Text and Data: Digitization, Preservation, Management, and Mining

文本和数据
数字化、保存、管理
挖掘与分析

CALIS 2020

Ann Arbor, Michigan

October 14, 2020

2020 Update from ProQuest and Ex Libris

ProQuest 和 Ex Libris 今年的最新进展

Ex Libris and Alma

- There are 1,800 Ex Libris Alma customers globally, 300 more since last CALIS.
- **全球已有一千八百多家Alma客户，从去年CALIS至今净增三百多家。**
- All Primo & Summon will be unified with Central Discovery Index in 2020.
- **年底前将完成用Central-Discovery-Index对Primo 及Summon的整合。**
- In Greater China, 150 institutions are in the ExLibris Higher-Ed Cloud.
- **150多所大中华区的院校已经加入ExLibris 的高教云端服务，其中有30多家Alma 客户。**

Ebook Central

- EPUB Reader was launched with in-book search and download in April 2020.
- **今年四月开始为用户提供 epub 图书**
- DDA Budget Tracker has been available since September 2019.
- **去年九月开始为图书馆员提供DDA经费使用的追踪管理。**
- Rapid Fulfillment with GOBI and Oasis since March 2020.
- **今年三月开始为用户提供快速的GOBI 和 Oasis 图书合同履行服务。**

ProQuest Platform

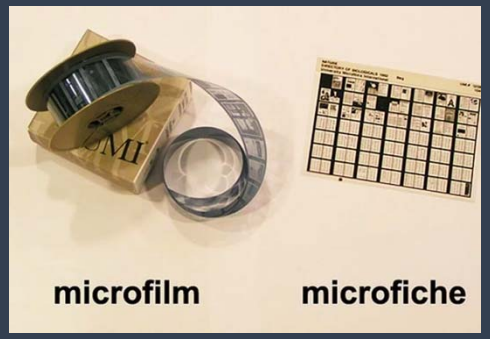
- After launching the world largest database ProQuest 1 Academic for multi-disciplinary studies, the company has made its discipline-specific product “ProQuest 1 Literature” available for literary studies and will deliver “ProQuest 1 Business” in December for business students.
- **继去年元月推出全球最大的综合文献库 ProQuest 1 Academic 之后，公司又为专业学科的本科生研制出了ProQuest 1 英美文学库和商业学科库 。**
- Supporting the need for the US and UK Government documents in China
- **协助国内院校图书馆及时订购所需的美国英国政府文献。**

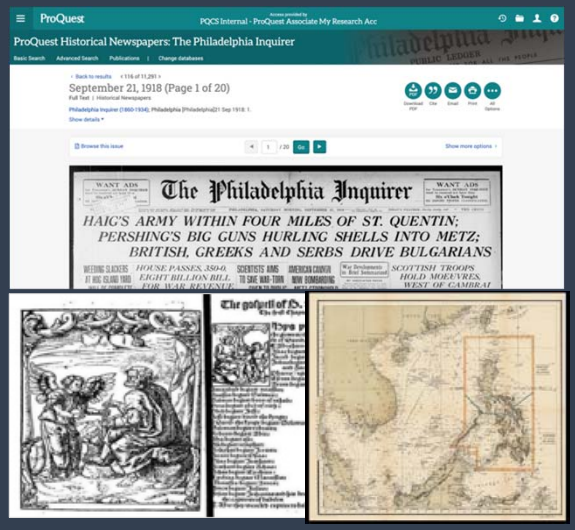
Text and Data
Digitization, Preservation, Management

