中国国家图书馆国际敦煌项目的创立与前景

高奕睿　林世田
（英国国家图书馆　中国国家图书馆）

一、敦煌遗书的发现及流散

敦煌莫高窟千佛洞始建于公元四世纪的东晋时期。1900 年 6 月 22 日，道士王圆箓无意间发现了藏经洞，其中堆满了 4 至 11 世纪的中国古代遗书及其它文物。藏经洞发现后，英国人斯坦因、法国人伯希和、日本大谷光瑞探险队、俄国奥登堡探险队、美国人华尔纳等闻讯后纷至踏来，采取种种手段，将大批敦煌遗书、绢画、壁画及雕塑掳掠运往国外。敦煌遗书的流失零落令人扼腕痛惜，成为中华民族与中国学术界之一段伤心史。由于敦煌遗书分藏于中、英、法、俄、日等地，至今缺乏一个完整的联合目录，藏经洞内到底藏有多少遗书，至今仍无确切统计。现在中国国家图书馆藏 16000 余件；英国图书馆藏 13000 余件；法国国家图书馆藏 5700 余件；俄罗斯科学院东方学研究所圣彼得堡分所藏 10000 余件。另外敦煌研究院、中国历史博物馆、故宫博物院、甘肃省博物馆、敦煌市博物馆、北京大学图书馆、上海图书馆等单位都有收藏，总数在 5 万件以上。这五万件汉文文献大致分为佛教类文献与非佛教类文献两大类，其中尤以佛教文献数量为最多，占总数的近 95％左右。

敦煌藏经洞发现以来，社会上乃至学术界一直流传着一种误解：认为敦煌遗书之精华部分已然被外国“探险家”们搜刮殆尽，中国国家图书馆所藏者均是价值不大之弃余糟粕，这种看法大谬不然。三十年代，国学大师陈寅恪先生在《敦煌劫余录序》中曾经列举大量事例，着重批驳了这种不实说法，如西凉建初十二年（417 年）写本《律藏初分》，是馆藏敦煌遗书确切纪年最早的一件，历经近 1600 余年仍完好如初；舞谱、摩尼教经典等为国内硕果仅存之资料，其珍贵稀有不言自明；《辩亡论》、《姓氏录》等皆为敦煌遗书中之稀世精品；解放后数十年来，通过各种途径不断有散落的敦煌遗书回归于中国国家图书馆，其中不乏大量珍贵文献，如《尚书》、《毛诗》、《春秋》、《老子》、《庄子》、《文选》等抄本。今天中国国家图书馆所藏敦煌遗书不但在数量上占据世界第一位，而且在内容质量方面与世界上任何一所敦煌遗书收藏机构相较毫不逊色，有过之而无不及。

国家图书馆在敦煌遗书入藏之初即派专人负责，整理编目。约于 1912 年编撰完成《敦煌石室经卷总目》，著录 8679 号敦煌文献。但其编排方式为流水草目，不便学者使用。1922 年陈垣先生在《敦煌石室经卷总目》基础上主持编撰分类编目《敦煌劫余录》，1931 年 3 月作为中央研究院历史语言研究所专刊第四种出版，著录 8653 号。1929 年成立的写经组，负责编撰馆藏敦煌文献目录，至 1935 年完成具有较高学术价值的《敦煌石室写经详目》及《续编》。1981 年 7 月，善本组将新字号部分整理编目，完成《敦煌劫余录续编》，著录 1065 号。1990 年，在馆长任继愈先生亲自主持下，《中国国家图书馆藏敦煌遗书总目录》的编纂工作正式启动，现在目录初稿已经完成，正在定稿。总之，敦煌文献作为国家图书馆的专藏之一，

