手写或印刷在薄薄宣纸上的方块汉字走出“深闺”，古籍数字化像一把钥匙……

打开“用”与“藏”环环相扣的铁锁

文化观察

□ 本报记者 卢昱

40余种珍贵宋元刻本、写本，著名藏书楼嘉业堂、密韵楼的抄本，文澜阁《四库全书》零本……近日，这批珍藏于美国加州大学伯克利分校的中文古籍善本，以数字化方式进驻“汉典重光”古籍平台（网址为https://wenyuan.aliyun.com/home）。

古籍中那些手写或印刷在薄薄宣纸上的方块汉字，经过数字化，飘起在“云端”，文化积淀又有了新的保存与光大的途径。

钱钟书的敏锐与远见

据统计，目前全国各公藏单位拥有古籍总量超过5000万册，需要修复的古籍约1500万册。即使在古籍不再继续遭到破坏的前提下，以当前的修复人才及修复条件计算，要完成全部修复工作仍需数百年。

古籍文献集文物价值和学术价值于一身。从保护的角度看，古籍应在合适的环境中收藏，尽量减少在普通环境中的时间，降低使用时可能带来的损伤。有测试表明，一部宋元古籍，离开专用书库，置于普通阅览室中供人翻阅一小时，其寿命就会缩短数月。

从利用的角度看，古籍若沉睡在库房，就无法发挥其价值，而且许多学者的研究与古籍内容息息相关。此时，古籍数字化像一把钥匙，打开了“用”与“藏”环环相扣的铁锁。

事实上，“古籍”与“数字化”已相遇三十余年。

古籍数字化，最初扎根在红学研究领域。在1980年国际红学会议上，美籍华裔学者陈炳藻提交《从字汇上的统计论红楼梦的作者问题》，提出用计算机统计《红楼梦》的字词，以辅助确定《红楼梦》尤其是后四十回作者的问题。国外的这些信息激发了国内部分学者的兴趣，他们开始关注并尝试将计算机技术应用于人文研究。

受此启发，镇江的红学研究者彭昆仑开始利用计算机程序探讨《红楼梦》人物年龄的问题。1985年他调到镇江市科委后，又与东南大学（原南京工学院）合作完成《红楼梦》数据库。此后，深圳大学建成“红楼梦多功能检索系统”。

1980年前后，钱钟书的女儿钱瑗到英国访学，看到英国学者用电脑储存、查阅莎士比亚的资料。回国后，她把这一信息告诉钱钟书。钱钟书敏锐地意识到这一新鲜事物的价值，让助手栾贵明从事相关研究。

在钱钟书的指导下，栾贵明主持的课题组先后完成“《论语》数据库”“《全唐诗》速检系统”等课题，并荣获1990年“国家科技进步奖”三等奖。这些都是利用计算机进行人文研究的早期实践。

当时，古籍数字化还在萌芽状态。在1987年12月人民日报出版社出版的《论语数据库》一书卷首，钱钟书写道：“从理论上来说，计算机和人类使用过的其他工具没有什么性质的不同。它在还未被人广泛使用的时候，除自身尚待完善以外，总会遭到一些抵拒。惯用旧家什的人依然偏爱着他们熟悉的工具。有了纸墨笔砚‘文房四宝’，准还有人用刀笔和竹简；有了汽车、飞机、电报电话，也还有不惜体力和时间的保守者。对新事物的抗拒是历史上常有的现象，抗拒新事物到头来的失败也是历史常给人的教训。”——当前古籍数字化的潮流，可说是对他远见的最好褒奖。

数字古籍为“母本”代言

上世纪九十年代后期，古籍文献数据库的建设步入快车道。

1996年，书同文公司启动的文渊阁《四库全书》电子版是一个标志性工程，被誉为大型中文电子出版工程的典范。该工程动用300名校录人员、60名技术、学术和管理人员，历时三年多完成。

而今，国家图书馆的“中华古籍资源库”已在线发布超过3.3万部的古籍影像；中华书局的“中华经典古籍库”已发布3000多种、15亿字的点校本古籍；爱如生公司的“中国基本古籍库”收书1万种，既有可供检索的全文，又提供古籍原版图像；像家谱、方志、中医药等专类古籍在多地兴起……

