 17:32
jzpz.pdf
43.12 KB | 2021/7/28 17:32
f0cf233e-38d3-43b0-bf55-528da99bff6e.pdf
214.5 KB | 2021/7/25 17:06
~$大学毕业设计(论文) 手册毕业论
162 B | 2021/6/15 22:21
~$大学毕业设计(论文) 手册毕业论
162 B | 2021/6/15 22:21
1.34 MB | 2021/6/15 21:15
1.34 MB | 2021/6/15 21:15
基于混沌脉冲位置调制的数字通信系统的设计与实现
304.08 KB | 2021/6/15 21:04
基于混沌脉冲位置调制的数字通信系统的设计与实现
304.08 KB | 2021/6/15 21:04

1.4 网盘文档管理

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

目录

CONTENTS

2 PDF格式文档解析

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

PDF格式文档解析

1 PDF文档概述

2 PDF文档对象

3 PDF文档结构

4 PDF文档解析

展示人：李少飞

PDF是Portable Document Format的简称，意为“可携带文档格式”，在1993年被美国排版与图像处理软件公司Adobe首次提出。PDF文件以PostScript语言图象模型为基础，无论在哪种打印机上都可保证精确的颜色和准确的打印效果，即PDF会忠实地再现原稿的每一个字符、颜色以及图象。

PDF 文档兼有电子媒体和纸质介质的优点，**高效**、**可检索**，而且信息传递又**可靠**、**完整**，与传统印刷品相比有无法比拟的优势，所以大多数人将PDF文档作为各类软件系统中存储的解决方案。

PDF文档对象

构成 PDF 文档的基本元素是对象（Object）。PDF 对象可以分为两种，分别是**直接对象**（Direct Object）和**间接对象**（Indirect Object）。由于 PDF 是由PostScript发展而来，所以它的数据信息类型和编程语言比较相似，也可以分为多种类型。PDF对象的基本类型如下表所示：

PDF 的对象类型	描述	示例
数组（Array）	有序的列表	[1　2　3　4　5]
布尔值（Boolean）	逻辑真/假值	false
数值（Number）	两种数值对象，整数和实数	1.2
名称对象（Name）	一个元子符号，被字符序列唯一定义	/ Type
字符串（String）	字符串	（abc）或<Aabb>
空对象（The null object）	由关键字“null”表示	null
字典对象（Dictionary）	由许多对象组成的表，用一对双尖括号包围	<< / Type /Catalog /Page 3 0 R /Outline 2 0 R >>
流对象（Stream）	通常表示压缩后的数据流	13 0 obj <</Type /Xobject>> stream 030004040404040 endstream

PDF文档对象

数组、**布尔**、**数值**、**名称对象**、**字符串**、**空对象**这六种结构都比较简单，**字典对象**和**流对象**则结合其他简单类型对象用于存储更加复杂的结构。

- 数组列表中的类型可以是这八类的任意类型，包含数组类型本身；
- 名称对象的最大长度是127，并且名称对象是不可分割且唯一的；
- 字符串有两种表示方法，可以用“（）”或“< >”来表示，用“< >”来表示时，两位表示一个字符，不足用0补齐；
- 字典对象通常用来说明一个对象的多个属性特征，字典对象的第一个元素是键，用名称对象表示，第二个元素是键对应的值，为任意PDF中的对象。
- PDF的页面和字体都是带有特殊属性的字典对象，可以通过查阅PDF参考手册来了解特殊属性的含义和类型说明。

PDF文件结构

1、文件头。通常出现在文件的第一行，标识PDF文件版本号。

2、文件体。PDF文档的主体部分，主要包括了组成PDF文档的各种对象。

3、交叉引用表。包含文件中间接对象的信息，存储了间接访问对象地址的索引表。

4、文件尾。给出交叉引用表的地址和某些特殊对象（如Catalog）的地址。

PDF文件结构

文件头:

```
%PDF-1.7
```

文件体:

```
1 0 obj
<< /Type
/Catalog
    /Outlines 2 0
R
    /Pages 3 0 R
>>
endobj
```

交叉引用表:

```
xref
0 7
0000000000 65535 f
0000000009 00000 n
0000000074 00000 n
0000000120 00000 n
0000000179 00000 n
0000000300 00000 n
0000000384 00000 n
```

文件尾:

```
trailer
<< /Size 7
    /Root 1 0 R
>>
startxref
408
%%EOF
```

解析PDF文档的过程也是再现PDF文档结构的过程。对PDF文档内容的访问是从目录对象开始，进一步访问页面对应的内容流对象。每个对象都有数字标号，作为引用值，对象之间通过引用联系起来，这样的话这些对象就可以被其他对象引用。通过按照文件结构，按照其规则，层层访问，保存其结构，保存其每一个页面，每一个对象来实现的。

其中一些PDF文件为了文档安全可能会存在加密的情况，可以采用第三方软件进行解密，然后再对PDF文档进行解析。

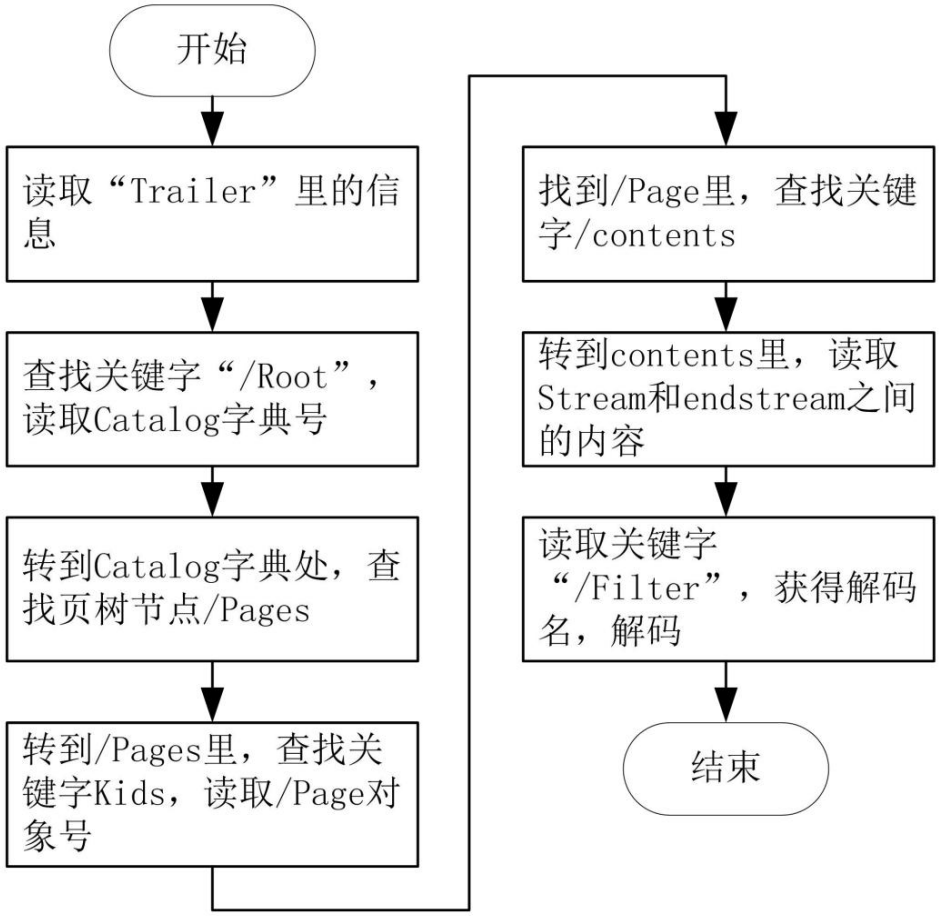
①从trailer中找到Root关键字，Root是指向Catalog字典，Catalog是一个PDF文件的总入口，它包含Page tree，Outline hierarchy等。

②从Catalog中找到Pages关键字，Pages是PDF所有页面的总入口，即Page Tree Root。

③从Pages中找到Kids和Count关键字，Kids中包含Page子节点，Count列出该文档的总页数。到这里我们已经知道PDF文件有多少页了。

④从Page字典中获取MediaBox、Contents、Resources等信息，MediaBox包含页面宽高信息，Contents包含页面内容，Resources包含页面所需要的资源信息。