历来倍受重视，在妥为珍藏的基础上，进行了长期的整理、修复与编目，并对研究者开放阅览。这些对推动敦煌学的发展起到了积极而重要的作用。

二、国际敦煌项目的创立

敦煌在中国，敦煌学在世界。敦煌莫高窟的艺术品及藏经洞所出古代文献是全人类珍贵的文化遗产。近百年来，收藏敦煌文献文物的各国图书馆、博物馆和研究机构，为保护这一遗产做出了不懈的努力，各国专家学者也为敦煌学研究的发展做了积极的贡献。尤其是近二三十年来，随着交流的不断增进、合作的日益加强与信息的逐渐畅通，人们获得敦煌资料的条件大大地改善了，世界敦煌学研究呈现出欣欣向荣的局面。然而不容忽视的是随着国内外敦煌写卷的相继公布，以及敦煌学研究的深入发展，学者们希望充分利用敦煌文献的要求越来越强烈，这就突出地显现出敦煌学研究中存在的三个问题。首先，收集整理资料是研究工作的基础和起点，敦煌研究也不例外。由于敦煌文献分藏世界各地，一般学者很难全部看到，这使得研究工作受到很大的局限，难免不深入，或者留下遗憾。第二，研究所借助的缩微胶卷，限于拍摄时的技术条件，许多写卷影像不清。而且，由于没有敦煌学专家的指导，拍摄的胶片上漏掉了许多重要信息。而根据胶卷印成的《敦煌宝藏》所含的信息内容还不及胶卷，无法满足学者研究的需要。第三，这些记录人类文明的写卷是世界文化的遗产，它们需要永久的保存与保护，为避免原件受损，应尽量减少流通，这就造成了研究与保存保护之间的矛盾。为了解决这些问题，更好地为学界服务，国家图书馆开始考虑如何使用更先进的手段保存保护敦煌文献，并且不影响为读者提供服务。当然据我们了解，这三个问题也普遍存在于敦煌文献的收藏单位。

随着时代发展，海内外学术界越来越认识到，将散落的敦煌文献珠联璧合，将是世界学术史上的一件盛事，也是人类文明史上的一件大事。因此，敦煌学国际合作呼之而出。1993年，中国国家图书馆、大英图书馆、新德里国立博物馆、法国国家图书馆、圣彼得堡东方研究院、柏林国家图书馆倡议成立国际敦煌项目。1994 年在大英图书馆设立秘书处，每年出版三期通讯，每两年举办一次会议。1994 年开始在专门设计的国际敦煌项目数据库中录入文献资料，1997 年开始文献图像数据化。1998 年 10 月数据库上网，网址：http://idp.bl.uk，目前上网图片已达上数十万幅。

三、中国国家图书馆国际敦煌项目的创立与发展

自 1997 年，国家图书馆即开始与英国国家图书馆磋商合作开展国际敦煌学项目（简称IDP）。经过反复商讨，最后经文化部审批，2001 年 3 月 7 日中英双方签订了合作谅解备忘录，开始了中国国家图书馆与英国国家图书馆为期五年的 IDP 项目合作。这个项目的目的就是通过数字化、网络化技术将敦煌文献的编目数据和手稿图像按统一的标准和格式整合成数据库，放在互联网上，无偿地供研究人员使用。值得注意的是，对于合作数据库的版权，

备忘录做了明确规定：各馆所做的图像和数据的版权归制作者所有；任何一方不得修改和删除对方数据；中英双方和第三方可以存取图像，但不得复制，也不得用于其它目的等。

为使学者们可以看到与原卷一样逼真的图像，项目使用了专门设计的 4D 数据库，用精密的 PHASE1 数码扫描设备将敦煌写卷制成一幅幅高清晰度的图像。图像将展示写卷的全部内容——正面、背面，甚至没有文字的地方，它比实际尺寸要大，图像的清晰度与看原卷没有区别。学者可以从任何地点、在任何时间通过网络检索到高质量的彩色图像。图像放大之后，还可以观察到过去用普通放大镜不易观察到的字的细部、墨的层次、纸张的纤维等问题。学者查阅敦煌文献既不必再有舟车劳顿之苦，也无需接触珍贵又容易损坏的原卷，解决了保护与研究的矛盾。

经过中英两馆的共同努力，2002 年 11 月 11 日国际敦煌项目中文网站在国家图书馆正式开通，网址：http://idp.nlc.gov.cn，目前上网书目数据 10773 条，写卷 500 余件，图片 6000 多拍，学者档案信息 400 余条，编辑 4 期敦煌学通讯。近年我们还制作了敦煌文献研究索引近 4 万条，暂时没有上网。今后几年内，将中国国家图书馆馆藏所有敦煌文献将陆续数字化后上网提供读者阅览。

中国国家图书馆国际敦煌学项目的最终目标是将中国国家图书馆所藏的写卷全部数字化，放在网络上让全世界的学者自由读取，以促进学术研究的发展。合作以来，中英双方本着求同存异的原则，发挥各自所长，相互支持，相互配合，取得了令人瞩目的成果，引起了国际上的广泛关注。很多外国专家专门向我们了解这个项目，国际会议也特别邀请该项目人员演讲、演示。

敦煌文献是人类珍贵的文化遗产，如何保护好这份遗产是学界关注的热点，也是图书馆员艰巨的责任。用先进的技术无损伤地揭示人类文明史中最古老的纸本文献，这项工作是极具挑战性和创造性的。我们已经解决了许多遇到的问题，我们还将继续遇到问题，解决问题。