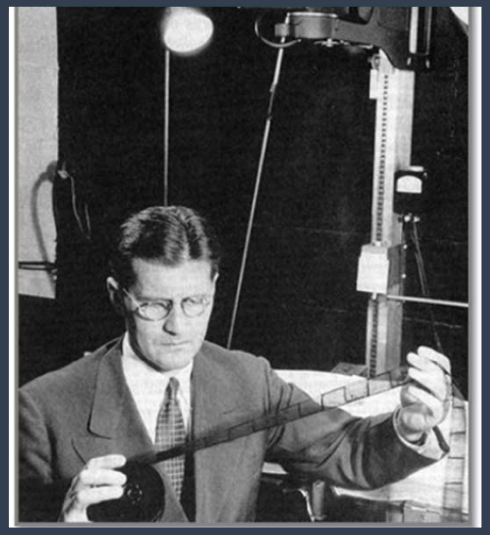
文本和数据
数字化、保存、管理

Paper Document: Preservation, Storage, and Digitization
纸质文档的保存、存储与数字化

- Preservation and storage of paper documents rely on microfilm and microfiche.
- 纸质文档的保存与存储需要缩微胶卷和缩微胶片。
- The technologies are mature for permanent preservation as microfilm and microfiche can last more than 500 years in a proper storage vault.
- 纸质文档永久存储的技术已经十分成熟，缩微胶卷和缩微胶片如果储存得当，保存期能够达到500年之久。
- Digitization from microfilm and microfiche is well established as well, and technologies still improve.
- 将缩微胶卷胶片上的内容转换成数字文档的技术也很成熟。技术还在不断完善。







Digital Document: Preservation infrastructure, Data Management, and Software
原生数字文档的保存所需的基础设施、数据管理及软件

- Born-digital files: Preservation requires computer storage and network systems, and tools for ingestion and processing, file management.
- 原生数字文档的保存需要计算机存储和网络系统、数据接入和处理工具、文档管理工具。

- Integrity of data: File inventory control, fixity checking, auditing and repairing.
- 数据文档保存的完整性需要文档库存管理、数据稳定性检验、数据审计和修复。

- Sustainability: Validation of file formats and media formats, and plan for those that will be obsolete in a near future.
- 数据保存的可持续性需要验证文档格式和媒体格式，对在近期内有可能被淘汰的格式，需要尽早做好准备。

5

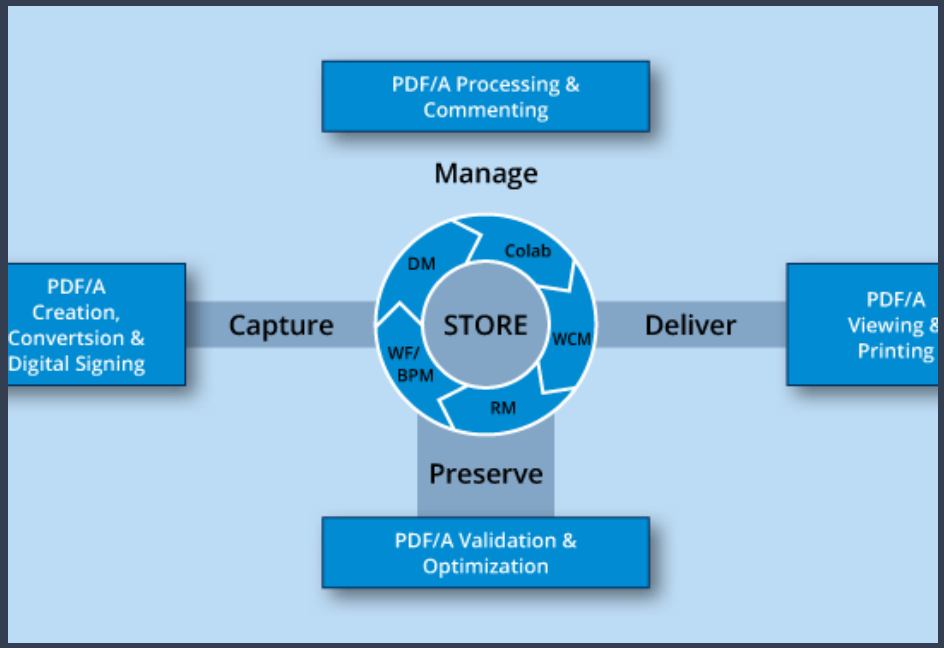
Digital Document and PDF: Preservation and Reuse
PDF在数字文档的保存及复用中的作用