⑤从Contents指向的内容流中获取页面内容。


开始
读取“Trailer”里的信息
查找关键字“/Root”，读取Catalog字典号
转到Catalog字典处，查找页树节点/Pages
转到/Pages里，查找关键字Kids，读取/Page对象号
找到/Page里，查找关键字/contents
转到contents里，读取Stream和endstream之间的内容
读取关键字“/Filter”，获得解码名，解码
结束

PDF文档解析过程

PDF文档解析


```
1 0 obj                          %obj 是对象开始的标志
<< /Type /Catalog
/Page 30 R
/Outlines 20R >>
Endobj                           %对象结束关键字

3 0 obj
<< /Type /Pages
/Count 1                         %页码数量为 1
/Kids [4 0 R] >>                 %Kids 对象说明它的子页对象为 8
endobj

4 0 obj
<< / Type / Page
/ Parent 30 R
/ Resources<< / Font << /F1 70 R>> / Proc Set 60 R >>
/ Media Box [ 0 0 612 792]       %页面显示大小，以像素为单位
/ Contenets 5 0 R >>             %页面内容对象的对象号为 5
Endobj
```

```
5 0 obj                          %文档内容对象的对象号，标志对象的开始
<< /Filter /FlateDecode          %流对象的压缩方式为
/FlateDecode /Length 148         %流对象的长度
>>Stream                         %流对象
BT
/F13 24 Tf                       %表示字体和字体大小
60 60 Td                         %表示这一行文字在页面上的开始位置
(Hello World)Tj                  %表示该页面的文本内容为“Hello World”
ET
Endstream                        %流对象结束标志
endobj

Trailer
<</Size 20                       %文件体对象数目
/Root 10 R                       %根节点引用
>>
Startsref 3000
%%EOF
```

Trailer → Root → Catalog → Pages → Page → Contents

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

目录

CONTENTS

3 Word格式文档解析

展示人：陈家辉

- OpenXML简介
- Word文件结构
- document.xml主体结构
- Word文档解析流程
- 应用现状

OpenXML简介

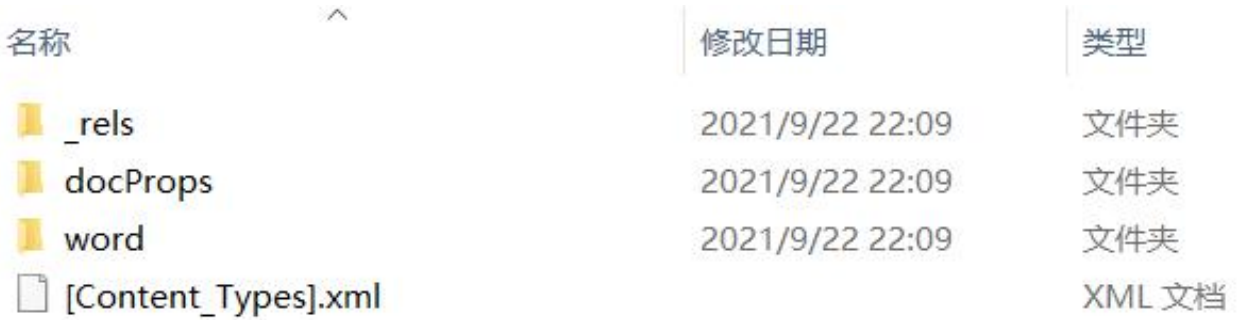
OpenXML(OOXML)是微软在Office 2007中提出的一种新的文档格式，在Office 2007及之后的版本中的Word、Excel、PowerPoint默认均采用OpenXML格式。

.doc ⇨ .docx

.xls ⇨ .xlsx

.ppt ⇨ .pptx

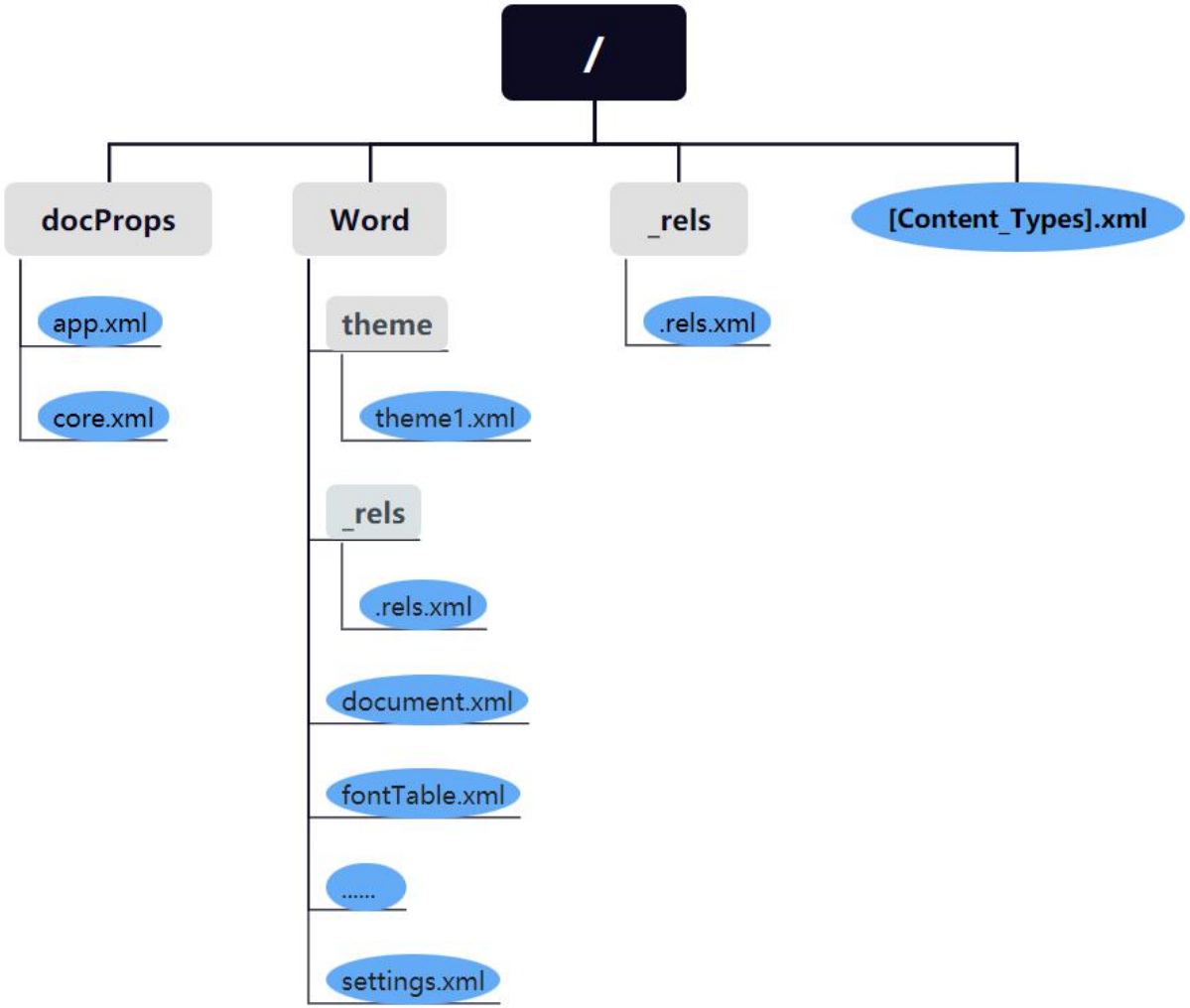
以Word为例,在OpenXML格式下,docx文件比doc文件所占用空间更小，docx文件本质上可以看作一个XML文件的集合,采用通常的压缩软件可以对其进行解压