四、中国国家图书馆敦煌数字化计划

中国国家图书馆与大英图书馆合作的国际敦煌学项目为中国国家图书馆敦煌文献数据化提供了实践和经验，经过几年的积极探索，我们根据中国国家图书馆的自身特点，制定了中国国家图书馆敦煌文献数字化的总体发展思路，即：以国际敦煌学项目为契机，建立具有国际水平的敦煌吐鲁番学研究数据信息中心，提高国家图书馆在学术界的地位；以国家图书馆的基础业务工作支持国际敦煌学项目的发展，使两者成为一个有机整体。在这个指导思想下，在我馆所做的 IDP 数据库基础上，我们设计了中国国家图书馆的敦煌数据库结构，包括如下内容：

- **中国国内散藏敦煌文献联合目录**

由于众所周知的原因，敦煌文献分藏世界各地，人为地造成整理研究的困难。编辑一部敦煌文献总目是中国老一辈敦煌学家挥之不去的梦想。王重民先生在《敦煌文献总目索引》后记中就提到应编辑一部“新的、统一的、分类的、有详细说明的敦煌文献总目”。当时世界各国馆藏尚未全部公布，要编成一部“总目”的时机还不成熟。但是，70 年代以后中国国内散藏目录相继公布，特别是 90 年代以后俄藏、北京大学、天津艺术博物馆、甘肃等地

所藏敦煌文献相继出版[1]，编辑国内散藏敦煌文献联合目录时机业已成熟。我馆于 2001 年 10 月份开始了这一工作，现在已经完成。

- **研究论著目录数据**

敦煌学作为国际显学之一，研究论着不断增加，研究者往往有望洋兴叹之感，编制目录便成为研究者殷切的期求。早在 1994 年敦煌吐鲁番资料中心成立之初，中心就把收集资料、编辑目录作为首要任务。根据学界的需要该中心正在编辑两个专题书目数据库：

敦煌文献研究论着目录数据库（含中、英、法、俄及其它馆藏）：资料中心于 2001 年出版了《国家图书馆藏敦煌文献研究论着目录索引》，收录了 1910—2001 年国内外发表的有关国家图书馆藏敦煌文献研究论着 8576 条，已经制成电子文本，即将在中国国际敦煌学网站上公布。从今年始，中心日夜加紧编辑英、法、俄及其它馆藏敦煌文献研究目录，至今已

---

[1] 七十年代以后出版的敦煌文献目录及图录如下：

1、敦煌文物研究所藏敦煌文献目录/敦煌文物研究所资料室编/文物资料丛刊第 1 期/1977；

2、关于甘肃省博物馆藏敦煌文献之浅考和目录/秦明智编/1983 年全国敦煌学术讨论会文集·文史·遗书编/1987；

3、西北师范学院历史系文物室藏敦煌经卷录/曹怀玉整理/西北师范学院学报（社科版）/1983 年第 4 期；

4、敦煌县博物馆藏敦煌文献目录/荣恩奇整理/敦煌吐鲁番文献研究论集第三辑/1986；

5、上海图书馆藏敦煌文献目录/吴织、胡群云编/敦煌研究/1986 年第 2—3 期；

6、天津市艺术博物馆藏敦煌文献目录/刘国展、李桂英编/敦煌研究/1987 年第 2 期；

7、北京大学图书馆藏敦煌文献书目/张玉范编/敦煌吐鲁番文献研究论集第五辑/1990；

8、重庆市博物馆所藏敦煌写经目录/杨铭编/敦煌研究/1996 年第 1 期。

9、上海博物馆藏敦煌吐鲁番文献（1-2）/上海古籍出版社、上海博物馆编/上海古籍出版社/1993。公布上海博物馆所藏 80 件敦煌文献。

10、北京大学藏敦煌文献（1-2）/北京大学图书馆、上海古籍出版社编/上海古籍出版社/1995。公布北京大学图书馆收藏敦煌文献 286 件。

11、天津艺术博物馆藏敦煌文献（1-7）/天津艺术博物馆、上海古籍出版社编/上海古籍出版社/1996——1998，公布天津艺术博物馆藏敦煌文献 350 件。

12、甘肃藏敦煌文献（1-6）/段文杰主编/甘肃人民出版社/1999，其中影印敦煌研究院、酒泉市博物馆、甘肃省图书馆、西北师范大学、永登县博物馆、甘肃中医学院、张掖市博物馆、甘肃省博物馆、敦煌市博物馆、定西县博物馆、高台县博物馆所藏敦煌文献共计 696 件。