除了以上大而强的综合数据库，在网络上，很多古籍爱好者，出于热爱和自觉，建立古籍数据库，如“书格”“殆知阁”等，与以上数据库多头掘进，共同成为诸多文史研究者的助手。

近年来，古籍数字化在服务于学术研究方面，立功颇多。比如中南民族大学王兆鹏主持的“唐宋文学编年系地信息平台”、浙江大学徐永明团队与哈佛大学共建的“学术地图发布平台”、中国社科院刘京臣的“宋代文学地图数字分析平台研究”等值得关注的数字人文成果，其平台的建设离不开数字化古籍的基础作用。

而数字化之后，古籍“母本”不再需要冒着各种风险“抛头露脸”。与此同时，数字化的古籍，可以走出“深闺”，像孙悟空一般实现七十二种变化，在不同时间满足不同地域读者的阅读需求，实现一对多、点对面、虚对实的变化。

在山东，古籍数字化的步伐也在同步跟进。据山东省图书馆历史文献部主任、研究馆员杜云虹介绍，省图在2013年发布“山东省图书馆古籍珍本数据库”，收入数字化古籍资源近1000种，共计10万余拍，内容涵盖从明代至民国不同时期、不同类型的经史子集四部类古籍资源；2018年，省图将数字化的馆藏《永乐南藏》1600余部佛经、204592拍、587764页，在网络上公开发布；目前，省图已完成“易学古籍数据库”建设，实现6164种易学古籍书目在线检索和其中900种易学古籍的数字化……

“现在，古籍普查工作还没有最终完成。我们要摸清家底，古籍数字化不是一朝一夕的事儿，要在保护好古籍的前提下，做好规划，清楚哪些工作是最迫切的，不能零打碎敲地做，更不能盲目开发利用。”杜云虹说。

对于如何用好“在云端”的古籍宝库，杜云虹分析道：“怎么让古籍里的文字活起来，让大家觉得不很遥远，有很多工作要做。央视的节目《典籍里的中国》，讲述典籍传承文明的故事，是很好的尝试。”

当阿里涉足古籍

“电商巨头阿里涉足古籍行业。这在电商行业意味着什么我不清楚，但在古籍领域确实算得上一个大新闻。”网友“人生五味”评价道。

阿里巴巴达摩院院长张建锋表示，达摩院自2017年起接触古籍数字化领域，2019年正式参与由阿里巴巴公益基金会、四川大学、美国加州大学伯克利分校、中国国家图书馆、浙江图书馆合作开展的“汉典重光”项目，旨在寻觅流散海外的中国古籍并将其数字化、公共化，让普通人也能亲近古籍，通过古籍与先贤对话、与优秀传统文化对话。

目前，首批20万页古籍已完成数字化，并沉淀为覆盖3万多字的古籍字典，公众可通过“汉典重光”平台翻阅、检索古籍。记者打开平台网页，试着搜索“山东”“济南”等关键词，跳出《战国策》《通鉴纲目》《河防一览》等结果，皆可定点查询、锁定。相较于其他成熟的数据库，“汉典重光”后台的数据量还偏小，在使用时也有一些不够流畅之处。

新潮的阿里似乎对陈旧的古籍还不太熟悉，但这种“相逢”正探寻着古籍数字化的新路径。

据悉，古籍数字化大概有以下流程：采集侧，将纸质书变为电子扫描版；生产侧，将电子扫描版变为文字版；应用侧，将文字版变为古籍研学系统，涵盖检索、字典、知识图谱等功能。

目前，古籍数字化在采集侧、生产侧有两种方法。第一种是纯人工录入，如一本书有10万字，人工把10万字输入计算机。像《四库全书》的编修，就是纸书时代的“人工录入”，当年在乾隆皇帝的主持下，纪昀等360多位高官、学者参与丛书编修，一共用了3800多人、耗时13年才完成。《四库全书》包含3462种书、7.9万余卷、3.6万余册，总字数约10亿。在当下，已很难找到并组织众多精通古文字的专家，如此专注、数十年如一日地来做录入工作。

