DOI:10.16262/j.cnki.1000-8217.2012.06.014

·基金纵横·

PDF 文档解析在国家自然科学基金项目档案数字化验收过程中的应用

冯 奇 吴 宁

(国家自然科学基金委员会，北京 100085)

1 引言

随着我国科技体制发展和国家财政对科技支持力度增大，科学研究项目在数量上不断增长，科研活动中形成的文字、图表、计算、影像等文件资料也日益增加。这些文件资料除具有凭证和查考等档案一般性价值以外，还具有知识储备、促进科技成果转化等利用价值，是档案的重要组成部分。

信息技术的发展为档案的数字化管理和利用提供了有利条件。借助现代信息技术手段，实现科技档案资源的存储数字化、管理信息化和利用网络化，是优化科技档案查询和利用效率、使科技档案充分为社会服务的有效解决方案，也是档案工作的发展趋势。

国家自然科学基金项目(以下简称科学基金项目)档案是受国家自然科学基金项目资助的、在自然科学研究活动中形成的文件资料，这些文件覆盖科学基金项目的组织申请、中期管理和结题验收 3 个阶段，具有很强的专业性和现实使用性。国家自然科学基金委员会(以下简称自然科学基金委)作为项目管理部门，十分重视档案数字化工作，不断推进档案信息资源库和档案信息化网络的建设。科学基金项目档案数字化存储、信息化管理和网络化利用将提高科学基金项目档案利用效率，并在一定程度上缓解科学基金项目申请量大幅增长引起的管理工作量激增的压力。

2 科学基金项目档案数字化工程

科学基金项目档案具有较强的系统性，即每一项目案卷均为围绕一个独立科研项目展开并随项目逐步进行有规律形成的一系列文件的整体。科学基金项目档案的案卷一般包括项目申请阶段的项目申请书、评审与评议意见、项目批准通知书、项目计划书，项目管理过程中的进展报告、中期检查报告、变更申请，项目结题验收过程中的结题报告、结题验收意见等材料。项目档案的系统性使档案整理和数字化过程可以暂时摆脱档案内容的专业性，为整理立卷和数字化处理提供了方便。

科学基金项目档案数字化工程主要任务包括档案纸质件的整理移交、数字化和复卷验收 3 个方面的工作。

2.1 项目文件的整理和移交

项目文件的整理和移交是项目档案数字化加工的前期准备工作。科学基金项目文件在归档之前由各学科项目管理人员管理并暂存积累。项目文件移交之前，须经过整理组卷、编写页号、编制档案卷内目录和填写移交清单等步骤。项目文件在项目结题后集中移交至自然科学基金委办公室文档处。

2.2 数字化工作流程

项目档案信息系统和网络利用平台是数字化项目档案的载体，其运行数据主要为项目档案数据库和数字化后的电子档案。

(1) 项目档案著录。档案著录是在档案信息管理系统中著录档案信息，著录内容主要由项目信息和项目档案案卷信息组成。项目信息包括项目批准号、项目名称、项目负责人及单位、学科编号、执行时间、中英文摘要、主题词等项目基本信息，课题组成员、单位及证件号等课题成员信息和成果名称、说明、主要完成者等成果目录信息，可以由自然科学基金委项目管理系统通过数据接口导入档案库；案卷

本文于 2012 年 8 月 9 日收到.

信息包括目录号、案卷号、盒号、页数、保管期限、立卷人及备注等。

(2) 项目档案扫描加工。形成电子档案是数字化过程中一个基础性环节。如果纸质文件没有电子副本，获得电子档案的方法是将纸质项目档案通过扫描转化为数字形式的原文影像文件。在数字化办公条件下，与纸质件一致的电子副本也可作为电子档案的直接来源。

在项目档案数字化过程中，如果把项目档案的每一页都进行扫描将是一件耗时费力的工作。我们考虑到这样一个事实：自2003年开始，科学基金项目管理工作均在项目管理信息系统中进行，项目基本信息及相关报送材料的电子文件(PDF格式)也同时上传到项目管理系统中，但是与纸质件不同之处在于这些电子文件不包括签字和单位公章。我们注意到PDF电子文件是由多个页面组成并且是页独立的，可以单独处理每个页面，因此为了充分利用现有资源，节约数字化成本和降低项目档案的扫描工作量，数字化过程仅对没有电子原文的纸质页面进行扫描。

