数字化条件下古籍整理的基本问题（论纲）

史睿

（国家图书馆善本特藏部敦煌吐鲁番资料中心）

近年来各地的数字化图书馆建设日益兴盛，古籍的数字化也有一日千里之势，但是必须指出，在相关基础理论问题尚未解决之前，任何古籍数字化，乃至一切文献数字化的努力都可能陷入南辕北辙的尴尬局面。这些基本理论问题是，古籍数字化的基本性质是什么？在古籍数字化的过程中谁是主导，内容专家还是技术专家？古籍数字化应该以什么为依归，衡量数字古籍优劣的标准是什么？保障古籍数字化走向正确路径的基本要素有哪些？古籍数字化与其它一切文献数字化的关系如何？

首先要明确的是，古籍数字化属于古籍整理和学术研究（或称校雠学）的范畴，而不仅仅是图书载体的转换或商业炒作的噱头。故必须以相关领域的学者（即内容专家，而非技术专家）为主导，才可能向正确的方向发展。纯粹的技术专家不可能将古籍数字化，甚至数字化图书馆领向一条康庄大道。技术是形式，内容是核心，内容决定采取何种形式，形式只能服务于内容，而不是相反。只有熟悉对象（古籍）内涵的主体，即内容专家，才有能力决定实现古籍数字化的基本路向和基本框架，技术专家的作用是在既定的框架内如何最便捷、最优化地实现目标。故在古籍数字化领域中，内容专家和技术专家的关系应该如同建筑工程师与建筑工人，这样才能形成人力资源的最佳配置，有效地发挥各自的功能。IT 技术永远是工具，没有内容专家的构建和引领，再好的 IT 技术也难以带来真正的利益。

其次，既然古籍数字化属于古籍整理和学术研究的范畴，那么就必须遵循古籍整理的基本原则，懂得学术研究的基本思维过程。古籍整理古称校雠学，涉及目录、版本、标点、校勘等一系列的学问，承担着“辨章学术、考镜源流”的学术任务。整理一部古籍，要选择善本为底本，又要广校众本，之后精心标点，与所引之书和引用此书之书一一校勘，还须广徵群籍，拾遗补阙，最后提要勾玄，界定其学术地位与价值。实际上经过整理的古籍乃是一部融入学术研究成果的作品，还附有各种索引数据库，以便检索，而并非原有任何版本古籍的复制。数字古籍应该遵循古籍整理的一般过程和一般规则，而现有任何版本的数字古籍都没有达到古籍整理的最低标准。其次，关于人文学术研究的一般过程与计算机信息处理过程的关系，笔者曾发表《试论中国古籍数字化与人文学术研究》（《北京图书馆馆刊》，1998 年 2 期，28-35 页），可供参考。学术研究处理文献的方式往往是突破其原有结构，将原文献划分为若干基本单位，提取其中指向内部含义的关键词，依照它们的属性进行排序、筛选、统计和分类，比较相关文献中的关键词，寻求他们之间的相关性。这一过程在手工查阅纸本文献的时代，需要学者具有深湛的功力；近代以来多以编纂各类古籍索引方式，将经验转化为

知识。这正是电子媒体需要继承的重要方法，为此我们必须将隐藏于学者大脑中的经验和智慧加以总结，建立模型和序列，将无法比较的学术关键字赋以数值，例如编制具有规范控制年号与公元纪年对照表、历代官阶序例表、家族世系表、姻亲关系表、地名沿革表、人名字号表等，然后再以这些模型和序例为准标引古籍文本，使之完成经验到知识的转化，建立人文学术研究的科学内核，有效积累和传播人类知识，让每次学术研究行为都从前人的终点开始。如果数字古籍其关键词的标引和规范控制水平比不上传统索引，其存在的价值势必受到强烈置疑。

