

小语种文献数字化对网络文献集成揭示的实践意义

□ 施明红 / 机械工业信息研究院 北京 100037

摘要: 文章通过对小语种文献数据处理解决方案的研究与应用回顾,展示了国家科技图书文献中心(NSTL)在小语种文献加工及多语种数据处理方面的技术成果。该方案的实施,进一步提升了NSTL网络服务系统的功能,使得NSTL在国内文献服务领域率先解决了小语种文献的数字化加工和网上文献服务的多语种显示、检索等问题,对于网络服务系统多语种信息集成揭示具有重要的实践和示范意义。

关键词: 数据处理, 小语种文献加工, OCR识别, 字符集编码, 网络信息服务

DOI: 10.3772/j.issn.1673-2286.2010.10.015

1 概述

国家科技图书文献中心(NSTL)科技文献网络服务系统自2000年12月开通以来,通过其不断丰富信息资源、方便检索的文摘数据库和快捷的原文传递系统,为广大用户提供了全新的获取国内外最新科技文献的服务模式,从而真正实现了科技文献的资源共享。

目前,NSTL网络服务系统已经为更多的读者所认知和利用,业内影响不断扩大。全新改版上线的3期系统暨国家科技数字图书馆,各类文献的数据总量近1.2亿条,成为国内最大的文献信息网络服务平台。

在NSTL十年来的发展建设中,中心领导始终意识到新技术的应用和升级进步,是保持网络服务系统生命力的重要手段,并通过2期和3期的改造,不断完善和提升了网络系统的服务功能。与此同时,网络系统揭示的外文资源,也从初期的单一英文语种,扩展到英、法、日、德、俄等主要语种,既丰富了资源内容,也更好地满足了科技工作者的文献检索需求。

2 需求的提出

2000年6月NSTL成立后,经过近3年的发展,到

2002年底中心各成员单位订购的外文期刊及会议文献已经达到了一万余种,其中小语种占有一定比例,主要为德、法、日、俄等工业发达国家的原语种文献。毫无疑问,这些原文文献对我国的科研发展和技术进步有着重要参考借鉴作用,但当时在同一系统环境下实现多语种文献的全面揭示和整合利用在国内还是空白,技术上存在一定困难,致使这类文献未能在NSTL网络服务系统得到充分揭示,影响了用户的利用。

如何尽快解决小语种文献的加工和网上揭示服务成为NSTL面临的一个紧迫问题。为此,2003年初,中心领导决定开展小语种文献数据加工解决方案的可行性研究,并成立了由各成员单位业务及技术人员组成的课题组,力争半年内拿出切实可行的操作方案,并在试点单位试运行。

课题组的初步调研结果显示,虽然许多国家都在积极探索多语种文献的数字化加工揭示方法,但尚没有一种被广泛采纳的成熟解决方案,参考文献也很匮乏。此时,我们感到了很大压力,也深感责任重大。

3 研究原则与目标

3.1 坚持实用性原则

面对这个当时业内的共同难题,课题组的同志们并没有退缩,大家的共识是以融合并充分利用各种适用技术为主要思路,努力寻求在现有中英文环境下适用的集成解决方案。

2003年3月,在NSTL袁海波主任主持下,课题组在前期调研基础上,经过认真讨论,确定了坚持实用性的研究原则和阶段目标。

3.2 目标策略

3.2.1 从易到难,分阶段实施

充分利用和借鉴NSTL现有英文数据处理模式和成熟方法,既考虑加工,更要考虑用户应用的可操作性,从易到难,分阶段解决,逐步实施推广。

3.2.2 操作要简便易用

重点从加工方法的易用性、操作系统与应用系统平台对现有环境及字符集的可支持性、用户检索的便捷性等方面入手进行研究,提出可操作的实施方案。

4 研究方法

明确了方便实用、可操作的原则和目标,课题组根据从现状调研—方案构建—模拟测试—分析论证—调整完善—实地运行的计划步骤,重点进行了国内外小语种文献数据处理状况调研,数据加工流程与文本获取方式分析,解决方案与可操作加工环境构建,实际模拟加工测试等几方面工作。

调研的内容主要包括小语种文献数据处理的方式、层次深度、系统平台、软件环境、数据处理的标准规范、检索与输入方式、时效与成本投入等。

4.1 技术应用状况

针对国内外小语种文献数据处理技术及应用状况,我们首先查询了相关主题的国内外文献。由于可借鉴的参考文献很少,课题组主要采用了调查走访的方式,调研了国内较知名的多家图书馆、文献服务机构和应用系统。结果表明,国内对小语种文献的数据处理主要应用于图书馆馆藏文献编目,但在数据输入