为保证扫描图像的质量和有效利用，将扫描文件替换至原电子文件前，还应对扫描图像进行纠偏、去污等图像处理和OCR(Optical Character Recognition)识别工作。OCR识别可在图像文件中识别文本信息，是对档案扫描文件进行全文检索的基础。自然科学基金委项目档案电子文件本身就是一种具有双层结构的PDF文件，文件内容包含位置对应良好的文本层和图像层，因此仅需要对扫描页面进行OCR识别并转换成双层PDF格式即可。

最后将扫描页面替换到原PDF文件的相应位置，产生完整的项目档案电子文件。

以项目结题报告为例，结题报告纸质件和电子文件的内容、版式一致，最后3页分别为负责人签字及审核意见表、经费决算表和经费使用说明表。纸质件与电子文件的区别为这3页中有无课题、项目管理和财务部门负责人签字和单位公章。仅需对这部分内容进行扫描处理后并替换到PDF文件中即可形成与纸质件一致的电子文件。采取这一方式，一份项目档案的扫描工作量可降低到10页左右。著录内容及扫描文件质检合格后，即可通过档案管理系统对案卷和电子文件进行挂接和关联。

2.3　项目档案纸质件复卷装订

项目档案纸质件装订是将项目文件整理排列并用档案卷皮封装的过程。装订的纸质档案需经过验收后才可以存入档案柜。

2.4　验收

科学基金项目档案数字化验收工作主要包括档案著录信息质检、扫描图像质检和复卷装订质检3个方面，是整个数字化过程中的重要一环。著录信息包括项目信息和案卷信息等内容。由于案卷信息是批量生成的，因此验收时主要对项目信息著录准确性进行质检，同时校验数据格式和枚举型数据的著录规范性。验收工作还需要检查扫描图像是否清晰、PDF页面替换是否正确等。最后对纸质件装订、装盒情况进行人工检查，保证整个数字化工程质量。

3　人工验收系统的设计

验收时，通常采用人工校验方式对比纸质档案与档案系统各著录项、检查电子档案扫描页面图像质量等来完成著录信息和扫描图像的验收工作。2012年第一季度，我们改进了项目档案人工验收系统，使验收系统尽可能方便人工校验。这一版本的验收系统采用左右两栏版式设计，界面左侧展示数字化著录项，右侧显示PDF电子原文。为方便检查著录项，在左侧设置封面、摘要、课题组成员、成果信息和扫描页面等选项卡，每点击一个选项卡时直接定位到PDF相应页面，左右两栏同时对照显示相应内容，数字化著录项的展示风格与PDF文件排版格式保持一致，例如封面卡片中按照项目结题报告的页面格式显示了数字化著录的项目基本信息，包括项目批准号、申请代码、资助类别、亚类说明、附注说明、项目中英文名称、负责人、依托单位、资助金额和执行年限等关键信息，其显示位置与PDF文件相同(图1)，使得验收人员容易检验著录内容的准确性及规范性。