13、上海图书馆藏敦煌吐鲁番文献（1-4）/上海图书馆、上海古籍出版社编/上海古籍出版社/1999。公布上海图书馆藏敦煌吐鲁番文献 187 件。

14、浙藏敦煌文献/《浙藏敦煌文献》编委会编/2000。公布浙江省博物馆、浙江图书馆、杭州市文物保护管理所、灵隐寺等单位藏品 201 件，附录温州博物馆所藏浙江出土五代以前写经 2 件。

经收集近 4 万余条。

敦煌吐鲁番学中文论着目录数据库：此部分内容已经完成上网，读者可以在中国国家图书馆网上查询。本库收录 1908 年至 2001 年中国大陆及港台地区出版的报刊、论文集中有关敦煌吐鲁番学的论文和专著目录，以公开发行的图书、报刊为主，兼收部分内部资料。为方便检索，发表在不同刊物上的同一文章一并收录。现有资料 2.5 万余条，以后每年补充新资料。

- **敦煌吐鲁番学学者档案数据库**

敦煌资料中心 1994 年开始建立学者档案工作，得到学界同仁的支持与襄助，为一百多位学者建立档案，专架存列，与所藏书刊资料相互补充，为学界服务。现在，中心将参考 IDP 数据库中的中国学者资料库，相互补充，建立更加完备的中国学者档案数据库。该数据库已录入学者档案信息 400 余条。

- **敦煌文字数据库**

敦煌写卷不仅仅包含了丰富的历史文化社会信息，也是中古时期汉字字形演变的实物载体。先秦文字处在汉字形成发展时期，字体多变，且与现代字体有很大的差异，所以近年来早期汉字字形研究逐渐兴起，以至于产生了一门新的学科——汉字古字体学。考古发现的文献材料绝大部分经过学者识读，整理出版。一批综合性的著作相继出版，如《楚系文字编》、《楚简帛文字编》、《包山文字编》、《郭店文字编》、《吴越文字汇编》、《战国文字编》，这些均是研究战国时期各国文字的综合之作。

与古代文字研究成果迭出相比，中古写卷文字的研究则相对非常冷清。之所以如此，主要是学者普遍认为中古时期文字与今天通行文字基本相同，差异甚少。所以尽管敦煌文献发现已经有一个多世纪，但是利用这批文献进行文字研究所做的工作还很少。目前仅有的成果有潘重规《敦煌俗字谱》、张涌泉《敦煌俗字研究》，应该说它们还不是很全面，比如潘著仅是利用台湾所藏的 144 件敦煌文献。

作为敦煌学数据库的一部分，敦煌文字数据库的目标是利用敦煌写卷建立一个互动的中文文字数据库，实现文字检索和文字与写卷的关联,建立一个互动的中文文字数据库。

目前大部分文字数据库仍采用纸本字典形式，提供的字形多使用描摹、复印剪贴等手段，有明显人工痕迹或模糊不清，读者很难确认其精确的字形，而且文字与原书相脱节。敦煌文字数据库克服了这些缺点，它既能反映某一个字字体的演变，又能把字和原来的卷子关联在一起，进行综合的研究。同时充分利用电脑处理文字的优势，完全使用精确扫描的图像制作的，原原本本的反映手写字体原貌，并且提供各种途径的检索。

研究敦煌字体另外一个重要意义是敦煌文献是中国现存的唯一中古时代写本，这项研究可为古代字体演变提供一个坐标。敦煌写卷拥有如此庞大的数量，将此制作成一个完整的数据库所提供的信息是其它文献所不能比拟的。敦煌文字数据库完全可以建立一个可靠的、有代表性的中国某一特定时期的文字数据库，而用其它文献构建的数据库其精确性、广泛性根本不能与之相提并论。研究敦煌文献已经发展成为一个专门的学问，但是没有充分注意敦煌字体的研究。敦煌文字数据库则旨在帮助学者填补这一空白，并促进相关研究。

敦煌文字数据库用户可以通过互联网使用，其基本功能如下：

1、数据显示：文字和写卷使用高清晰图像，用户无需再查阅其它文献；点击某一字可立即看到其所在的原卷；可以将某一字按照一定的顺序排列起来，以便进行比较。如按照时代的顺序排列可以发现字体演变规律；可以将同时期内佛经、道经以及世俗文书中的相同文字进行比较；也可以将传统字典中某一字与敦煌文献中的字相对比。敦煌文字数据库可以使学者根据不同的需要进行各种比较研究，这项工作过去学者主要靠手工操作来实现，费时费力，现在学者可以在几分钟内完成，因此它会成为一个非常有用的研究工具。