第二种是计算机与人工结合，计算机利用文字识别技术提取一部分文字，计算机无法识别的文字则由人类专家手动录入，最终再由人工进行检校。这一技术路线虽探索多年，但始终没能让识别效率大幅提升。原因主要在于：计算机能识得的古籍文字极为有限，若用传统的机器学习方法“教会”计算机海量的古籍文字，得先提供海量的标注数据，用于训练识别模型。而古籍文字没有现成的标注数据，需要懂古文的专业人士手动标注，可能比人工直接录入的工作量更大、成本更高。

面对海量无标注的数据，如何让AI（人工智能）快速批量识别古籍，始终是古籍数字化领域的技术瓶颈。对此，阿里巴巴达摩院技术团队与四川大学专家联手，在第二种技术方法的基础上，研发了一套全新的识别系统。

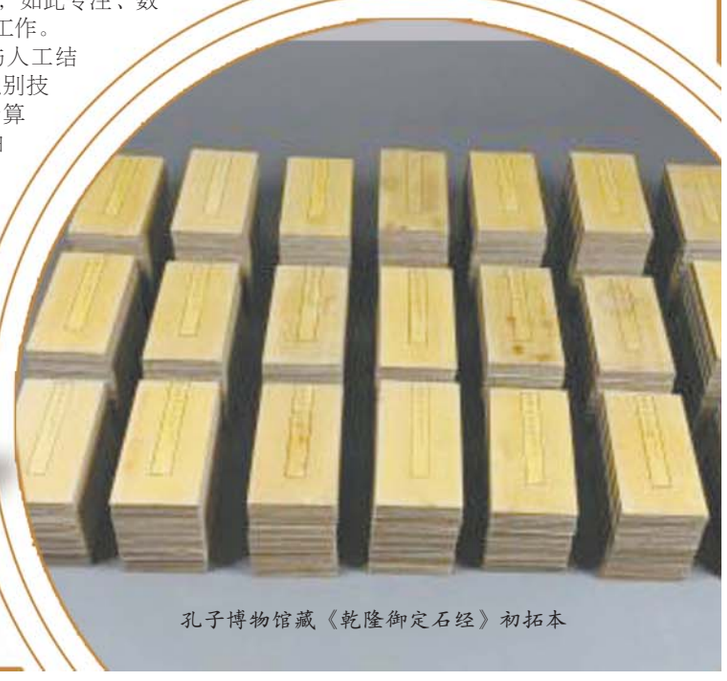
首先是全书检测，把古籍正文中的每个字都抠出来，作为单独的一张图；然后进行聚类，一本古籍总字数可能有10万字，但其中有很多字是重复的，比如“之”“乎”“者”“也”等，聚类就是让机器自动把字形笔画一致的字归为一类，接着再由专家进行标注。原本全部要人工标注10万字的书，经过聚类，只需要对二三千字类进行标注即可，一类字只需标注一次。

聚类和人工标注，不仅完成了每一类文字的认字过程，还收获了更多新的训练样本，可以继续喂给机器学习。古籍里有很多生僻字、异体字、异形字，出现概率极低，几乎找不到样本。对此，达摩院团队使用字体迁移方法，让机器自动为每个字合成几个新样本，确保单字样本量达到10个，用来训练少样本识别模型。

从聚类到少样本模型识别，走完一轮，全书70%左右的文字可以被打上正确的标签，余下的部分将从头再来一遍，进行第二轮迭代，又能解决余下文字中的70%。经过两轮迭代，一本书91%的文字可以被识别。如此，通过不断的学习，训练数据越来越多，机器的认字能力也越来越强。