第三，也是非常重要的一点，古籍数字化，乃至一切文献数字化，必须采取以应用为指针的原则，一切工作都以这项原则为起点，同时又以它为评价工作成效的指标，要实现以应用为指针的原则，就必须懂得应用者的基本诉求。为此，我们首先需要确定数字古籍应用者的范围，古籍是为学术研究服务的，而非供大众消遣的余兴节目，数字古籍也不例外。古籍数字化必须全面借鉴以往的学术成果，明了纸本形态古籍在学术研究应用中的长处与局限。学者对于应用的要求是古籍数字化的起点，任何从事这项工作的机构或个人如果不了解这些要求，都必将导致全部工作的失败。以往，当计算机工程师开始设计会计软件时，对于会计的原理和应用要求也是完全陌生的，但现在会计软件工程师已经成为一个专门的行业，会计软件也与应用日益吻合。既然会计软件能与应用合拍，那么文献数字化也应将应用的要求放在第一位来考虑。实际上就其本质而言，学术研究的应用要求与其它领域并无二致，一言以蔽之，曰：“知识发现”。所谓知识发现（Knowledge Discovery in Database,简称KDD），与我们常说的数据资源再生相近，又称数据挖掘技术，是指从大量数据中提取出可信的、新颖的、有效的并易于理解的知识的高级处理过程[1]。它已广泛应用于市场行销、产品制造、通信网络管理、金融投资、自然科学研究等许多领域[2]。我们相信，数据挖掘技术运用于人文研究领域，必将创造出更卓越的业绩。纸本索引的目标就是数据资源再生，但问题在于纸本检索工具不能随读者的要求提供多种排检方式，故其再生资源的可用性有限；此外，研究者对文献本身的认识是随着研究工作的深入而逐步清晰起来的，其工作初期往往难以明确提出与自己的研究题目完全切合的全部关键词，而是要在较大范围内进行模糊查询或渐进式查询，这更是纸本检索工具所不能解决的。数字古籍目标应以纸本索引为向导，以应用为目标，将“知识发现”进行到底。应用是我们衡量古籍数字化工作的指标，

第四，为了实现知识发现，古籍数字化，乃至一切文献数字化必须建立在深入标引和严

---

[1] 见高文《KDD：数据库中的知识发现》，载《计算机世界》1998 年 37 期，8 月 28 日，技术专题版，D1 页。又朱廷劭《数据挖掘——极具发展前景的新领域》，载《计算机世界》1999 年 1 期，1 月 4 日，产品与技术版，C14 页。

[2] 见朱廷劭、王军《数据挖掘应用》，载《计算机世界》1998 年 37 期，9 月 28 日，技术专题版，D5, 8 页。

格规范控制的基础上。无标引、无规范控制的文本，其价值为零。因为只有经过深入标引和严格规范控制的数据库才能产生再生资源，而再生资源经过有效的排序和筛选，才能实现知识发现。当然，这必须以既往的学术研究为基础，以现代 IT 技术为工具。关于标引和规范控制，原本是现代图书馆学的题中应有之义，但近来 IT 技术的神话冲淡了相关的学术研究，现在我们才发现，深入标引和严格的规范控制是实现知识发现的必要手段。所以，我们必须破除 IT 技术的迷信，重新估价 IT 技术的功能与价值，并努力补上传统学术中标引和规范控制这一课。和其他文献相比，古籍的标引和规范控制更为复杂，可以认为，古籍的数字化是一切文献数字化的特例，如果我们对于解决这个复杂特例有了完整的方案，那么其他文献数字化解决方案就迎刃而解了。

胡适之先生认为传统的经史研究存在范围太狭窄，注重功力而忽略理解，缺乏参考比较的材料等积弊，故以清代三百年间第一流人才的心思精力，都用在经学的范围内，却只取得了一点点的成果，关键是缺少对古籍的系统整理，又不注重学术成果的积累，两千四百多卷的《清经解》，大多是一堆流水烂帐，没有条理，没有系统，人人从“粤若稽古”、“关关雎鸠”说起，怪不得学者看了要望洋兴叹了[3]。针对清儒治学方法的缺陷，胡适之先生着重提出，必须系统地整理古籍，包括索引式、结帐式和专史式的整理。此后，学界编纂了多种引得、通检、索引、汇编等工具书，部分完成了索引式整理的目标，拜前辈学者之赐，我们查阅古籍不知享受了多少便利。但是我们也发现，中国古籍汗牛充栋，经过系统整理的毕竟只是少数，方便的检索工具也还嫌太少，离胡适之先生的标准还有相当的距离。即使是已有索引的古籍，我们用来解决具体问题时仍会感觉到种种不便。至于结帐式的整理，则尚未受到学术界的普遍重视，而在未有结帐式整理之前，所作的专史研究，其完整性、可靠性都值得怀疑。为了促进人类知识的有效积累和有效传播，使我们的后代不必研究任何问题都从“粤若稽古”、“关关雎鸠”说起，我们才有必要从事文献数字化的工作；此外，积极建设网上中文资源库，打破某些国家或某种语言对网络资源的垄断，这将有利于中外学术文化的交流，树立中国人的学术自信心和自尊心。总之，我们一切都应从长远目标出发，而不应被暂时的商业利益所蒙蔽。

[3] 胡适《〈国学季刊〉发刊宣言》，原载《国学季刊》一卷一号，1923 年 1 月，此据欧阳哲生编《胡适文集》三，5—17 页，北京大学出版社，1998 年 12 月。