时,仍需挂接相应语种的输入法来处理小语种文献的母体数据,并未见用于小语种文摘数据处理的应用。

4.1.1 国外应用

国外主要国家的大型文摘检索系统和数据库,如日本的科技文献速报、俄罗斯的文摘杂志、英国的科学文摘、美国的化学文摘、医学文摘、工程索引等,多年来一直采用其本国母语报道世界各国科技期刊文献中发表文章的文摘信息。对非母语的文章,或者选择有英文篇名、摘要的收入揭示,或者将相关信息翻译成母语,更多是取其发音的拉丁音译,并通过英文关键词、主题词标引等辅助手段揭示其主题内容。这种方法周期较长,投入很大。

4.1.2 国内应用

国家图书馆启用的ALEPH 500系统,为引进以色列开发的图书馆集成管理系统,支持UNICODE字符集,但当时没有进行小语种文献数据处理;高校CALIS系统中,仅有北京农业大学成员馆对日语文献编目的输入法应用做过有关研究;深圳图书馆开发的ILAS图书馆自动化集成管理系统,支持UNICODE字符集,能够处理少数民族文字,但对小语种文献的加工仅限于书目级别的采编功能,数据量较小且应用并不广泛,不能提供对文摘级别大数据量处理的实际应用。

4.2 调研结果分析

上述这些单位均为国家级大型文献服务机构,或在文献信息处理方面技术较成熟,但仍未较好地解决小语种文献深加工的数据处理问题。

调研结果表明,对小语种国外科技文献的数字化加工处理和网络数据库建设,当时在国内基本还是空白,也未见到哪个系统能在同一平台环境提供小语种文献信息服务的,这对我们来说既是挑战,也是机遇。

5 技术路线与难点

5.1 流程描述

NSTL各成员单位采用的英文数据加工流程虽不尽相同,但大致可分为以下几个环节:

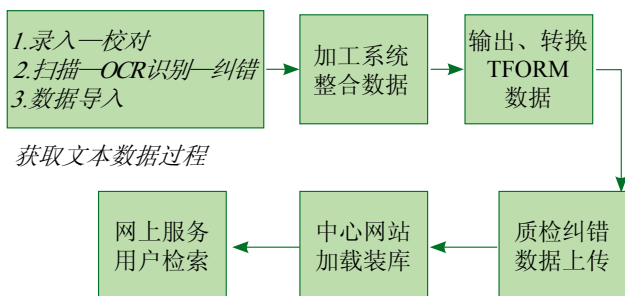


图1 数据加工流程示意图

图1中斜体字表示的框为获取文本数据的过程,主要有录入、扫描识别、数据导入三种方式。它处于加工流程的最前端,可在加工系统平台外部完成文本的获取,而不受该加工系统对不同语种字符集支持与否的影响。

5.2 技术难点分析

文本的正确识别、获取,与OCR软件的功能和性能密切相关,它也是小语种文献数据处理流程的主要环节。通过对加工流程各环节的分析,我们发现,只要找到相应语种文本数据获取的工具或方法,并解决各语种专用字符集的同平台显示和索引问题,小语种文献数据处理的主要难点就可以克服。

参照NSTL各成员单位应用系统环境和英文数据加工的经验,课题组对加工流程中影响小语种文献处理的主要因素进行了认真分析,认为主要的技术难点是如何获取相应语种的文本数据,并能够支持同一操作系统平台环境所应用的字符集。

5.3 技术路线

解决方案立足于流程加工的可操作性,以短期内可实现应用、对系统运行环境影响最小为目标。基于此,我们认为在中英文Windows系统环境下,利用中文GBK字符集支持日、俄文语种字符的编码及显示功能,将获取的小语种原编码文本数据转换为GBK字符编码是最简便可行的办法,经过实际测试,结果证明基本上是可操作的。

6 方案的确定与测试

经过对德、法、日、俄等小语种文献数据反复进行实际操作测试及可行性分析,课题组确定了在Windows系统平台下进行小语种数据加工的解决方案,并在各成员单位构建了加工测试所需的软件环境,分别进行了可操作性测试。

6.1 方案描述

该方案的主要思路是在各单位应用的英文数据加工系统环境下,通过相应语种的OCR扫描识别软件,对小语种文献加工数据进行图形—文本转换,获取其可编辑的文本数据,同时利用Word 2000支持UNICODE字符集的特性,将获取的含有各语种专用字符集的文本数据转换为国标GBK字符集的文本数据,以满足现行中英文加工系统的字符显示及排序要求,实现按原语种进行数据加工的目的。