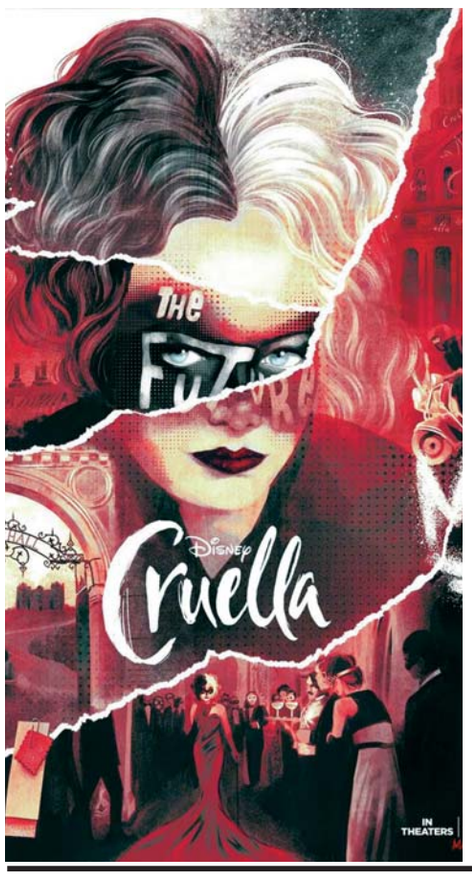
在复杂的算法养成过程中，人工标注的工作量被大大降低。“经过反复的学习和提升，目前达摩院系统对伯克利20万页古籍的整体识别准确率达到了97.5%。这套人机交互的识别方案，录入效率比纯人工输入提升了近30倍。”张建锋说。

张建锋表示，守护中华传世典籍，是科技工作者和文化工作者共同的使命。阿里计划将这套技术工具连同古籍数字化平台一并捐赠，交由权威公共机构长期运营；同时，阿里仍将在古籍数字化工作上持续投入人力、物力。



孔子博物馆藏《乾隆御定石经》初拓本

《黑白魔女库伊拉》：迪士尼的知黑守白



阿玮点映

□ 王文玨

“坏人”，如今越来越受银幕欢迎。坏人们似乎心比比干多一窍，又好像在情感空间里比别人多几个维度，破坏性强，力道十足，在黑夜里闪闪放光。尤其是那些坏不到底的坏人，黑白不定，内心徘徊里走出了人的彻底，也走出了魔的欲望。

《101忠狗》里，那对护狗小夫妻随着时光的流逝已面目扁平模糊。但那个披着狗皮大氅，叼着长烟枪的红唇库伊拉，反而成为坏人界经典。6月6日起在内地上映的新片《黑白魔女库伊拉》，主攻坏女人的青春时代，力图带给人们一个疯魔的天才，一个屡屡被亲情伤害背叛的普通女孩。

“坏人”成戏不简单。要坏，还要没那么坏，坏出来龙去脉，坏出气贯长虹的“我不得不”，在拿捏人类情感的阴阳两面之间反复横跳。迪士尼为库伊拉设计了黑暗心酸的童年：头发天生黑白两色的小孩走到哪里都被霸凌，母亲只好带她奔赴伦敦，希望一个开明城市能容纳“怪胎”。行至山穷水尽，母亲去昔日雇主、时尚大师男爵夫人那里求救，却被男爵夫人唤狗扑下了悬崖。小库伊拉以为母亲的死是自己的过错，从此沉默又疯狂。

伦敦也一样冷。孤儿库伊拉与小偷朋友一起长大，成了小偷女王。而天生的时尚DNA始终不甘，她雄心勃勃进入男爵夫人的时尚帝国打工，从小喽啰做起，如同安妮·海瑟薇之于《穿普拉达的女魔头》。当惊世才华已足以与男爵夫人抗衡，她也发现了母亲之死的真相。更加反转的是，眼前冷血恶毒到极点的时尚大师竟是自己的亲生母亲。一场关于亲情的复仇，就这样在“你黑，我更黑”中展开……

近几年，迪士尼摸到了市场的脉，屡屡到自己的“对面”去借助黑暗的力量讲故事。尤其在“大女主”类型片中，黑暗给人物内心冲突带来十足的澎湃动力。库伊拉被母亲抛弃，又被环境霸凌，还始终夹杂着自我否定，汇聚于复仇中攒起怼天怼地的能量。黑化的过程，反而成了洗白的过程——记忆中那个苍老贪婪的库伊拉不见了，观众眼中只有这个瘦弱躯干，黑瞳闪泪，燃烧着反骨烈焰的库伊拉。与迪士尼过去四平八稳的故事相比，人们在这种类型片里随角色求爽得爽，宣泄感十足。