解压docx文件后文件夹的内容

Word文件结构

docx文件目录结构



部分XML文件描述

组件名	描述
app.xml	应用程序特定属性
core.xml	文档格式的通用文件属性
.rels.xml	存储父文件夹所有xml信息和索引id
document.xml	文档中所有文字的内容和属性
fontTable.xml	文档所使用字体信息
setting.xml	文档总体设置信息
style.xml	文档整体样式信息
……	……

document.xml主体结构

例:Hello World!

```
- <w:body>
  - <w:p w:rsidRDefault="00E17673" w:rsidR="00F7256C" w14
    - <w:r>
        <w:t>Hello World!</w:t>
      </w:r>
      <w:bookmarkStart w:name="_GoBack" w:id="0"/>
      <w:bookmarkEnd w:id="0"/>
    </w:p>
  - <w:sectPr w:rsidR="00F7256C">
      <w:pgSz w:w="11906" w:h="16838"/>
      <w:pgMar w:gutter="0" w:footer="992" w:header="851"
      <w:cols w:space="425"/>
      <w:docGrid w:linePitch="312" w:type="lines"/>
    </w:sectPr>
</w:body>
```

XML节点标签描述

节点标签	描述
<w:p>	表示一个段落
<w:pPr>	表示此段落属性
<w:r>	表示一个样式串
<w:rPr>	表示此样式串属性
<w:t>	真正的文本信息
<w:pgSz>	页面大小
<w:pgMar>	页边距
……	……

Word文档解析流程

直接读取

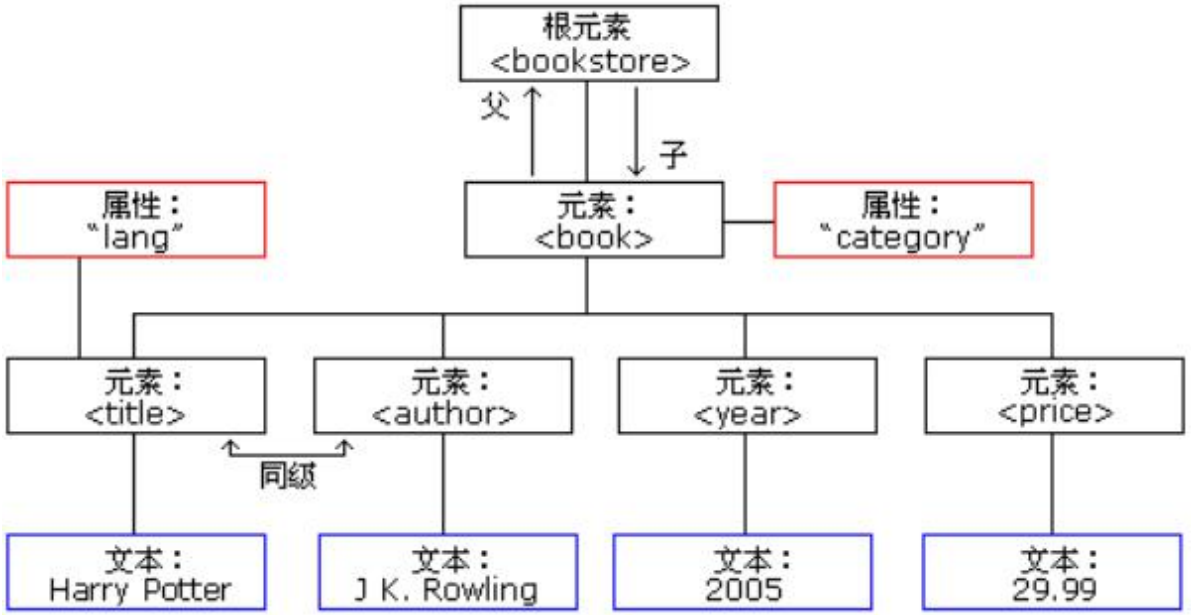
Java POI组件的xwpf

Python的python-docx

解析XML

- DOM
- SAX
- JDOM
- DOM4J

XML树示例


根元素
<bookstore>
父
子
属性：
"lang"
元素：
<book>
属性：
"category"
元素：
<title>
元素：
<author>
元素：
<year>
元素：
<price>
同级
文本：
Harry Potter
文本：
J K. Rowling
文本：
2005
文本：
29.99


开始
建立DocumentBuilderFactor
建立DocumentBuidler
建立Document对象，获取树的入口
建立NodeList
Node信息获取
结束

1

Word解析与脱敏技术

解析XML,采用BMHS和Word2vec
关键词匹配算法进行敏感词处理

2

文档恶意性检测

恶意代码、图片等可嵌入XML，采用
机器学习相关方法进行文档恶意性检测

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

目录

CONTENTS

4 在线文档管理

展示人：介来拉石

由来及发展

起源于协同办公

1984年，MIT和DEC公司的人提出了利用计算机辅助不同领域的人进行协同合作。
两年后开始每两年举行一次CSCW国际研讨会，汇集不同行业人士讨论计算机的辅助协同工作。

协同编辑

异步协同编辑

异步协同编辑允许多人在时间上分离地对同一个文档进行编辑，并通过加锁、版本控制工具如Ｇｉｔ、ＳＶＮ等机制来保证文档数据是一致的

同步协同编辑

同步协同编辑是让协作用户实时感知其他成员的编辑操作，既增加了用户并发度又提高了协同编辑效率，达到所见即所得的效果

一致性模型

因果一致性

两个前后操作 i 和 j，如果因果顺序相反，就产生了因果不一致性问题。

结果一致性

在协同中各节点执行了所有操作且都不编辑时，所有节点的共享文档的副本数据是一致的。

操作意图一致性

对于任意一个操作，其在协同系统中的任意节点执行之后的效果应该与其希望达到的效果一致

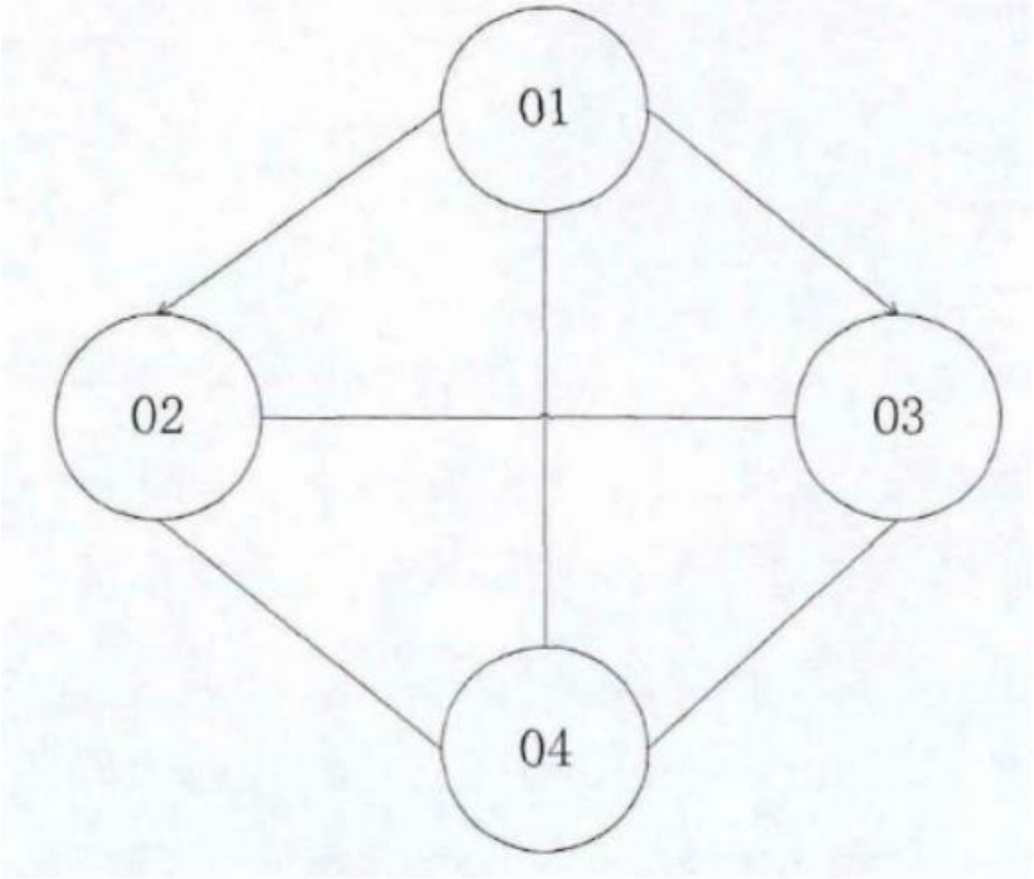
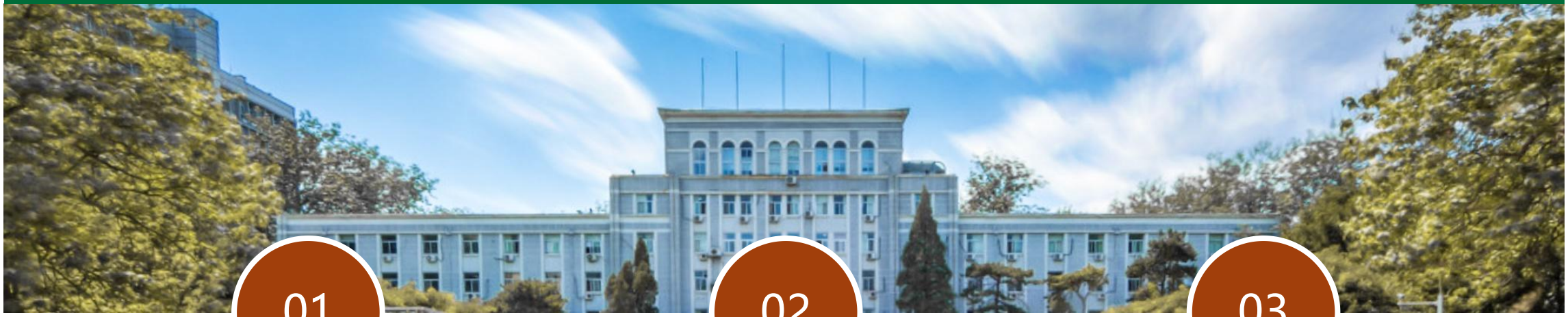