- PDF, a standardized digital document technology, delivers the key qualities of paper in a digital format.
- PDF是一种标准化的数字文档技术，能够以数字化形式展现纸质文档的核心特质。

- PDF is fixed, self-contained, readily shareable and relatively hard to change.
- PDF格式具有固定性、独立性，可以随时分享且相对不易被改动。

- PDF/A is an ISO-standardized subset of PDF, to ensure that users see the exact same document both today and for years to come.
- 根据 ISO 标准，PDF/A 是 PDF 的一种特型，用于电子文件的长期保存。

  - Unlike normal PDF, PDF/A requires that everything necessary to precisely render the document is contained in the file, including fonts, color profiles, images and so on.
  - 与普通的PDF文档不同，PDF/A在自身文档内部嵌入显示该文档所需的信息（字体、颜色、图像，等等），以达到长期保存的目的。



USC Shoah Foundation

南加州大学犹太大屠杀基金会

The Foundation has created 4 “mirror sites” for the Visual History Archive, the audio-visual interviews with survivors and witnesses of the Holocaust and other genocides, guaranteeing that a fully-functional Archive will exist in perpetuity outside its home at USC.

影像历史档案库保存了大屠杀中幸存者和目击者的音视频采访记录。基金会建立了4个“镜像站点”以确保影像记录的永久保存。

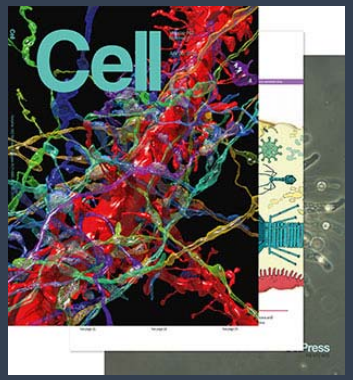
7

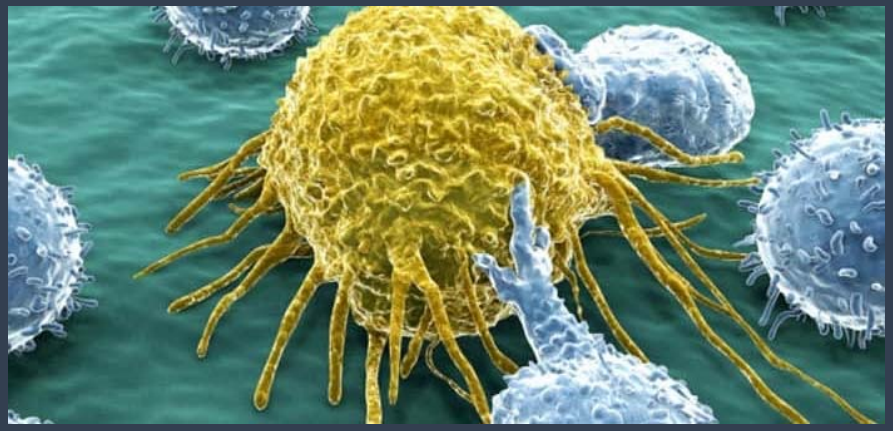
Audio/Video: Preservation Process and Infrastructure
音视频文档的保存过程及基础设施

- Inventory a collection of video files and their containers.
- 查点音视频库里的音视频文档及音视频容器。
- Capture the technical information related to a video file so its format can be identified, such as lossless codec, resolution, frame rate, etc.
- 采集音视频文件的技术信息（包括无损压缩、分辨率、帧率等）， 确认音视频格式。
- Assess what may be at risk for obsolescence, including media player software.
- 判断何种格式的音视频和媒体播放软件有可能被淘汰。
- Build a computing infrastructure to process and store huge volume of A/V data with rapid growth.
- 由于音视频数据量的庞大和迅猛增长幅度， 需要有规模型的计算机基础设施和镜像站点来处理和保存海量的数据。

Text and Data : Mining and Analyzing

文本和数据挖掘与分析





Why Do We Need TDM?