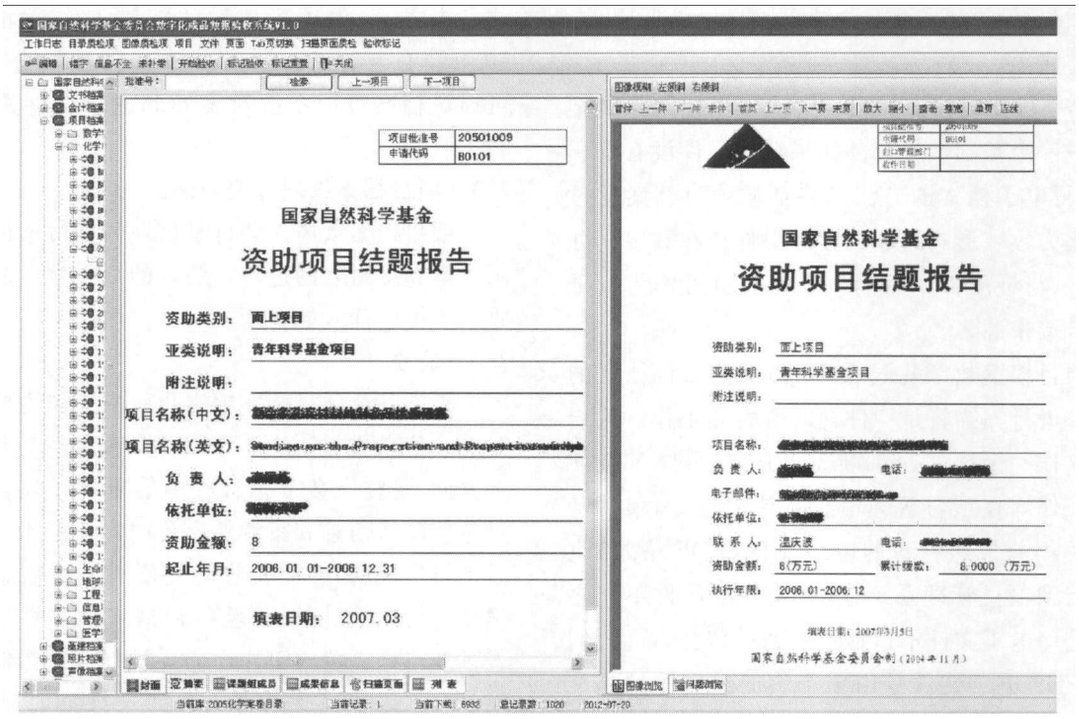

图 1 人工验收系统展示

自然科学基金委科学部验收人员利用该验收工具进行项目档案验收，需由档案管理人员对质检人员及其验收对象授权后方可登陆系统进行验收。验收过程中可以方便地标记著录项和图像的常见错误，也可以根据实际需要录入非典型的错误信息或直接修改。系统会自动记录验收过程并反馈给数字化工作人员以便及时改正项目档案数字化过程中的错误。

尽管人工验收系统软件功能较为完善且使用较为方便，但由于项目档案著录内容覆盖项目文件的诸多方面，课题组成员信息和成果信息等著录数量繁多，人工验收复杂度大大增加且人工验收效率大幅降低。各学科工作人员项目管理任务繁重的现状，也增加了开展人工验收的难度。

如何减少各学科工作人员的验收工作量、便捷地完成档案著录项和扫描图像质量的验收工作是本次项目档案数字化工程考虑的焦点。

4 PDF 文档解析在数字化档案验收过程中的应用

4.1 基于 PDF 文档解析的自动验收系统

PDF 文件是一种定义了版式位置属性的电子文档，文档内容同时依赖于版式位置信息。PDF 版式信息有利于提取文件的内容，使得通过既有 PDF 电子文件校验项目档案著录信息和扫描图像替换正确性成为可能。

科学基金项目档案自动验收系统基于 PDF 电子文件解析来实现档案著录信息的准确性校验。项目基本信息均来源于项目档案的各个文件，而项目档案电子文件与纸质件一致，因此可通过对比项目基本信息著录项和电子文件相应内容的一致性来检验档案数字化的著录质量。例如，科学基金项目基本信息中的项目名称、资助类别、负责人、依托单位、资助金额和执行年限等均与结题报告 PDF 文件首页相应位置提取的字符串进行匹配；项目中英文摘要、关键词、研究成果目录信息则与结题报告相应页面的内容进行对比；项目课题组成员信息则利用项目申请书 PDF 文件的课题组主要成员部分来判断项目基本信息著录的准确性。