图 2-1 因果执行图

并发控制算法



01

串行化悲观延迟操作：直到其前面的操作全部执行完毕再执行，另一种是乐观执行操作，而后通过撤销重做。串行化方式不保证各节点的因果一致性和操作意图一致性。

02

令牌传递和加锁方式都是控制系统在任意时刻只能有一个正在编辑文档的用户来保证数据一致性，但这种方式限制了用户的操作，没有体现实时协作编辑系统多并发的优点。

03

操作转换(乐观)思想是让在节点生成的本地操作立即在本地副本执行，来自其它节点的远程操作到达本地后经过某些变换后再执行。

dOPT算法

dOPT算法是最早出现的基于OT的并发控制算法，引入了状态向量来解决因果一致性的问题。基础思想是将并发操作进行并发转换后顺序执行各个节点的操作。但这个算法采取了简单的节点优先级策略（优先级和节点标识相关，节点标识越大优先级越高）。

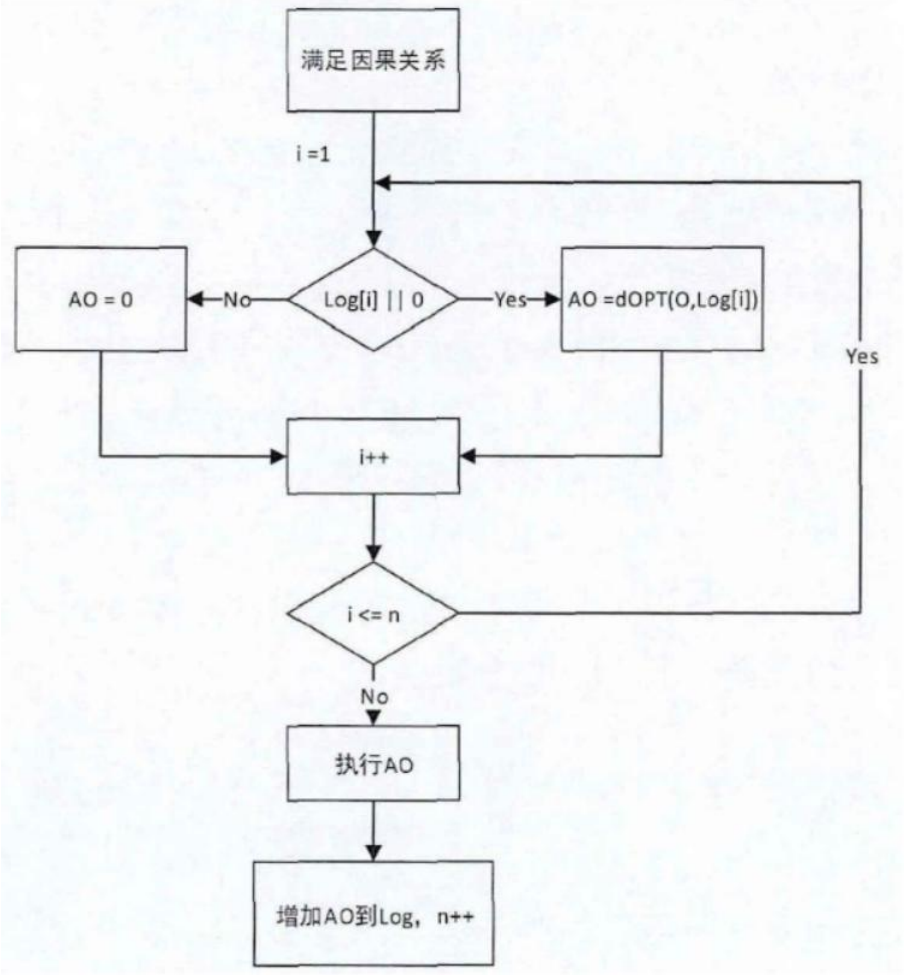
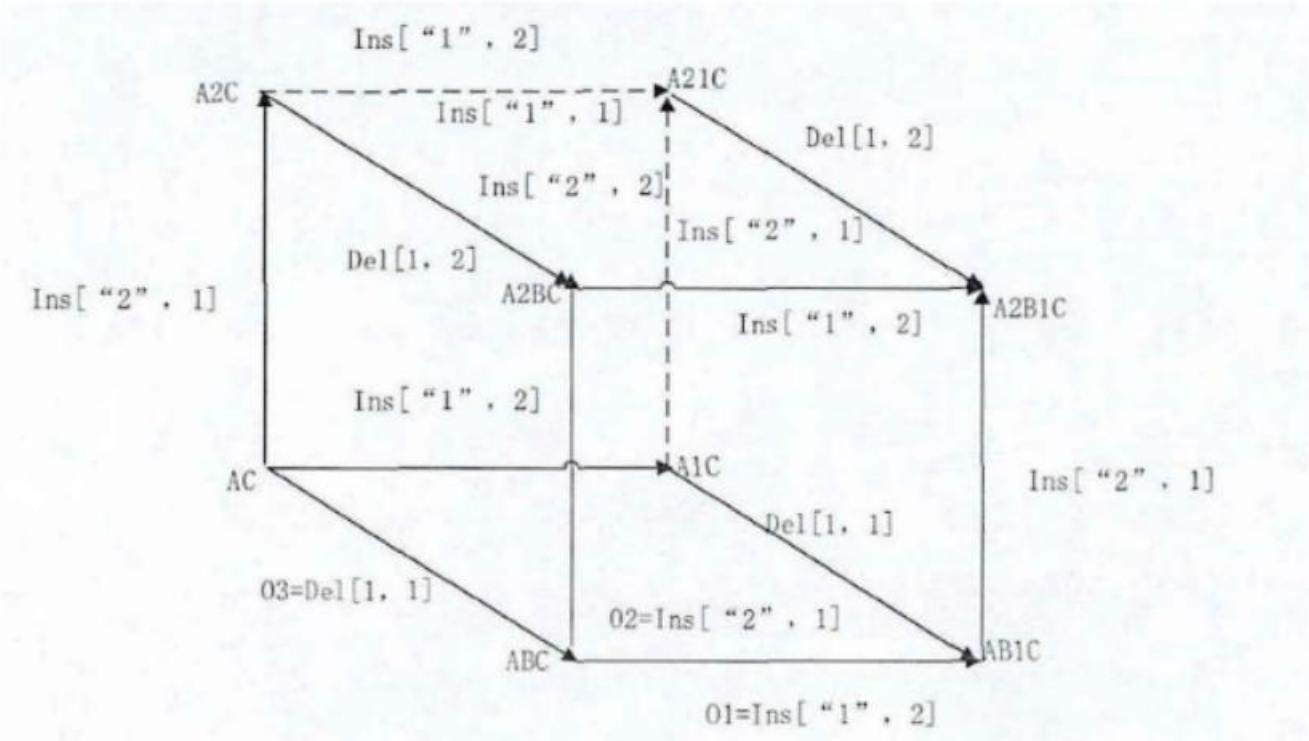