为什么需要文本与数据挖掘？

- Want to cure cancer?
- 在治疗癌症的研究领域有海量的文献

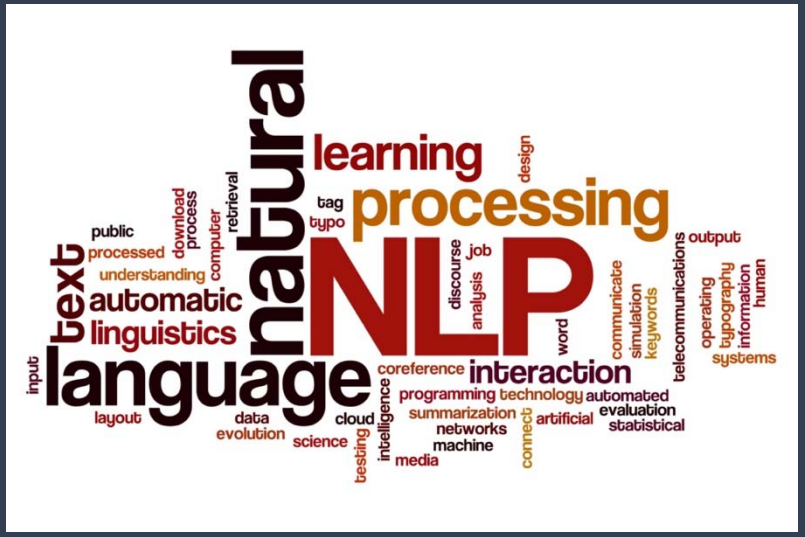
- 312, 308 articles in last 10 years
- 过去十年内有31万多篇论文

- Read 85 per day = 10 years
- 每天阅读研究85篇需要十年

- TDM = weeks or months
- 凭借 TDM 分析仅需几周或几个月

What is Text and Data Mining?
文本与数据挖掘的概念

- **Content and Data 内容与数据**
  - Select and format large amounts of text and data
  - 海量文本数据的**选择**与格式的**处**理
- **Analysis 挖掘和分析**
  - Search and identify relationships in data, and discover patterns
  - 通**过计**量型数据挖掘，**识**别数据中潜在关系和**规**律模式
- **Knowledge Discovery 知识发现**
  - Develop new knowledge and insights
  - **发现**新知**识**，形成新**观**点



Customer Pain Points In TDM
用户的TDM痛点

Access to Sought-After Content + Mining Tools

不易获取的重要内容、挖掘工具

Collecting Content & Converting it to Data for Analysis

费时费力的内容处理与数据转化

Library's New Role in Supporting Faculty and TDM

TDM教学研究仍然是图书馆的一个新职能

Teaching & Learning Digital Literacy in College Courses

如何在人文社科教学中引进数字分析素养

ProQuest TDM Studio and the Research Workflow
ProQuest TDM Studio及其研究流程

Select Content 1.选择内容

PQ Content , OA Content , Patents, User Uploaded Content

PQ内容、开放获取内容、用户上传的内容

ProQuest DISSERTATIONS & THESES GLOBAL

WSJ

Perform Analysis
2. 开展分析

PQ Methods
PQ分析方法

User-Generated Data & Methods
用户自己的数据和分析方法

python
常用分析工具

R

Manage Results & Export Findings
3.管理与导出研究结果

Sample export options (Graph, Table, Feature Set)