自动验收系统还通过解析 PDF 文件实现了判断扫描页面替换原有 PDF 页面正确性的校验功能。科学基金项目管理过程中所有文件均为根据标准模板撰写的，具有统一且相对固定的格式，因此可以通过检验 PDF 文件的关键页面位置是否与模板一致来判断扫描页面替换及替换页序是否正确。例如数字化后项目档案的结题报告 PDF 文件最后 3 页依次为带有签名和公章的项目负责人签字及审核意见表、经费决算表及经费使用说明表的图像扫描文件，自动验收工具能够对替换页面及替换页序的正确性做出判断。由于其他 PDF 文件只有 1 页签字盖章页，只需要检验替换位置是否正确即可。

此外，自动验收工具仍保留人工验收系统的规

范性校验功能。规范性校验主要包括检验著录格式是否符合数据格式设置、枚举型著录项是否合法及项目执行年限等著录项是否符合逻辑 3 个方面，保证著录内容在形式上的正确性。自动验收系统可以完成检验项目档案著录及扫描文件替换过程正确性并记录质检结果的验收工作，基本可以代替人工校验工作内容。但是，遇到 PDF 文件跨页表格中逻辑上应在某页面表格最后一行的文字跨页打印在下一页第一行的情况，PDF 文件解析仍然存在困难，因此自动验收检查项目档案成果目录信息遇到这种情况时，系统会标记该处原文跨页，待人工检查。

4.2　自动验收实验及结果

为了检验自动验收系统的性能，在一台装有档案管理系统和 SQL Server 2005、2.6GHz CPU、2G 内存的 PC 机上对自动验收系统进行了测试。适逢 2012 年 4 月化学科学部工作人员利用人工验收系统验收了申请年度为 2002—2005 年期间的项目档案。因此，我们以化学学部申请年度为 2005 年的 1020 卷项目档案为实验数据比较自动验收与人工验收的效率和准确性。

人工和自动验收均对项目档案数字化档案著录和扫描图像进行了质检，质检内容覆盖了除实体验收以外的验收工作，共计检查 48 961 项内容，其中档案著录准确性检查 36 735 项，规范性校验 8160 项，图像质检 4066 项，具体质检数量见表 1。

表 1　2005 年化学项目档案验收统计

质检类型	质检内容	质检数量
著录准确性	项目基本信息	12 204
	课题成员信息	6 403
	成果目录信息	18 128
著录规范性	著录格式	4 080
	枚举型著录项	3 060
	项目执行年限	1 020
扫描页替换	替换位置	3 046
	替换页序	1 020

由表 1 可见项目档案验收时需要质检的内容繁多，验收工作量巨大。自然科学基金委化学科学部的一名工作人员利用设计良好的人工验收工具审核 2005 年的项目档案耗费 10 个工作日，检查数字化与原文不一致问题 48 处，其中经数字化工作人员排查确为数字化错误并改正的 14 处，其他不一致是由于档案信息来源于项目管理系统而档案中无对应文件的原因产生的。

利用自动验收工具对同一批项目档案进行验收，用时约 6.5 小时，验收效率明显提高。自动验收工具共计检查出档案著录和图像替换与原文不一致之处 2232 项。数字化工作人员对不一致内容逐一排查，发现确为错误的 1258 处，占项目档案数字化质检总项数的 2.57%。

表 2 显示了数字化工作人员对自动验收不一致问题排查的统计结果，并给出了与人工验收结果的对比。其中不一致数量表示验收过程中对比档案著录项和页面替换与档案电子文件的差异数量，但不一致是由多种原因造成的，其中一部分确为数字化错误。表 2 也给出了数字化错误数量及错误数占各项质检总量的比例，错误率越高表明验收质检的效果越好。由此表可以看出，成果目录信息出错率较高，原因是成果目录信息中包含大量希腊字符和上下标等特殊符号，这些符号因数据库字符集限制和 OCR 识别率低而容易出现录入错误。此外，部分不一致内容并非著录错误，主要是由验收工具本身无法处理跨页成果信息或 PDF 页面定位不准确等原因造成的误判。

表 2　自动验收与人工验收结果比较

验收方式		项目基本信息	课题成员信息	成果目录信息	著录规范性	替换位置	替换页序
人工验收	不一致	19	10	19	0	0	0
	错误	8	4	2	—	—	—
	比例(‰)	0.7	0.6	0.1	—	—	—
自动验收	不一致	197	232	1684	0	24	91
	错误	84	8	1162	—	4	0
	比例(%)	0.69	0.12	6.41	—	0.13	0