图 2-2 dOPT 算法流程图

adOPTed算法

adOPTed算法是Ressel等人在CSCW大会中提出的，算法的创新之处在于引入了 L 变换函数(L-Transformation),以及多维交互图来存储操作所有可能的形式，将所有操作的生成以及中间形式用线性日志存储，也保证了用户操作意图一致性与因果一致性。



adOPTed 算法多维交互图

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

目录

CONTENTS

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

5 基于区块链的文件管理

- 基于区块链文件管理的意义
- 区块链简介及发展历程
- 基于区块链文件管理的主要方法
- 基于区块链文件管理的项目展示

展示人：肖克

基于区块链的文件管理的意义

- 在第八届中国电子文件管理论坛上，“国家档案局副局长付华指出，近年来不断涌现的新技术对档案事业的发展提出了许多新挑战，其中之一便是区块链技术作为电子档案的保真技术是否可行”[1]

- 2017、2018、2020和2021年，国家档案局在科技项目立项选题指南中均设置了区块链技术研究选题，据统计2020年立项的120个项目中，区块链相关项目数量排名第二[2]

[1] 中国档案报，把握科技脉动，探索管理创新[EB/OL].[2018-4-13.http://www.saac.gov.cn/news/2017-12/15/content_216584.htm
[2] 国家档案局.2020年度国家档案局科技项目拟立项项目公示[EB/OL].[2020-07-17].http://www.saac.gov.cn/daj/tzgg/202007/09ea034a5cb.shtml

区块链简介

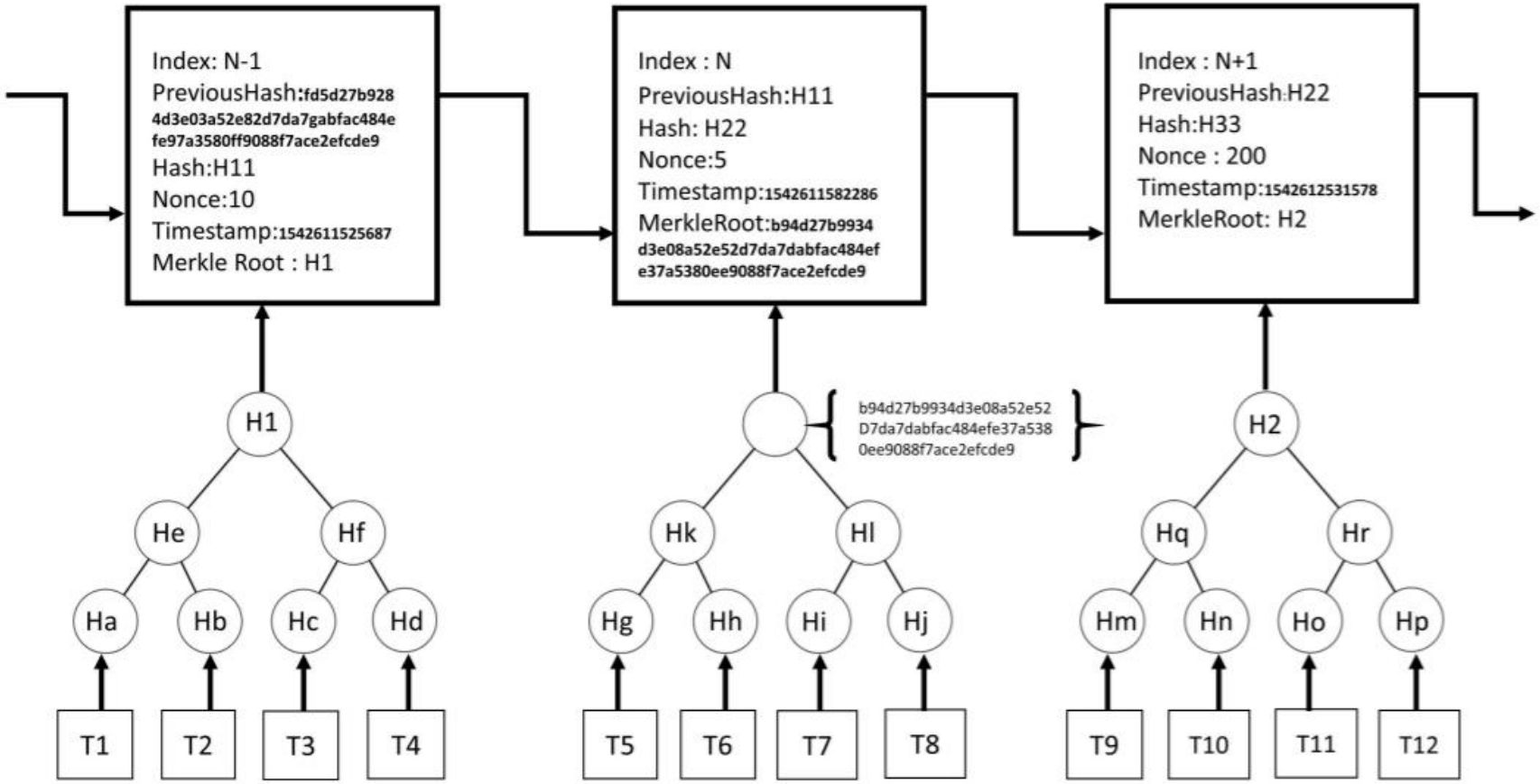


Fig1. Blockchain structure diagram

五、基于区块链的文件管理

区块链发展历程

Bitcoin: A Peer-to-Peer Electronic Cash System

Satoshi Nakamoto
satoshin@gmx.com
www.bitcoin.org

Abstract. A purely peer-to-peer version of electronic cash would allow online payments to be sent directly from one party to another without going through a financial institution. Digital signatures provide part of the solution, but the main benefits are lost if a trusted third party is still required to prevent double-spending. We propose a solution to the double-spending problem using a peer-to-peer network. The network timestamps transactions by hashing them into an ongoing chain of hash-based proof-of-work, forming a record that cannot be changed without redoing the proof-of-work. The longest chain not only serves as proof of the sequence of events witnessed, but proof that it came from the largest pool of CPU power. As long as a majority of CPU power is controlled by nodes that are not cooperating to attack the network, they'll generate the longest chain and outpace attackers. The network itself requires minimal structure. Messages are broadcast on a best effort basis, and nodes can leave and rejoin the network at will, accepting the longest proof-of-work chain as proof of what happened while they were gone.

bitcoin

eosio

- 区块链1.0：
  数字货币去中心化
- 区块链2.0：
  智能合约数字资产
- 区块链X.0 …

基于区块链的文件管理的主要方法

电子文件可信性要素：[1]