6.2 加工环境构建

加工环境的搭建由客户端系统环境和配套的各种应用软硬件组成。

6.2.1 系统环境

客户端系统环境采用Windows XP或Windows 2000 Pro,由于对小语种文本数据的OCR识别、提取和校对处于整个数据加工流程的前端,即需要在导入加工系统前,在外部完成对小语种数据的文本处理,故对原有英文加工系统程序的运行环境没有特殊要求。

6.2.2 软硬件配置与应用环境

- (1) 理光1035高速扫描复印机
- (2) 美国ABBYY公司FineReader 7.0 Professional Edition OCR识别软件
- (3) 日本A.I. SOFT, Inc.公司KOKO 9.0日文OCR扫描识别软件
- (4) 沈阳格微软件公司“日语一点通”及中日文输入平台
- (5) Office 2000

- (6) Adobe公司Acrobat 6.0版
- (7) NSTL数据加工系统及TFORM转换系统
- (8) NSTL数据加工质检系统

6.3 文本获取步骤

文本获取主要分为四个步骤:

- (1) 利用相应语种OCR软件对小语种扫描数据文档进行识别。
- (2) 文本转换, 将经过纠错后的文本数据传送到Word 2000, 实现各语种专用字符集到中文GBK字符集的自动转换。
- (3) 校对过程中选择应用相应语种输入法。
- (4) 提取完成校对加工的数据, 转换为TFORM格式, 并经质检软件检查无误后上传中心网站。

6.4 OCR识别效果测试

在利用OCR获取文本的识别测试中, 各语种OCR的识别效果在原件质量良好的情况下, 用300DPI以上分辨率扫描, 识别率能够达到99%以上。

OCR识别样例如图2-3所示。

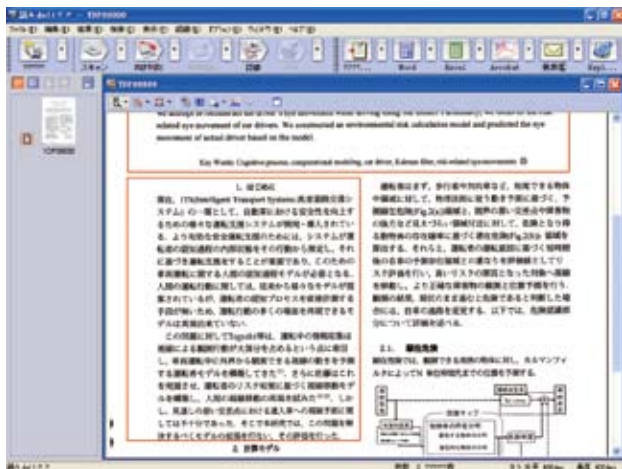


图2 框选识别区域—日文

识别步骤与加工英文数据基本相同, 但在处理识别错误时, 需在OCR的文本结果窗口进行双窗口对比校对, 无法像处理英文时, 可利用字典的辅助提示功能。在切换相应的输入法输入正确字符时相对稍慢, 加工速度与英文相比有一定差别。

以上加工流程的各个环节, 均经过德、法、日、

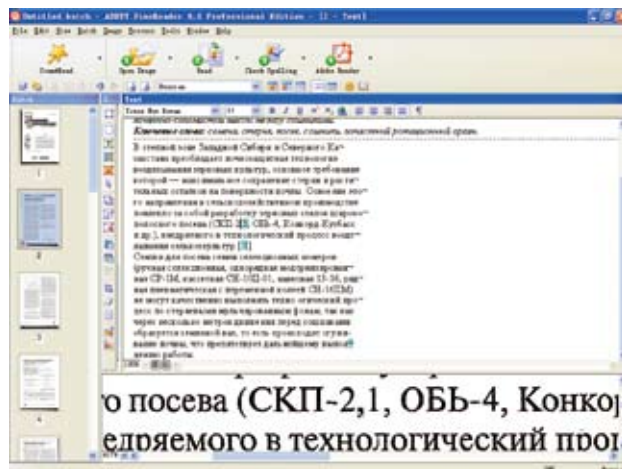


图3 OCR文本校对窗口—俄文

俄语种文献的实际模拟加工, 并成功进行了符合中心网站数据加载要求的TFORM格式数据转换, 通过了NSTL质检程序检验、数据上传、网管中心试验库加载等测试。

6.5 WEB端用户查询方式

由于加工中获取的文本数据已经是GBK字符集, 在现有中英文系统平台能够正常显示和排序, WEB端用户可利用Windows配置的常用中文输入法, 选择对应语种的软键盘输入检索词。日文汉字可用中文全拼输入法或支持GBK的中日文输入法输入, 无需考虑字符集的问题。由于通常需输入的检索词字符不多, 软键盘输入方式对读者是否熟悉相应语种字符的输入影响并不大, 只要在帮助页面中对输入方法加以说明即可。

