第五届两岸四地中文数字化论坛发言

中文数字化，勿以初级而不为

张轴材 北京书同文数字化技术有限公司
joe.zhang@unihan.com.cn

2007 年初第四届两岸四地中文数字化论坛（澳门）之后，我和我的团队，在包括两岸四地在内的众多朋友的帮助下，在中文数字化特别是汉字数字化的几个初级的、基础性的项目上做了一些工作。现汇报交流如下：

一、汉字基础项目

a) 中日韩常用汉字对比分析：该项目从汉字文化圈社会生活与教育界最常用的汉字入手，选中国大陆的《现代汉语常用字表》、《HSK 汉语水平考试字表》、中国台湾地区《国小字表》、中国香港特别行政区《小学生用字一览表》、日本《当用汉字表》和韩国文部省指定的汉字表为对比对象。将它们纳入 ISO/IEC 10646 的国际标准的框架下的数据库，分别进行覆盖率统计与字形异同的对比分析，并按照汉字简化、正形、异体代换映射的结果，分列出清晰的“同形同码”、“微差同码”和“简化异码”等多组对比字表，及其诸子集的频率权重统计图表。《中日韩常用汉字对比分析》作为一项处于语言文字与信息技术边缘的研究项目，由张轴材先生主持，依托北京书同文数字化技术有限公司，充分利用数字化技术，开展中日韩和港台地区的多边合作，“取得了具有实用价值的成果”。日前该项目通过了教育部语言文字信息管理司主持召开的鉴定会，获得了来自北京大学、北京师范大学和商务印书馆等教育出版界的专家的好评。根据鉴定组专家也提出的意见与建议，该项目组利用书同文公司的数字化资源进一步开发，完成了从 V3.30 版到 V4.0 版的大规模更新。目前已决定纳入《国家语言生活绿皮书》由商务印书馆出版。 见“CJK 求同询异”http://hanzi.unihan.com.cn/CoolHanzi/