- 准确性
- 可靠性
- 真实性

区块链技术中共识机制、加密算法、时间戳等技术的应用能够在文件创建、保存、传输和存储阶段确保其凭证和记录作用的有效性。

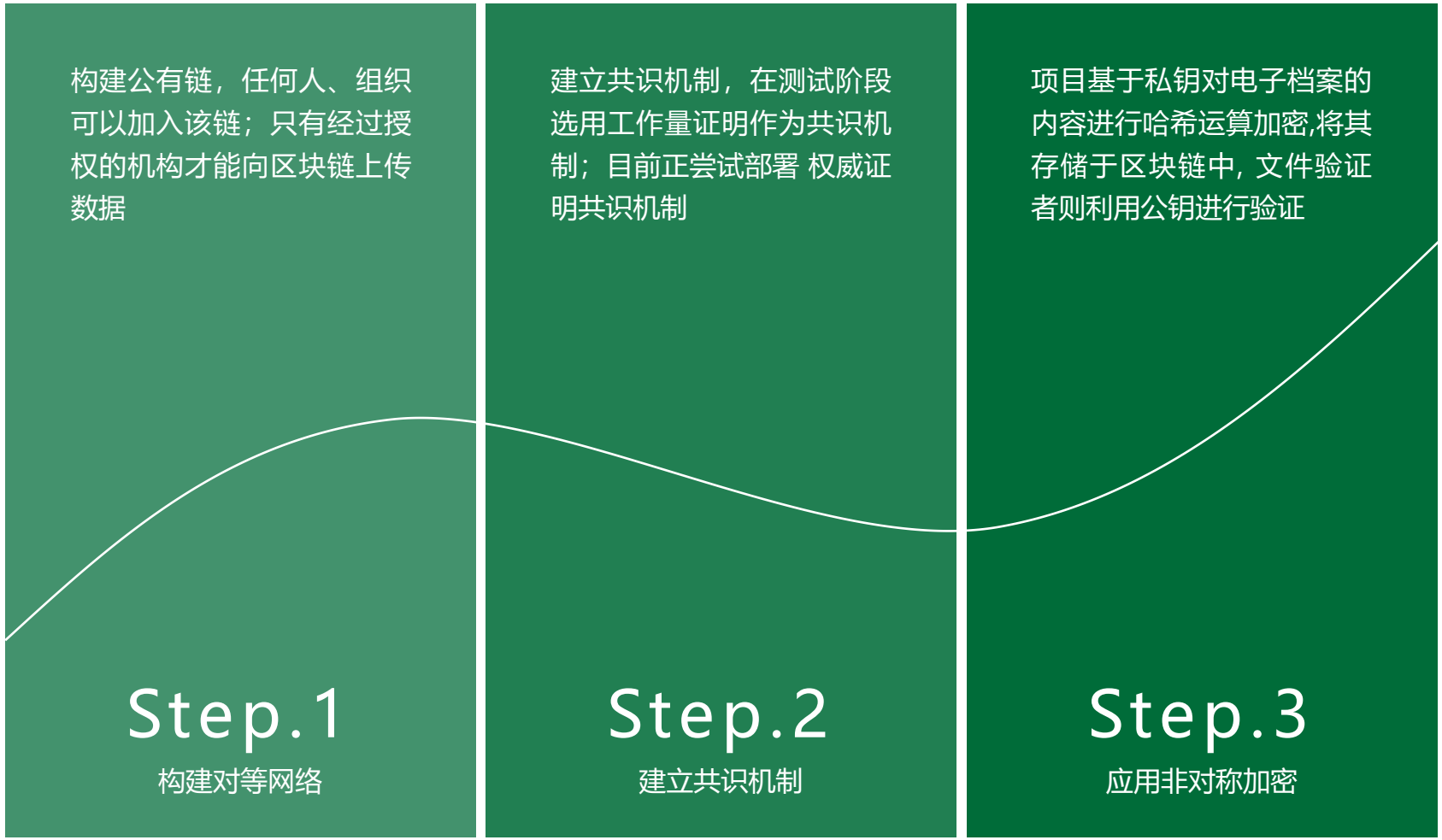
[1] Blockchain and Distributed Ledgers as Trusted Recordkeeping Systems: An Archival Theoretic Evaluation Framework[C]//Future Technologies Conference .2017.

五、基于区块链的文件管理

基于区块链的文件管理的项目展示

- **ARCHANGEL**

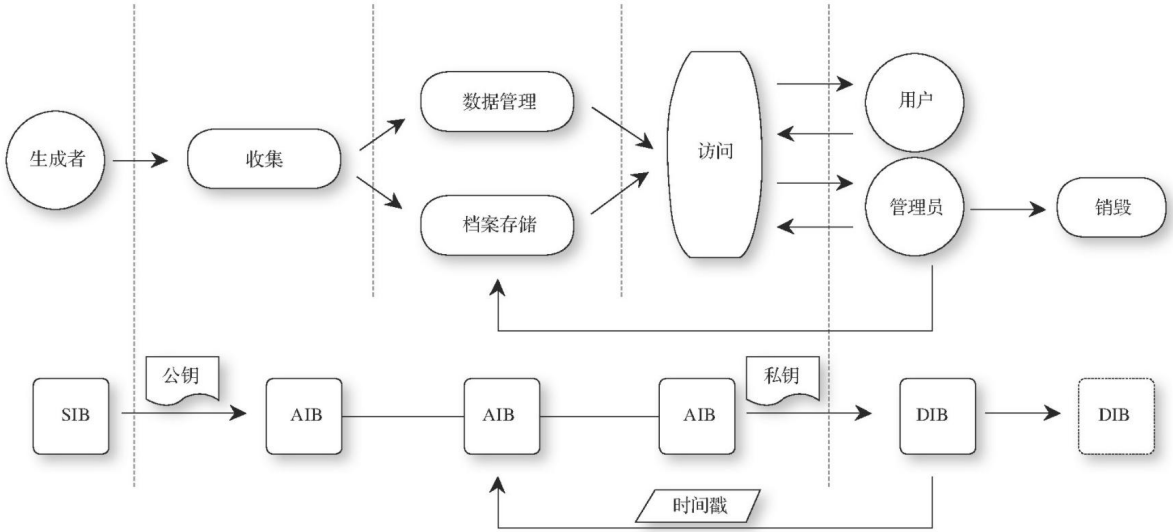
ARCHANGEL项目由萨里商学院 、英国国家档案馆、开放数据研究院等机构共同协作，保障文件完整性的分布式架构。