导出格式包括图片、表格、特征集等

Customer Private Cloud
用户各自的云端

ProQuest TDM Studio

Did Shakespeare Work with other Authors?
悲喜剧《两贵亲》是莎士比亚与别人的合作成果？

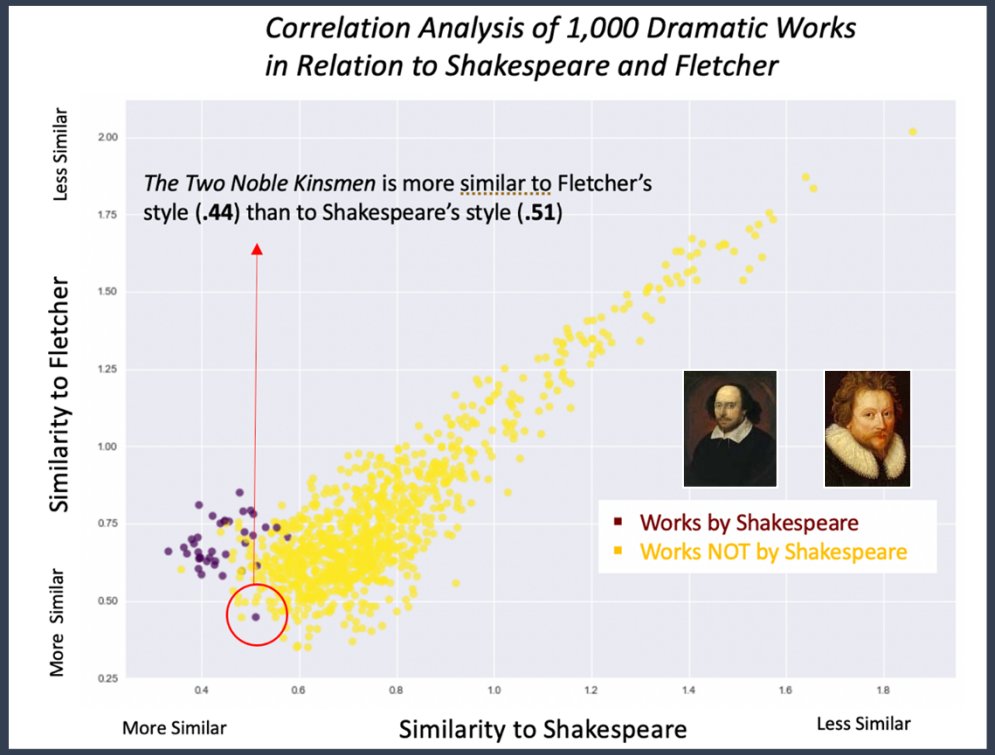
- Did Shakespeare collaborate with John Fletcher?
- 莎士比亚是否与约翰·弗莱彻（John Fletcher）一起创作了《两贵亲》？

- The tragicomedy *The Two Noble Kinsmen* has been attributed to Shakespeare, but it is closer in statistical style to Fletcher in this TDM study.
- 这出悲喜剧一直被认为是由莎士比亚独自创作的。但是根据TDM研究，该剧的文字计量风格与弗莱彻的更为相似。

- More likely, *The Two Noble Kinsmen* was co-authored by Shakespeare and Fletcher.
- 《两贵亲》极有可能是莎士比亚与弗莱彻共同创作的。

A Study Using ProQuest TDM Studio

使用 ProQuest TDM Studio 的研究案例



Is Firm's Underperformance Related to its CEO's Overconfidence?