7 研究解决的主要问题

7.1 建立小语种记录唯一标识生成规则

在方案测试中我们也遇到了大大小小的不少问题, 主要问题是要解决小语种记录唯一标识的生成策略, 以便于质检程序进行查重。

此前, 在西文文摘数据上载到NSTL网络系统后台数据库的流程中, 质检程序提供了数据查重功能, 避免冗余数据, 排除重复上载。其规则是依据国家科技图书文献中心连续出版物及其文献标识创建细则, 通过篇名单词首字母、ISSN、母体刊名及页码段等数据

组合的算法生成唯一记录标识。

由于日、俄文数据篇名字符的特殊性,原质检程序因未建立相应的字符对应规则,无法生成准确的唯一记录标识(SICI)。为此,课题组分析了日、俄文等不同小语种字符集的特点,提出了以《中国机读目录格式使用手册》附录的“日文假名罗马字拼音表”及“俄语字母和拉丁字母音译对照表”映射到西文字符的解决办法。

7.2 质检程序改造

在上述规则确立后,课题组依据新的对照规则,编写了小语种记录唯一标识生成程序,并对原有质检系统进行了优化改造,使其完全兼容了对多语种数据的质检处理。

7.3 解决的主要难题

经过广泛、深入的技术调研,在系统研究国内外关于小语种文献数字化加工和网络信息服务成果基础上,课题组对加工流程、输入方式、字符集应用进行充分分析和反复试验,比较、分析和评测各种方案的可操作性,解决了日、俄语等不同语种文献加工所需OCR识别软件的遴选、文本输入及校对、流程优化、集成配套应用等一个个难题,按时提出了在中文Windows系统环境下实现小语种文献数据处理的解决方案。

8 研究成果



图6 日文检索页面

2003年10月上述解决方案通过了中心组织的专家组评审,其后经过机械工业信息研究院、中国农业科学院图书馆、中国化工信息中心等三家成员单位对试用系统3个月的运行实践,进一步总结了经验,完善了实施方案。

8.1 研究中取得的主要成果

(1) 配套制定了小语种文献相关数据项的加工细则和质检规则;

(2) 日俄文小语种字符对照表及唯一识别符(SICI)的生成规则;

(3) 优化完善了记录唯一标识生成程序和小语种文献专用质检软件;

(4) 培养了一批小语种文献数据加工业务骨干。

8.2 方案实施

在小语种文献数据加工系统试运行取得良好结果的基础上,2004年初,课题组对各成员单位的数十名加工人员进行了小语种数据加工的专项培训。至此,NSTL全面启动了小语种文献的数据加工工作。

8.3 页面展示

图6-图11为NSTL网络服务3期系统中小语种检索页面展示。



图8 日文检索结果文摘页面



图7 日文检索结果题录页面



图9 俄文检索页面



图11 俄文检索结果文摘页面



图10 俄文检索结果题录页面

9 结语

2010年6月12日,在NSTL成立10周年之际,新的网络服务系统3期平台已经成功开通上线。截至目前,上载到后台数据库中的日文期刊文摘数据已近120万条,俄文期刊文摘数据近48万余条。德、法文文摘数据由于按西文字符处理,直接装入西文期刊库。我们期待着NSTL小语种会议文献文摘数据能够早日启动上

载,为读者提供更加丰富的外文文献信息。

小语种文献数据处理解决方案的实施应用,实现了对多语种文献资源的集成加工揭示,进一步提升了NSTL网络服务系统的功能,使得NSTL在国内文献服务领域率先解决了小语种文献的数字化加工和网上文献服务的多语种显示、检索等问题,对于网络服务系统多语种信息集成揭示具有重要的实践和示范意义。

作者简介

施明红 (1958-), 现任机械工业信息研究院情报研究所副所长, NSTL资源工作组副组长。多年从事外文文献的采集、编目管理以及数据库建设等工作。研究方向: 数字图书馆发展、文献数据库建设、数据处理技术。通讯地址: 北京百万庄大街22号 机械工业信息研究院 100037。E-mail: smh@mail.machineryinfo.gov.cn

The Practical Significance of Non-English Original Language Digital Processing for Network Integration Showing

Shi Minghong / China Machinery Industry Information & Publication, Beijing, 100037

Abstract: This paper introduces the Non-English original language digital processing solution's research and application process, and shows the achievement of National Science and Technology Library (NSTL) in the multilingual data processing field. The implementation of this program further improved the capability of NSTL network service system, and for the first time solved the problems of Non-English original language digital processing, as well as multilingual display and data retrieval in the network services system. It is good practice and exemplary case to show multilingual information integration of network service system.

Keywords: Data processing, Non-English original language data processing, OCR recognition, Character set encoding, Network information services

(收稿日期: 2010-08-30)