五、基于区块链的文件管理

基于区块链的文件管理的项目展示

- 基于区块链技术的电子文件可信保护框架



移交和接收

将电子文件按照SIB形式经公钥加密发送给文件管理者；接收方使用私钥解密，完成接收流程

存储和管理

文件管理者将文件解密后，更新时间戳并对文件进行二次封装，形成的AIB存储在分布式账本中

利用和销毁

需要时，用户可以从分布式账本中获取AIB进行解密；管理员定期提取DIB进行鉴定，决定是否销毁

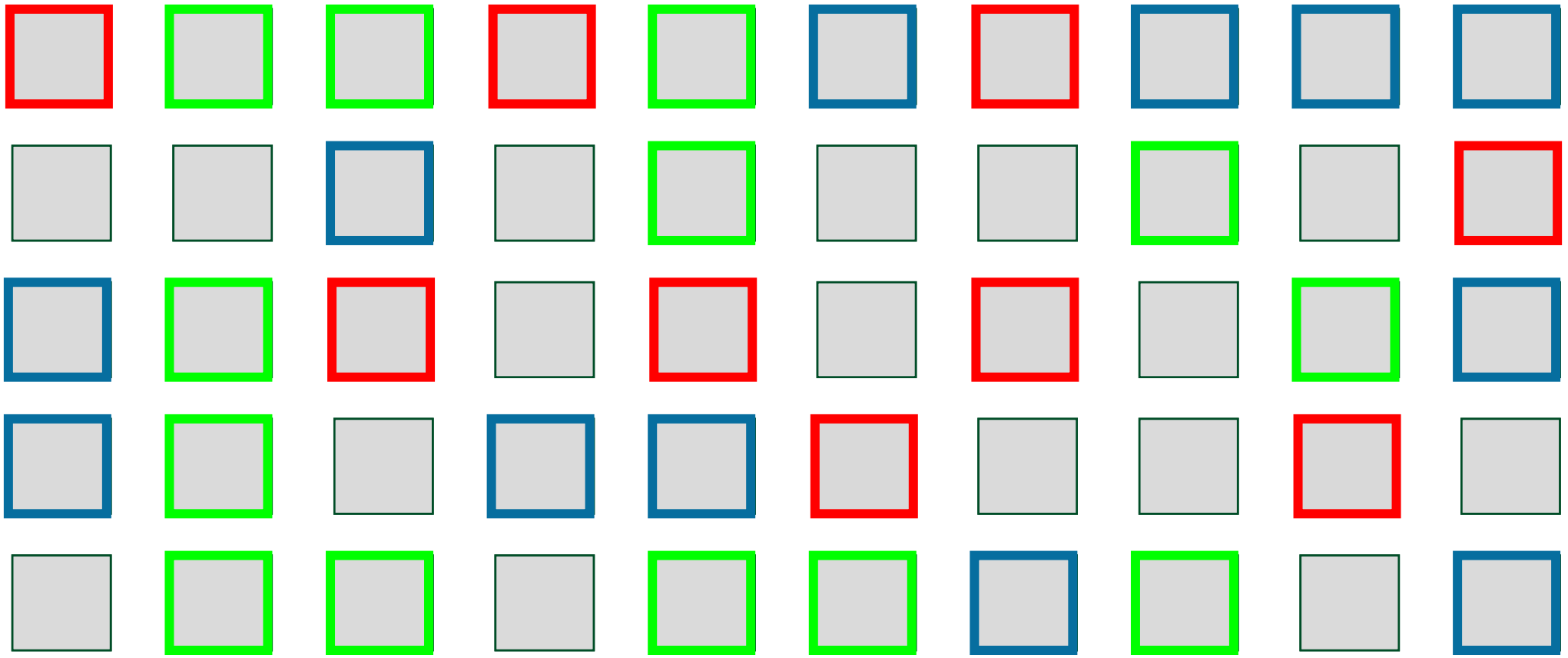
06

总结和代码演示

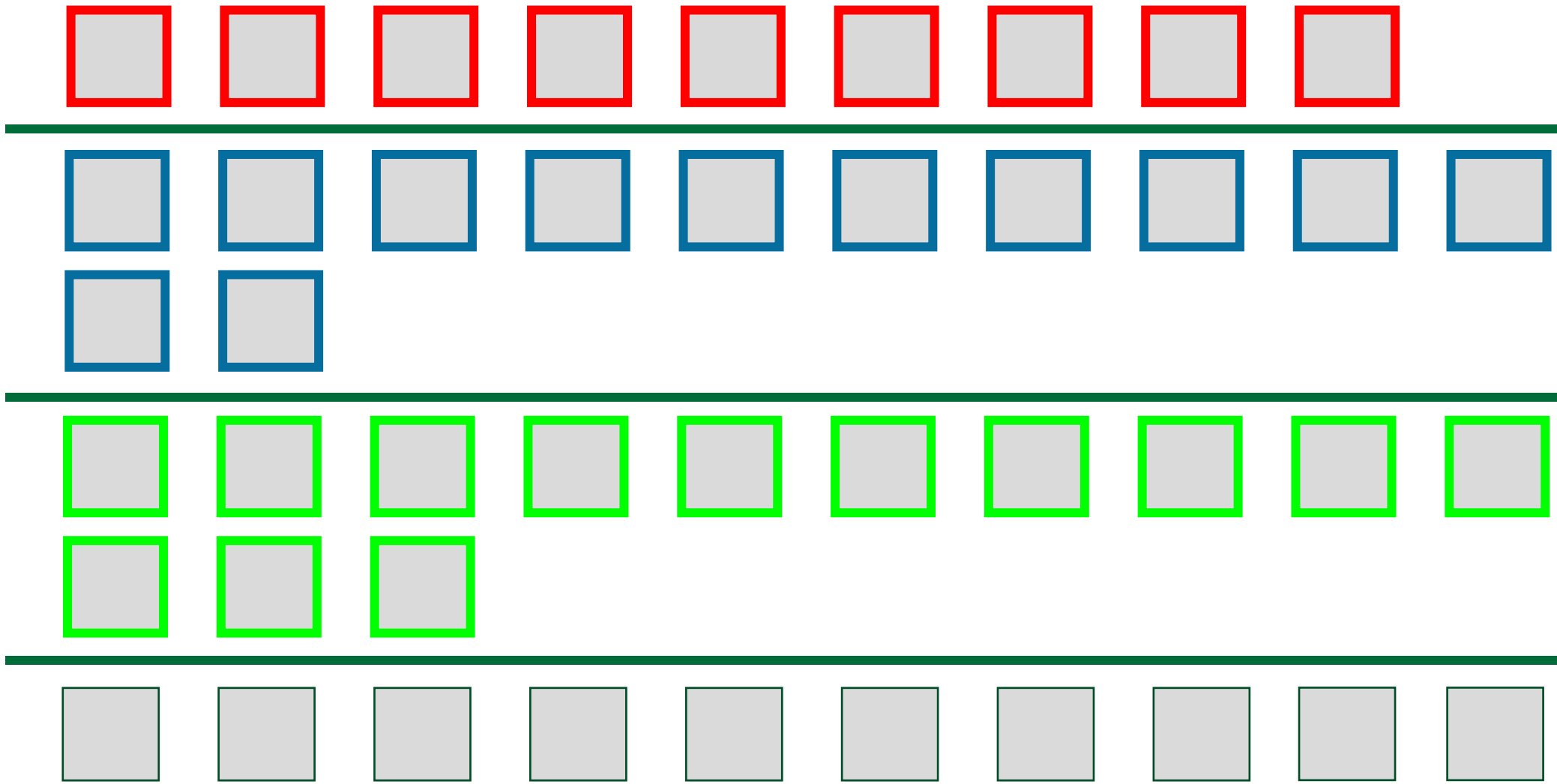
展示人：卫青

在代码演示前
我们来讨论一个问题

思考



思考



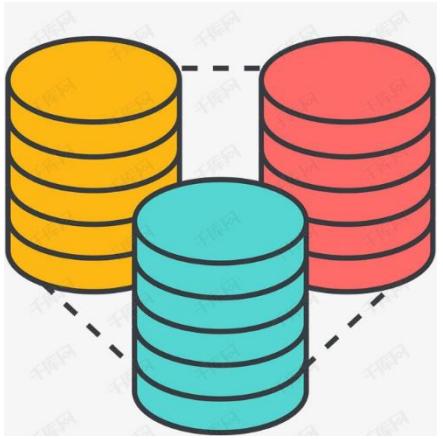
多格式文档解析与管理干了什么

多格式解析

统一管理

文档数据 → 信息 → 知识库

ID	文件	文体	描述
A16	ca16	新闻 news	Chicago Tribune: Society Reportage
B02	cb02	社论 editorial	Christian Science Monitor: Editorials
C17	cc17	评论 reviews	Time Magazine: Reviews
D12	cd12	宗教 religion	Underwood: Probing the Ethics of R
E36	ce36	爱好 hobbies	Norling: Renting a Car in Europe



