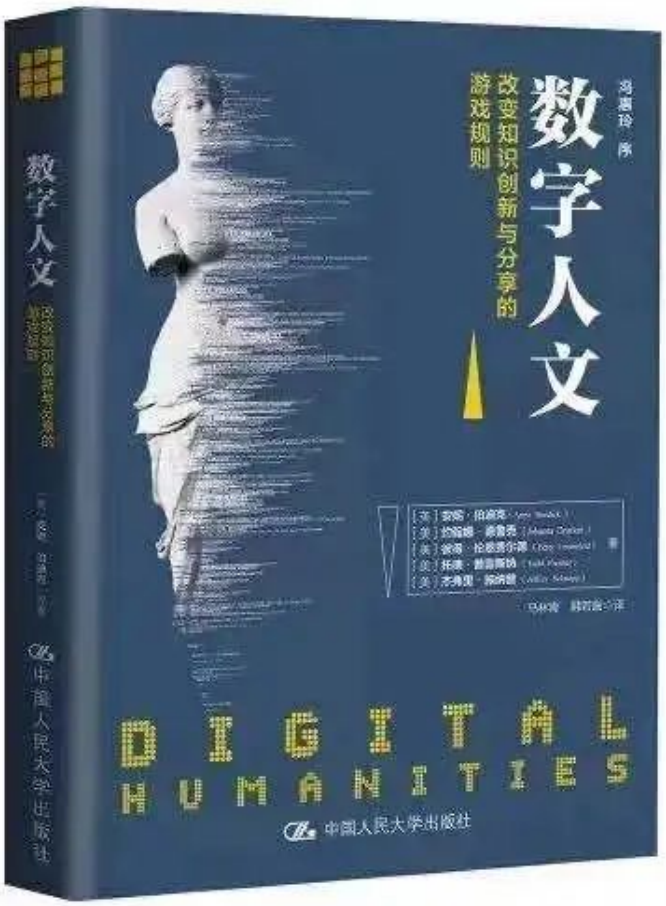
“DH精读”为数字人文优秀著作内容精选连载栏目。本文出自《数字人文 改变知识创新与分享的游戏规则》一书中的“案例研究2 来自亚历山大图书馆的纸莎草残片及文本语料库的扩展出版”一节。



作者 Anne Burdick， Johanna Drucker，Peter Lunenfeld，Todd Presner， Jeffrey Schnapp
译者 马林青 韩若画
中国人民大学出版社2018年出版

该文本语料库项目将建立一个链接现有资源库的集合，利用特定的基于文本的注释平台来讨论文本变体，执行概率性的自然语言处理，使用校对工具研究变体，并生成一个关于纸莎草残片的扩展的评述版。在此过程中，将用到几种不同的传统的和扩展的出版模型，从而使不同背景和专长的学者以适当的方式展示其研究成果。

一个新的纸莎草文稿残片贮存地发现于埃及的亚历山大。这些残片虽然由于受到忽视和磨损而遭受了相当程度地损害，但它们有望解答一些由来已久的问题，例如关于腓尼基字母的传播和整个北非尤其是西部沿海地区采用它的时间，以及它沿着贸易路线传播到印度的可能性。这些纸莎草令人惊讶的特点是，它们在使用的过程中被多次循环利用，有许多是重写本，有些还有着若干层辨识度不同的手迹。

目前，有不少可用的成像技术和可整合纸莎草数据的数字平台。这个项目的一项主要工作就是重用已成功用于西闪米特人碑文研究的一些技术。然而，这些纸莎文稿中的语言不只是闪米特人语言，令参

与该发现的图书馆员惊讶的是，这些文稿很好地记录了目前尚未被识别的几种语言的使用和传播。一位学者提出印欧语系根源于语言结构的组织，这意味着欧洲与印度次大陆有更早的接触，并且有比先前所认为的更显著的文化传播和影响。在使用任何自然语言处理技术之前，这些文本首先需要被译解（这是由于它们自身状况欠佳）。处理古代语言尤其是文本残片的NLP技术尚在实验阶段。此外，一些推测性和概率性阅读的方法也将被用于这些纸莎文稿。