整部影片的观感美艳惊人，光怪陆离。二十世纪七十年代，伦敦时尚界的朋克风刚刚兴起，不屑一顾的放肆与破坏，时刻准备手撕古典唯美派。母与女就这样夹杂着置对方于死地的欲望和审美的互相鄙视，轰轰烈烈地过招儿。全片的时尚大战化身美之对决——华丽朋克像妖王塞壬般，亮出来自叛逆未来的招牌。与男爵夫人的高贵清冷相比，黑白分明的库伊拉把贫民区带来的“脏乱差”升华为“酷烈”的力量，如野草生机勃发，疯狂扩张。在男爵夫人精心设计的晚宴上，库伊拉乘巨大的垃圾车呼啸而来，翻斗倾倒后，滚滚曳地的裙裾狂浪层叠，盛大到仿佛没有尽头。故事里原本单薄的起承转合，被美的杀气腾腾掩盖了几分。

尽管影片有“坏女人”“黑化”等头衔，迪士尼的理念在于，要想黑得漂亮，必须黑皮白心。再怎么作妖，主人公还是能守住迪士尼的老本行——真善美内核。即使它微妙地变换了形态，减少了分量，但那一点点对于朋友、亲人乃至理想世界的不舍，必须是人性中定盘的星。有了这一点，迪士尼才放心大胆地黑化自己的主人公，让她合理化复仇，合情化疯狂。但不管如何霹雳手段，最终仍要收起黑色羽翼，凭着对屈辱、不公的彻底洗雪，变成荣耀归来的胜利天使。

作为歌颂人类情感的老手，迪士尼一直知黑守白。它并非不能体察残酷的真正内核，却宁肯坚持真善美与暗黑元素之间的平衡。对它来说，放弃真善美，走冰冷黑的路子就是放弃自己的基本盘，风险不小。2009年的《鬼妈妈》是它最大胆的一次尝试，故事惊艳的不仅是暗黑系画面，更有对一个貌似合情合理的世界，现实之墙背后的无限怀疑。相比之下，欧洲一些新锐导演的作品“黑”得不那么硬，他们着眼的并非某个个体的黑色遭遇与治愈，而是把黑色溶解在童话故事的边边角角，不知不觉间，锤裂完美世界。法国动画《机械心》中，幽光明灭的浪漫永远伴随着“刀”，漫天白雪淹没终将破碎的爱情——爱，就会带来宿命的伤害，寒针般刺透了成年人的心。与这种“真黑真冷”派作品相比，迪士尼再黑也黑不出三里地。

新词解

“我emo了”

□ 本报记者 李梦馨

“‘我emo了’是什么意思？”“e（音同“一”）个人momo（即“默默”）地哭”？

这段不知夹杂着中英文还是文字拼音的对话，乍一看有些让人摸不着头脑。然而“我emo了”却是不少人当下状态的真实写照，可作如下解释——即“我抑郁了”“我颓废了”或“我情绪化了”。emo，全称“Emotional Hardcore”，即情绪硬核，最初是从硬核朋克中派生出来的一种有着艺术家气派的音乐。正如其名，emo是一种凸显情绪化表达的音乐，随着旋律递进，乐队在舞台上想笑就笑，想哭就哭，呐喊、嚎叫、抽泣，时常有之。后来emo逐渐演化为一种涵盖服饰发型、言行举止的风格。近来，随着“网抑云”的兴起，一种专指某种情绪的emo文化再次出现，emo通常带着颓废、抑郁和伤感的底色。

近义词：“网抑云”，以评论区闻名的网易云APP中，充斥着大量伤感的言语，最具代表性的莫过于“生而为人，我很抱歉”。不论是真情流露，还是无病呻吟，总能勾起同病相怜的人一起交流“病情”。“丧文化”，是指以“葛优躺”等为代表的一种青年亚文化。

幸福的人都是相似的，不幸的人各有各的emo。基金人看着满屏绿光，陷入emo；对于“学生党”来说，看不到尽头的期末季、毕业之际的分离和迷茫是他们的emo。工作的烦恼、爱情的不顺，不管处境如何，emo总有由头。一天之中，深夜总是人们集体沦陷emo的时候，但等太阳升起后，一切emo又都烟消云散。

