中国中文信息学会2012战略研讨会

民族语言信息处理云计算之思考

那顺乌日图

2012年4月13日

江西婺源

谈：

战略，还是战术？
问题，还是技术？

一、意义

云计算技术是IT产业继PC、互联网之后的第三次革新浪潮，国家十二五规划把云计算作为新一代IT产业研发与应用的重要领域之一。利用云计算技术现实语言文字生活在网络空间中的扩展与延伸，是语言文字信息技术处理与信息资源有效利用的主要途径，因为语言文字及其所承载的信息资源是网络空间的主要媒介。

国家提出要大力加强民族语言文字规范化、标准化、信息化建设，要大力推进少数民族语言文字信息技术的规范化、标准化和信息化进程，建立少数民族语言文字的网络数据及各类信息资源传输平台，共享各类信息资源。

当前，国内外信息技术发展迅猛，但少数民族语言文字的信息化水平依然处于较低层次的阶段。少数民族语言文字信息化建设虽然经历了长达20多年的发展，已初具规模，但与走新型工业化道路，实现“跨越式”发展和后来居上的战略目标，尽快缩小与东部地区的信息化差距，填平“数字鸿沟”的要求极不适应。

还是处在文字处理阶段，在这个层面徘徊。没有深入语言深层（如，语义理解）。

个人设想：下次少数民族会议主要探讨语义处理。

具有自主知识产权的民族语言文字信息处理的核心技术比较少，缺乏具有领先水平的多文种统一的操作系统，应用软件之间兼容性差，开发性工具软件更是少之又少；少数民族文字网站的技术含量低，网页内容匮乏；政府部门缺乏统一规划、管理和强有力的引导，资金投入非常分散；软件开发各自为政，低水平、重复开发严重。

解决这些问题的重要措施之一就是开拓少数民族语言文字信息化建设的新局面，建立民族语言文字信息资源集成云平台，开发信息资源支撑技术，搭建信息资源共享云服务体系，从而加快民族语言信息资源工程，这已成为一项十分紧迫的任务。

关键词：多文种、云技术；资源建设、服务功能；整合、提升；

依托已有信息化设施，通过建设民族语言文字云信息中心的数据库和共享发布系统，进一步提高民族地区语言文字信息资源方面的数字化能力，形成为民族地区提供少数民族语言信息资源的服务体系，以贯彻落实国家少数民族地区语言文字政策，促进边疆地区的经济发展和社会稳定。

设想：民文信息处理如何为中文信息处理提供参考？是否有这个可能？（可否有个逆向思维？）

如，语义计算，蒙古语有构词附加成分是有聚合（类聚）关系。

VRALIG(艺术的)

BEHILIG（坚固的）

BAGATVRLIG(勇敢的)

HUDERLIG (健壮的)

SAHVLIG（茂密的）

GVWALIG（美丽的）

BAYALIG （富裕的）

SIMELIG（营养的）

NARILIG （细致的）

D0RBILIG（雄壮的）

MIHALIG （丰满的）

BIDAGVLIG（笨拙的）

SURLIG （雄伟的）

HUCILLIG（酸味的）

SIHIRLIg（甜味的）

SULIG（奶味的）

CIGIGLIg（潮湿的）

TEMURLIG（金属的）

CINEGELIG（富裕的）

HOROSOLIG （健康的）

BIRALIG（有力的）

BUDUGULIG（鲁莽的）

DABAGVLIG（优越的）

C0B0G0LIG（活拨的）

SIGESULIG （有汁的）

DURSULIG（形象的）

数据重要？

还是语言自身规律重要？

二、目标

- 1.建设基础资源、教育资源、科技资源、文化资源和公共资源整合等数据库；
- 2.建设媒体资源管理、多媒体业务播发、数据采编、内容分发、国家级信息共享等应用系统和门户网站；
- 3.依托“全国文化信息资源共享工程”、“全国农村党员干部现代远程教育工程”已有的网络和服务体系，为民族地区基层群众提供民族语言信息资源方面的服务；
- 4. 建设相应的安全保障系统和其他配套环境。