代码演示

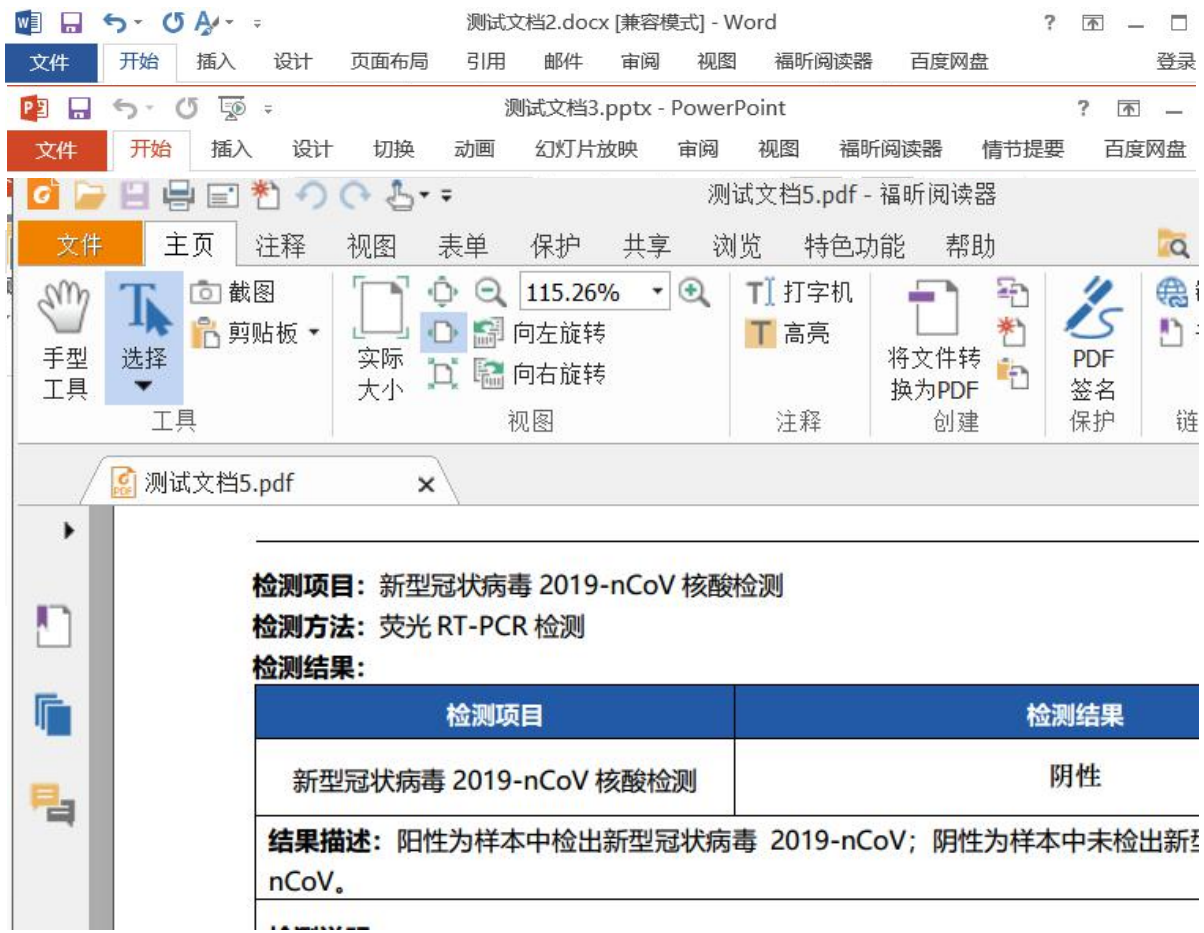
利用python_docx和win32com解析word文档

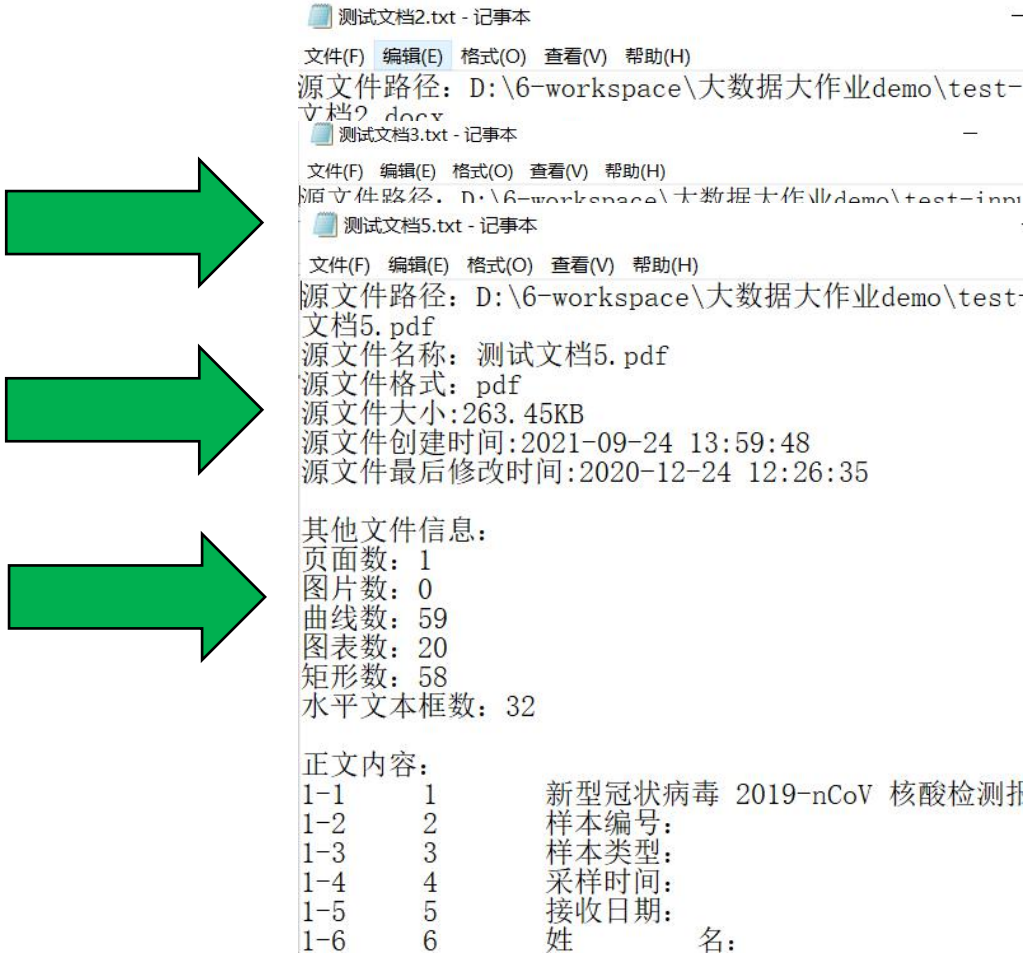
```
from docx import Document
from win32com import client as wc


def reSaveTodocx(doc_path):
    # 用此方法来兼容doc文件格式
    # 使用win32com模块调用word程序将doc文件转存为docx文件
    word = wc.Dispatch('Word.Application')
    doc = word.Documents.Open(doc_path)  # 目标路径下的文件
    doc.SaveAs(doc_path[:-4] + '.docx')  # 转化后路径下的文件
    doc.Close()
    word.Quit()
```

```
def read_docx(file_path):
    if file_path.endswith(".doc"):
        file_path=reSaveTodocx(file_path)
    # 打开docx文档
    docx_file=Document(file_path)
    # 正文之外的基本信息
    # # 读取页眉
    docx_head_pars=docx_file.sections[0].header.paragraphs
    docx_header=''
    for par in docx_head_pars:
        docx_header += par.text
    # 读取页脚
    docx_foot_pars=docx_file.sections[0].footer.paragraphs
    docx_footer=''
    for par in docx_foot_pars:
        docx_footer += par.text
    # 读取正文
    for i in range(len(docx_file.paragraphs)):
        # 按照段落分割标注构成文档内容
        content_para=str(i+1) + '\t' + docx_file.paragraphs[i].text + '\n'
```

代码演示


测试文档2.docx [兼容模式] - Word
测试文档3.pptx - PowerPoint
测试文档5.pdf - 福昕阅读器
测试文档5.pdf
检测项目：新型冠状病毒 2019-nCoV 核酸检测
检测方法：荧光 RT-PCR 检测
检测结果：
检测项目
检测结果
新型冠状病毒 2019-nCoV 核酸检测
阴性
结果描述：阳性为样本中检出新型冠状病毒 2019-nCoV；阴性为样本中未检出新型
nCoV。


测试文档2.txt - 记事本
源文件路径：D:\6-workspace\大数据大作业demo\test-
测试文档3.txt - 记事本
测试文档5.txt - 记事本
源文件路径：D:\6-workspace\大数据大作业demo\test-
文档5.pdf
源文件名称：测试文档5.pdf
源文件格式：pdf
源文件大小:263.45KB
源文件创建时间:2021-09-24 13:59:48
源文件最后修改时间:2020-12-24 12:26:35
其他文件信息：
页面数：1
图片数：0
曲线数：59
图表数：20
矩形数：58
水平文本框数：32
正文内容：
1-1 1 新型冠状病毒 2019-nCoV 核酸检测报
1-2 2 样本编号：
1-3 3 样本类型：
1-4 4 采样时间：
1-5 5 接收日期：
1-6 6 姓 名：

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

谢谢观看