不管什么时代，人们总需要一个释放情绪的出口。诗歌、校园民谣、青春疼痛文学……以不同形式背负了不同时代的痛点，不同阶段人的迷惘苦闷。在互联网时代，emo附着在音乐的种子上破土而出，在网络社区的你来我往间发酵膨胀。像这个时代的一切一样，情绪的表达也被精简至于近乎虚无。一句可以替代万语千言的emo被轻易地捡起来、灵活地运用，或许也会很快地被淡漠、被遗忘，让位于新的e或者m抑或o。

“破防了”

□ 本报记者 李梦馨

一句“我破防了”，道出百般复杂的情绪。

释义：原指游戏用语，即游戏中的装备、技能被打破，失去了防御效果。而在现实和网络语境中，破除的不是游戏防御，而是心理防御，是心理防线被击溃了。破防的缘由各有不同，也因此衍生出不同的意思。如果在对线中节节败退，跳脚急眼了，恼羞成怒，破防是一种嘲讽；如果是因过度悲伤而动容，破防又是一种崩溃；如果备受触动，破防又是一种感动。

近义词：“无能狂怒”，来源于游戏GTA5中的一款电视节目《无能狂怒人》，讽刺那些无能却情绪愤怒的人。“被刀”，即被虐到、被伤害的意思，完整的说法是“人被刀，就会死”，追的剧，看的小说，碰上虐心桥段，或者直接“bad ending”，对于观众、读者而言，就是“被刀”了。

当代人，似乎越来越容易遭遇让自己破防的事情。宏大的公共事件，琐碎细微的日常，屏幕里的温暖一幕，都是一幕幕破防瞬间。为什么当代人这么容易破防？是心理防线低了，脆弱得无力承受现实吗？或许只是从无数小事里，在他人的故事中照见了自己。

“藏狐表情包”

□ 本报记者 李梦馨

你永远不知道，下一个会火的表情包是什么。最近，有这样一张表情包刷屏了。图片中的男子有着一张神似藏狐的脸，恍若灵魂出窍一般，表情几度变幻，“狐疑中带着一丝茫然，崩溃中又略有几分挣扎”。

像大多数走红的表情包一样，越能表达言不尽意的情绪，用起来就越称手。苦熬到半夜付尾款、看着钱包被掏空的时候，人在屋外、钥匙被锁在房间里的时候……现实中总有够荒诞、够无奈的一刻，语言苍白无力的时候，抛去这样一张表情包便足以尽其意。

这张表情包的原主是B站up主@无穷小亮的科普日常，作为一名科普类博主，他的日常就是帮助网友鉴别各类稀奇古怪、来路不明的生物。但画风常常变歪，网友会时不时地发来一些恶搞视频，比如把插在树枝上的葡萄当成变异品种，将一片长成猴子形状的海藻看成水猴子等等。

在一期“网络热传生物鉴定”系列视频中，有网友发来几段荒岛求生的视频，自称在荒岛上求生的人在海滩上捡到了半人高的石斑鱼、龙虾、冬瓜，甚至还有活生生的女人……仿佛把观众的智商按在地上摩擦。于是乎，小亮皱眉沉思、捂脸崩溃、倚墙绝望，一张“神图”由此诞生。

但搞笑的背后，是科普真理、澄清误解的长路漫漫。在谣言丛生、伪科学充斥的环境中，荒诞有时候成了真理；当下的明白，在芜杂的信息轰炸过后，可能又会变成下一秒的糊涂。小亮曾在接受采访时说，科普其实就是和谣言抢夺阵地，造谣的人不停歇，他也不能放弃。

科普之路，可能是徒劳无功的，可能是一厢情愿的，但并不意味着是没有价值的。“一个树懒，你不告诉他这是树懒，他就认为这是‘水猴子’。但你一告诉他，窗户纸一捅破，他立马觉得这件事非常荒诞。我就要用一次又一次的努力，去对抗舆论的这种荒诞。”小亮说。

大众报业集团(大众日报社)出版 国内统一刊号:CN37-0001 邮发代号:23-1 社址:济南市泺源大街6号 邮编:250014 电话查询:(0531)82968989 报价全月45.00元 零售价:3.00元 广告许可证:鲁工商广字第01001号 广告部电话:85196701/6708 昨日开机5:20 印完8:10 大众华泰印务公司(大众日报印刷厂)印刷