麻雀虽小五脏俱全（只吃猪肝？还是鸡胗鸡翅多样化？）

三、建设内容及规模

1.建设多语（文）种数字化成品资源*（上百TB？）；*

2.语言文字信息化技术支撑体系建设需要多种软件，包括应用工具软件，其解决途径可有：充分利用开放软件，从国内外购置现成的软件，现有软件的升级，在已有软件产品的基础上二次研发等。

3.建设国家多语种云信息中心，搭建该中心所有软、硬件系统，达到国外一流水平。

（1）公共基础支撑系统：网络系统、服务器、存储系统、安全系统、系统软件及工具软件。

（2）资源建设与管理系统：音视频资源制作子系统、文字资源制作子系统、图片资源制作子系统、动漫及多媒体制作子系统、资源管理子系统、计算机培训机房、多功能厅。

（3）信息共享与发布系统：卫星播发子系统、门户网站子系统、门户子系统、办公自动化。其中门户网站子系统中又包括基础软件与基本服务、电子图书、内容分发与视频点播、学习系统、共享系统、电子词典等在线服务。

四、措施

措施：

找出突破口，确定攻克方向；

“走出去”：去北亚，去中亚，如有可能去欧美；

联合：第一是在圈内联合（联合实验室）

第二是同国外联合

步骤：对至今民文信息处理进行一个盘点或梳理；

从战略高度进行一个规划；

少数民族专委会于2011年12月24日-12月25日在广西南宁召开了少数民族语言文字信息处理项目研讨会。这次会议的主要任务是在去年“中华统一信息平台建设”项目规划的基础上，进一步明确申报思路，细化任务分工，确定进度安排。

近年几个少数民族地区均获得过工信部软件专项和国家其他部委、基金支持，在制定相关标准、编码字符集，建立单/双语文字/语音语料库、开发民文操作系统、办公套件、输入法、电子出版、文字识别、语音识别/合成、互联网信息搜索、机器翻译等方面有了一定进展。但是，与汉语言文字信息处理相比，研究开发深度和广度远远不够，获得国家支持的主动性欠缺，已有成果的推广应用情况也不容乐观。

地方政府部门在资金、政策上提供优惠甚少，甚至出现抢政绩，控经费的现象。

与会人员对下一步工作重点进行讨论，大家一致同意构造重点项目的原则为：符合国家发展规划、追赶高新技术发展步伐，以汉语信息处理为基础与榜样，以获得自主知识产权为重要目标、避免重复开发。

会议初步确定项目的名称为“中华多语种信息服务平台”，该平台是各民族语言资源与技术共享平台，涉及到的研究开发内容包括：国产汉字操作系统及支撑软件的民文化；少数民族语言文字资源建设（不包括字体字库）；少数民族语言文字服务体系建设——各语言间的机器翻译，跨语言、跨文字信息搜索，文字识别，语音识别与合成，资源管理；移动终端民文软件；相关标准；民文信息安全；民文已有成果的应用现状调查等。

- 为争取国家对少数民族语言文字信息处理工作的更多支持，与会人员一致同意：
- 通过政协、人大会议（以提案和建议形式）向中央领导反映；*这次会议上几个自治区分头找各自地区的人大代表、政协委员提交到国家相关部门。*
- 通过财政部、国家发改委、科技部、工信部、民委等国家机关申请项目；*有些地区、像新疆已经开始立项，内蒙古有些地区也在积极争取项目。*
- 通过向领域专家介绍和汇报项目内容，引起他们的关注，获得他们在纳入指南范围方面的支持。*希望这次会议引起大家对这一工作的关注和支持。*

中华多语种信息服务云平台

监控
调度
管理

民文
标准

民文
移动
终端

其它终端

多样化终端

机器翻译 | 辅助翻译 | 信息搜索 | 跨语搜索

文字识别 | 语音识别 | 语音合成 | 资源管理

少数民族语言文字的应用服务

少数民族语言文字资源（除字体字库外）

国产汉字操作系统及支撑软件的民文化

基于虚拟机
集群的应用
层解决方案

数据库集
群解决方案

私有云备份
解决方案

私有云灾难恢
复解决方案

系统层

云平台节能
解决方案

以虚拟化技术为基础，对计算资源和存储资
源进行整合，作为私有云的基础平台

民文
信息
安全

安全
保障
体系


谢谢大家！
2006 7 12