构建中华语言信息化大平台

（序　一）

李宇明

中华民族有三十余种古今文字，因此便有很多“动天地、泣鬼神”的仓颉式人物。土弥·桑布札，便是这仓颉式的杰出人物。1300年前，他创制了藏文，使灿烂的藏族文化得以用书面语的方式保存与发展。他还有八部藏语文论著，《文法根本三十颂》和《音势论》一直传留至今。《文法根本三十颂》是藏语语法著作，用偈颂体写成，重点讲虚词与词格助词；《音势论》或译作《字音强弱论》，结合音义对藏文做了阴、阳字性的分析，研究了字性阴阳交替所形成的语音变化和时态、语态等语法变化。他创制的藏文和他的煌煌著作，不仅是藏族的文化财富，也是中华民族乃至全人类的文化瑰宝。

而今社会进入信息化时代，藏语文也大步跨上信息高速公路。藏文字符已经有了国家标准和国际标准；计算机藏文处理正阔步前进，开发出了一系列应用软件，并可以利用互联网收集和交换信息。我们在对土弥·桑布札的历史贡献表示敬意的同时，对新时代的土弥·桑布札，也充满敬佩之情。

《藏文字符研究》是大量扎实研究的总结，对藏文字符相关的问题，如字母、读音、编码、字频、排序、图形符号、拉丁字母转写规则等，阐述详细，立论精当。江荻博士是面向信息化进行藏语文研究的重要专家，1992年就发表了藏文文本的统计成果，且逐渐从字母统计发展到结构统计，从静态统计发展到动态统计。他还提出了藏文熵值计算问题，给出了藏文排序的数学模型与算法，并领衔设计了藏文拉丁字母转写方案。有此功底，《藏文字符研究》的学术水平和应用价值，就是不言而喻的了。

当然，藏语文信息化的道路还十分漫长，中华语言信息化的道路都十

分漫长。我认为，一个信息化的完整方案，应当包括三个层面。

Ⅰ. 语言文字层面，包括语言文字的各种知识和规范标准，如语音图谱，大字典，大词典，语法大典，语料库，语言知识库，储存、显示、交换用的文字编码，计算机字库，等等。

Ⅱ. 操作层面，包括语言文字的操作系统、应用操作系统和内容处理工具。语言文字操作系统主要是语言文字输入、传输、输出以及翻译的各种软件；应用操作系统是办公、商务、教育、大众服务等领域使用的各种软件；内容处理工具如百科辞书、术语标准、适应不同人群的内容检索系统，等等。

Ⅲ. 内容层面，存放各种内容数据，如商品超市、农牧业知识、经济常识、医疗卫生知识，等等。

Ⅰ、Ⅱ是信息工作平台，代表着信息化的科技水平，Ⅲ是这一平台所承载的内容数据。建造平台的目的，就是要充分积聚内容数据，并对其进行科学加工，以最大限度地满足各行各业的信息需求。这三个层面所涉及的软硬件，都应当根据统一的标准相互兼容，协调合榫。

中国是多民族多语言多文字的国度，有56个民族，百种以上语言，十来种常用文字，还有一些常用的外国语文。这些语言文字及其所负载的知识，也应当在这三个层面整合兼容；这些语言文字的信息产品，也应当在这三个层面里集成共享。这样就会筑成中华语言信息化的大平台，每一种语言文字信息化的成果，立即就会成为中华语言信息化大平台的成果，每一个学者的贡献，立即就能使中华语言信息化大平台的功能有所提升。常积跬步而至千里，广聚小流而成江海。这样，可以从根本上解决语言文字之间不兼容、软件之间不匹配、重复投入、少数民族语言信息化总是被动追赶的问题。

这是一个仰望星空时产生的宏大构想。宏大的构想需要有统一的规划、统一的标准，而统一的规划和统一的标准，来自统一的意志和宽广的胸怀。拥有无数个仓颉和土弥·桑布札的中华民族，相信在信息化的时代能够实现这一构想，搭建起中华语言信息化的大平台！

2009年五一国际劳动节

序于北京惧闲聊斋