案例研究
教育

ABBYY FineReader® Engine (SDK) 将科研论文转换为数字知识

客户概况

中国知识基础设施工程 www.cnki.net

客户名称

中国国家知识基础设施工程

地点

中国

行业

教育

挑战任务

数字化编排多语种科研文献并保留文献布局格式。

解决方案

实施一套基于 ABBYY FineReader Engine 的解决方案。

最终结果

- 提高了效率和准确度；
- 完善保存了文档结构与布局。

中国国家知识基础设施工程（CNKI）是一项得到中国教育部、科技部、中宣部、新闻出版广电总局大力支持的电子出版工程。

该工程提供了中国超过 90% 的知识资源，其涵盖的文献主题、文献种类、地理范围和年份范围均为全国之最。数据库收录了众多科研领域的学报、学位论文、报纸、会议记录、年鉴、参考文献、百科全书、专利、标准、科技成果和法律法规。

中国九十年代末的大规模知识资源数字化工程开创了中国最为兼容并包的学术知识平台。1999年，清华大学与清华同方共同开发了中国知识资源整合数据库，并提出了中国学术期刊的标准体系。如今中国的每一位科研工作者都在使用这一平台，所有的学位论文与科研成果也都会援引该平台的知识资源。

挑战任务

CNKI 专注于教育领域，收录了大量的图书、文档、期刊、博士论文、报纸等中外文纸质资料，而这些资料都需要经过数字化处理后，整理编入易于检索的知识数据库。数据库每天都会收录数以百计的新档案，每个新档案中还有数以千计的新条目。

除了卷帙浩繁的收录资料，繁杂的资料语言也是亟待解决的问题。资料涉及的语言有中文、越南语、泰语及绝大多数的欧洲语言等。此外，科研著作与学位论文特有的大量插图、表格、方程、制图、图表等也至关重要，需要尽数保留。所有资料还需编入索引，保存为特殊的 CAJ（中国学术期刊）格式。

鉴于上述难点，采用人工录入将资料数字化费时费力，并给 CNKI 带来巨大的负担。所以 CNKI 采用了一家中国本地厂商的 OCR 解决方

ABBYY 解决方案合作伙

泰比 ABBYY

abbyy.cn

案来实现自动化并提高录入效率。与人工录入方式相比，OCR 方式显然更快、更好，但仍未达到预期效果。

首先，由于该系统仅支持中文，所以有相当数量的资料无法识别。其次，识别质量欠佳，校验结果花费了大量时间与人力。最后，该系统仅能捕获文本，无法保存文档布局和其他元素。

解决方案

为了寻找替换的 OCR 核心解决方案，CNKI 致函上海泰彼信息技术有限公司，全球领先的 OCR 与数据采集技术提供商 ABBYY在中国的代表处。

为了在最短时间内完成积压资料的数字化，泰彼公司建议采用 ABBYY FineReader Engine — OCR 软件开发工具包方案，以实现与 CNKI 现有环境进行深度无缝整合。

在数字化第一阶段，ABBYY FineReader Engine识别出文档中的所有文本。在第二阶段，该引擎从文档内容中抓取检索值（元数据）。利用元数据，可以实现知识数据库中数字化资料的快速高效检索。

与以往的 OCR 方案相比，ABBYY FineReader Engine能够保存文档的原始布局，并将经过处理的文档导出为 Microsoft® Word 文档、Excel® 文档、可检索的 PDF/A 文件，以及符合中国国家标准的 CAJ 本土格式文件。

只需一名操作员，即可快速轻松地校验 ABBYY OCR 识别结果，并确保检索结果 100% 精确。

最终结果

采用 ABBYY OCR 技术后，CNKI 显著提高了资料处理速度与精度，减少了人工干预。ABBYY FineReader Engine的智能文档分析功能保存了导出文档的结构与布局，确保文档将来在 CNKI 数据库中的有效使用和存储。

通过使用多核处理，资料识别速度得到了显著提升。在过去，同样的任务需要耗费数周时间，而现在仅仅需要几天。得益于自动化资料处理，CNKI 能够将原本从事人工录入和校验资料工作的数十名员工解放出来，投入其他项目的工作，大幅提高了生产力。

最重要的是，这类大规模数字化工程的最深远影响在于提升了使用者舒适度。现在这一全球平台的用户都可以更快的速度、更高的准确度搜索到所需信息。 ABBYY 的数字化解决方案令中国国家范围的知识更易于检索、便于使用，完美践行了 ABBYY 的企业使命 — 知行合一。

"ABBYY 是国际知名的 OCR 技术提供商，其 OCR 识别精度、甚至中文的识别精度都远远超乎我的期望。ABBYY 的技术使我们节约了大量时间，提高了工作效率。我们希望双方能够进一步合作，优化我们新的工作流程。"

-CNKI 技术总监吴先生

ABBYY 3A
Asia, Baltic, Middle East, South America, Africa
P.O. Box #32, Moscow, 127273, Russia
俄罗斯电话： +7 (495) 783 3700
传真：+7 (495) 783 2663
sales_3a@abbyy.com

ABBYY