CEO过分自信是否会导致公司业绩问题？

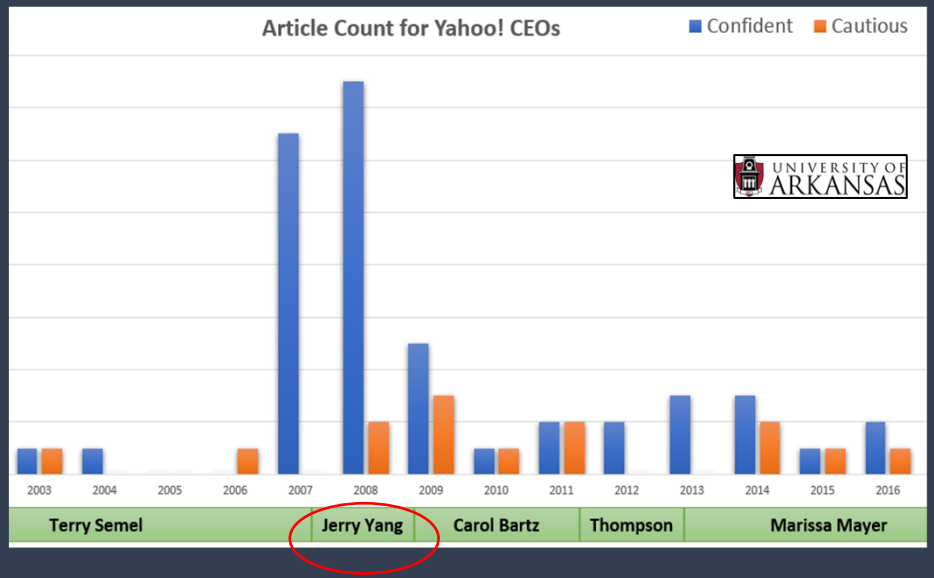
- Professor Caleb Rawson performed a TDM study, instead selecting a few CEOs and reading every article about them as business faculty usually do.
- 与多数商学研究人员不同，凯勒·罗森没有选择研读关于公司CEO的文章，而是开展了一项 TDM研究。
- He identified 2,500 CEO/firm pairs and used the list to select articles from a set of leading newspapers. Next, he created a corpus of 22,443 articles by filtering down an initial set of 323,453.
- 罗森教授选择了两千五百个CEO和公司的派对，在主流新闻报纸中选取了32万多篇有关CEO及公司的派对文章，然后从这些文章中筛选出22,443篇并组建了一个语料库用于计量分析。
- He concluded that overconfident CEOs talked more about their R&D activities, leading to worse performance of their own firms.
- 他的研究结果显示，过分自信的CEO喜欢透露自己的核心研发活动，会导致公司竞争力不佳的局面。

A Study at University of Arkansas Using ProQuest TDM Studio

阿肯色大学使用ProQuest TDM Studio的研究案例

Overconfident Jerry Yang Led Yahoo! to its Downturn

杨致远的过分自信与雅虎公司业绩的下滑



ProQuest TDM Studio and its Customers

ProQuest TDM Studio及其用户

“ProQuest has been an amazing partner in setting up a TDM infrastructure that is both powerful and easy to use.”

“ProQuest是出色的合作伙伴，ProQuest提供的TDM基础设施功能强大、操作简便。”

•耶鲁大学经济学助理教授John Eric Humphries）

“In the medical field, we often don’t have the manpower to go through tedious manual processes to analyze data, which can be frustrating. By text mining a large corpus of material, TDM Studio is helping us…”

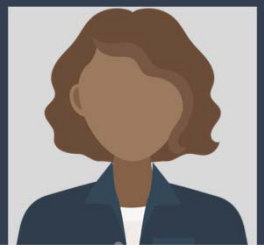
“在医学领域中，我们通常缺乏人手去人工处理分析大量数据。这十分遗憾。但是TDM Studio能够帮助我们，通过挖掘语料库中的文本数据……”

•哥伦比亚大学（Columbia University）副研究员Sunmoo Yoon

ProQuest TDM Studio Update
ProQuest TDM Studio的最新进展

Available to Customers since April
已经在今年四月推出

TDM Analysis Using R or Python
使用R语言或Python进行TDM分析



TDM SCHOLAR
TDM研究人员

TDM EXPERT
TDM专家

TDM BEGINNER
TDM新手

TDM LEARNER
TDM学习者

TDM Analysis with Guided Tools
使用导向工具学习掌握TDM分析

To be Launched in December
即将在年底推出

Thank You
谢谢

Allan Lu
Vice President
Research Tools, Services & Platforms
ProQuest LLC