注：比例表示质检中确定的数字化错误占各类型质检项目的错误率。

通过对比分析人工验收和自动验收结果，发现人工验收检查出的错误全部包含在自动验收检测结果中。自动验收具有效率高、对数字化错误查全率高的优势，因此是一种有效的验收手段。

5　总结与思考

数字化是档案信息化管理和网络化利用的基础，验收是数字化工程质量的保证。在科学基金项目档案数字化过程中，提出利用计算机辅助完成项目档案质检的自动验收方法，提高了数字化工程验收的效率和质量。通过项目档案数字化工作，我们发现实施文件、档案一体化管理的重要性和迫切性。在业务系统中对归档需要的元数据进行前端控制，实现档案基本信息和电子文件自动归档，可大大减少档案著录和图像扫描的工作量，既能充分利用资源，又能最大程度地保障电子档案的质量。科技档

案蕴含着大量的信息财富，提供利用是档案数字化的最终目的。如何利用数据挖掘等现代信息技术把档案信息库转变为社会科技发展和科学基金项目管理的知识库将是我们的更高目标。

参 考 文 献

[1] 王传宇，张斌. 科技档案管理学，第 3 版. 北京：中国人民大学出版社，2009.
[2] 薛四新，肖怀志，杜文晓. 数字化技术与存储解决方案. 山西档案，2004，(1)：26—29.
[3] 贾玲，周晓林，陆江等. 从档案实体管理、信息管理到档案知识管理. 中国档案，2012，(2)：42—44.
[4] 林祥振，陈亚军. 一种智能文档存储格式方案及其应用. 中国档案，2012，(6)：69—71.
[5] GB/T 3792.5-1985. 档案著录规则.

致谢 感谢宋晓哲和时晓东在软件开发和数字化方面的工作。

APPLYING PDF DOCUMENT ANALYSIS IN ACCEPTANCE CHECK OF NSFC DIGITAL PROJECT ARCHIVES

Feng Qi Wu Ning
(*National Natural Science Foundation of China*, *Beijing* 100085)

(上接第 348 页)

经验的同时，必须看到其所赖以生发的宏观制度环境，从整体上思考制度的借鉴或法律移植问题。这不仅要靠科学基金机构的努力，更要加强立法、行政甚至司法等各机关之间的协调互动，共同努力，从完善国家宏观的制度环境方面，加强立法及实施合作，才能奏效。

(3) 制度建设的重点在于分清各方权责、强化监督。期间必须明确资助单位、研究项目依托单位、研究人员、伦理审查机构，在项目的申请、立项资助、项目实施、结项及结项后等各个阶段，分别承担怎样的义务、责任，以及参与者享有哪些权利，以及如何在这些主体之间建立良性的、公开透明的、合理的监督与互动关系，是相关立法必须予以关注和解决的基本问题。

参 考 文 献

[1] The Belmont Report—Ethical Principles and Guidelines for the protection of human subjects of research, [EB/OL] [2012-9-11]. www. nih. gov
[2] Title 45 CFR Part 46 protection of human subjects. [EB/OL][2012-9-11]. www. gpo. gov
[3] National Statement on Ethical Conduct in Human Research (2007), [EB/OL] [2012-9-11]. http://www. nhmrc. gov. au/
[3] 2nd edition of Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans. [EB/OL][2012-9-11]. http://www. pre. ethics. gc. ca

SCIENTIFIC ETHICS SYSTEM OF FOREIGN SCIENCE FOUNDATIONS

Tang Weihua[1] Wang Guoqian[2]
(1 *The Law School of Qingdao University*, *Qingdao* 266071; 2 *National Natural Science Foundation of China*, *Beijing* 100085)

Abstract Scientific Ethics System is an important part of Foreign Science Foundations system. The developed countries have established comparatively systematic Scientific Ethics System, which can bring us some law-making experience.

Key words foreign, science foundation, scientific ethics