2、文字数据：传统文字数据库，例如字典、词典，其文字检索绝大部分采用标准字体。从敦煌文献上切割下来的敦煌文字其结构不同于现代文字结构，同一个字甚至有十几种写法，有的文字结构甚至没有一个现代的对等结构。另外即使文字的结构基本一致，其读音、字义仍是学者认真讨论的热点。因此，在敦煌文字数据库收入相关的标准字体，因此一个文字有时与若干个现代字相连接，而不是一个。此外，尚有一些字没有相对应的现代汉字，数据库仍然包括这部分。文字的图像信息包括写卷的日期、出土地、出土日期，收藏地以及研究目录。

3、数据库外部信息：数据库也将收集文字外部信息，标准字库包括三方面主要信息，标准字的构成，音标以及现代字典中的信息。

A.标准字的构成：这部分是文字的分解与检索。每一个字被分割成几个基本部分，允许研究者根据这些基本部分自由进行检索。一般来说一个字可以分成两部分，当然有的部分还可以继续分下去。比如，“體”可以分成“骨”和“豐”，“豐”还可以继续分成“豆”和“曲”。

B.拼音的价值是用来检索的，与其说是汉语拼音列表不如说是韵书。本课题最终成果将包括《广韵》等数据，但是目前我们只是利用现代字典的拼音检索。

C.数据库其它外部数据来自于像《干禄字书》、《五经文字》、《龙龛手鉴》等传统字书词典。这些字典包括 8 至 10 世纪使用的标准字和非标准字。在敦煌文字数据库中两者可以直接比较。

3、检索:用户可以依据以下原则检索特定的文字:一般文字的默认检索，可以检索出所要检索文字的图像；在传统偏旁部首检索的基础上，敦煌数字数据库加以扩大，可以根据字的任何一个部分来检索该字；因为敦煌文献中存在大量外来语，利用传统的拼音检索只能检索出其中一部分。敦煌文字数据库将设计一个更开放的拼音检索系统，尽量满足敦煌文字的需要；除了以上原则之外，用户还可以通过写卷的日期、存放地、文献的分类来限制检索结果。

敦煌文字数据库设计思路：

1、大容量：现存敦煌文献不下 5 万件，每件文献按 2 万字计算，总字数将远远超过 10 亿字，数据库的结构将必须处理这么多的文字。另外文件名命名的便利也必须考虑，因为每一个文字图像都需要一个唯一的名字和路径。这个数据库结构必须能处理这些内容。

2、可持续性：敦煌文献数据库是个长期的项目，要求不断地能增加新的数据。数据库将一直放在网上，并不需要大规模修改。

3、可延伸性：尽管敦煌文字数据库是专门为敦煌文字设计的，但它有非常强的可延伸

性。它还可以容纳金石铭刻、竹木简牍等，还可以加入西夏文、纳西文等少数民族文字。

4、检索功能的可灵活性：因为敦煌文字字体与现代字体相比多有变化，因此检索功能需要有一定的灵活性。

敦煌文字数据库内容：

1、输入文字外部信息：敦煌文字数据库首先需要输入正字及拼音。此外还需将《干禄字书》、《五经文字》、《龙龛手鉴》等三部中古时期的字书扫描、切分，并制作索引。

2、切字：过去像这样的文字数据库绝大部分的精力都花在切字上。而我们则在 Photoshop 软件基础上，设计了自动选字、切字、命名软件。使用这种方法，切字的时间大为缩短，经初步测定，一个生手每分钟可切 60 多个字。

3、文字与图像连接：这是另外一个比较费时的工作。敦煌文献的每一个字必须与至少一个标准字体连接。经常是一个对应若干个。假如没有文字库中没有与之相匹配的字符，就需要手工录入到文字库中。因为敦煌文献百分之九十以上是佛教文献，绝大部分佛经网上已经有电子版，我们挑选质量高的版本作为底本自动转入文字库中作为标准字库。

总之，敦煌文献作为中国国家图书馆最重要的特藏之一，从始至终都被格外看重。今天，我们将以 IDP 合作项目为基础，借助新的数字技术手段，结合管理人员的专业技能，完成一个包括书目、全文、影像、研究成果等综合信息的研究性数据库，为学术工作服务。这个数据库已初见端倪，今后还要不断完善。