标准、平台统一与资源整合

——多语言知识库建设的思考和建议

王铁琨

“多语言知识库”是教育部语言文字信息管理司立项支持的民族语言文字标准化、信息化重大项目，也是“中国语言资源数据库”建设和“民族语言文字信息化平台”的重要组成部分。

语言不仅是信息的载体和交流的工具，语言更是重要的、不可再生的文化资源。

基于珍爱中华语言资源的理念，中国正在酝酿开展新世纪的语言普查（后更名为“中国语言资源有声数据库建设”，作者注），以期建立可永久保存的中国语言多媒体语料库及相关数据库， 绘制详细、准确、可传至后代的多媒体语言地图， 建立需要保护的语言、方言目录，开发和利用好国家语言资源。

语言权利包括个人和群体的语言权利，涉及公民的生存权和发展权。应该采取有效措施，切实维护和保障公民和群体的语言权利， 如母语学习权、母语使用权和母语研究权，以及获得各种语言服务的权利。

语言文字资源库（包括“多语言知识库”）建设还受到标准、平台和资源等方面一些因素的制约。这三个方面是建库的前提，必须着力解决好。

一是统一标准。“多语言知识库”建设首先要坚持统一标准，否则无法做到兼容和共享。这个标准就是国际标准，以及在ISO/IEC 10646框架下的国家标准。随着信息化的发展，现在整个地球都成了一个“村”，我们在标准建设上也不能搞“窄轨铁路”，自我封闭。在建库步骤上，可以考虑已有国际标准的语种先建，从“双语”开始，逐渐形成“多语”。

除了标准的大的方面。数据库、知识库建设中标准、规范无处不在，其中包括选材（资源选择）规范、资料整理规范、文本录入规范、校对规范和建库规范等一系列内部的规范和标准，都需要先行统一。否则，各行其是，各自为政，库建起来也发挥不了预期的效用，甚至可能花钱“打了水漂”。

二是统一平台。“多语言知识库”建设要坚持统一平台。中国各民族文字多种多样，有纯表音的，有表意的，有表音兼表意的，有罗马字母式的，有汉字系的（如古壮文和西夏文），有的从左至右书写，有的从右至左书写，有的自上而下书写……现在信息技术发展越来越快，如何采用必要的技术手段，使上述不同体系、不同书写方式的文字在一个统一的平台上实现兼容和共享，是“多语言知识库”的平台建设要解决的问题之一。

三是资源整合。语言资源有语种之分（如汉语、藏语、蒙语、维吾尔语等），有地域之分（民族语言也有各种方言，如藏语就有三大方言），有古今之分（如蒙文有现代蒙古文与古八思巴文），有境内境外之分（许多跨境语言文字如傣文、苗文、哈萨克文，境内外有程度不同的区别），从载体上又可以区分为平面媒体语言、有声媒体语言和网络媒体语言。如何科学地采集和整合这些重要的语言资源，也需要研究和规划。

语言资源不存在统一不统一的问题，而有一个是否科学、合理和适用的问题，自然，采集语言资源的内部规范也需要统一。目的不同，资源库建设的类型和内容也会不同。资源库建设还有一个整合已有资源和开发新的资源的问题，力避重复建设。