该研究的问题是识别这些纸莎文稿中出现的语言组，利用已知的闪米特人碑文语料库匹配书写形式，利用支持跨越版本、翻译及书写形式进行数据挖掘和文本分析的数据库跟踪变体。这项工作产生的一个附带好处是建立一个纸莎文稿的数字化语料库。这项工作面临的其中一项困难是使用早期印欧语的一个主要人物年迈多病，不能长途跋涉，他的贡献必须完全依靠数字代理完成。我们必须建立一个可以注释和跟踪其贡献的平台，或者对现有的平台进行改造以达到这个目的。这有助于激励处于职业生涯中期的学者和年轻学者加入团队，他们对语言变化和传播问题的构想至关重要。他们需要一种扩展的出版形式，这种形式可以快速公布工作成果，同行评审周期短，承认不同程度的知识贡献，可以链接到其他现有的语料库和资源库，且鼓励竞争精神。

方法

为有效地开展项目，利用众包方式进行翻译、解释和编辑是一种可能的选择。对大型语料库的分析和对比而言，统计方法也是很必要的。因此，在对文本、人工制品和手稿的研究中，近距阅读技术和远距阅读技术都会用到。该项目的第一阶段将对成像和数字化进行充分整合，所有不确定的符号或图形元素都将被标注出来，使该项目凭猜测所做的部分工作易于识别。通过使用可显示大量对象的文化分析平台和图形识别软件，书写形式的相似性将被用于寻找语言的相似性。文本翻译尽量顺畅，其中变体和有争议的元素将被显著标记出来。众包工作中的零散成果将定期出版。资深学者的解读将在网上被分层展示，以便于他的工作能保持独立，且可于后期从网上去除后以纸本印刷的形式出版。

工作计划

· 确定图形技术和设备来源

· 为共享访问建立合作伙伴关系

· 创建图形文件并测试整合和对比技术

· 测试用于碎片化数据集合的概率文本分析方法

· 确定在需要时对来源文本进行翻译

· 链接到现有资源库和在线翻译

· 测试注释和版本控制平台

· 测试文化分析和模式识别软件

· 为短周期贡献引入同行评审系统

· 为翻译、编辑和解读创建出版和众包平台

· 进行成像、翻译和解读的迭代过程

· 以扩展版的形式连续发布成果，包含链接、与现有语料库的比较，以及文本的其他版本

传播和参与

创建Twitter订阅和RSS订阅宣传该项目并吸引更多人参与；以在线印刷+模式发布一个该项目的测试版，并建立工作流重新组合这些内容使之符合传统出版的要求；将翻译、比较和解释工作进行众包；基于此领域学者指出的与其他现有纸莎文稿、文本或古代手稿和语言碎片之间的关联或比较，持续扩展纸莎草文稿版本。该项目将通过聚合引擎链接到近东、中东、欧洲和美国的主要资源库，以建立用于调查和比较的更大规模的统计样本。通过分布式知识生产方式建立传统与新型学术参与模式之间的桥梁，使资深学者可以与年轻学者有效合作，并方便众包输入。

评估

学术的技术、知识和文化/制度等各个方面相互依存。成功在一定程度上取决于译解完成的程度、各图层的易读性，以及成像和文本分析技术的可信度。成功的另一个衡量标准是：可链接在一起的数字化片段的数量，和/或古代手稿的翻译数量。

注：本文为虚构案例，目的是为构建团队、配置必要的技术资源，以及在跨学科和跨机构环境中开展项目提供模型。

To be Continued

编辑 徐碧姗

排版 邵亚伟

公众号账号：rucdh2019

网址：http：//dh.ruc.edu.cn

邮箱：rucdh@ruc.edu.cn

中心简介

中国人民大学数字人文研究中心集人民大学多学科优势，秉持融合文理、协同创新之理念，开展数字人文理论研究、实践探索、人才培养和学术交流。