

第三届中文数字化论坛

少数民族语言文字信息化

中国科学院软件研究所

ར་གོ་ཚན་རིག་ཁང་མཉེན་ཆས་ཞིབ་འཇུག་མཉེན་འཛུགས་ལྷན་ཁག་གི་ལས་ཁུངས་ལྷན་ཁག་གི་ལས་ཁུངས་

جوڭگو پەنلەر ئاكادېمىيىسى يۇمشاق دېتال تەتقىقات ئورنى

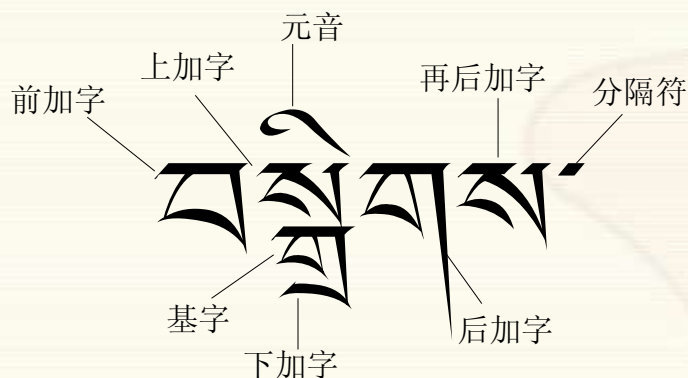
འཇུག་མཉེན་འཛུགས་ལྷན་ཁག་གི་ལས་ཁུངས་ལྷན་ཁག་གི་ལས་ཁུངས་

二零零五年十二月二十八日

民族文字处理的繁难

- 民族文字的特点
 - 藏文是一种拼音文字

- 藏文从左向右，从上向下两个方向拼读显示



བདམ་ལྷོད་མ་རིག་གནས་

雪域文化

- 维吾尔文——自右向左书写，字母位置不同，写法不同

- 维吾尔文字符H A H 独立使用和 在词首、词中、词尾不同位置的书写显现形式



存储顺序 →

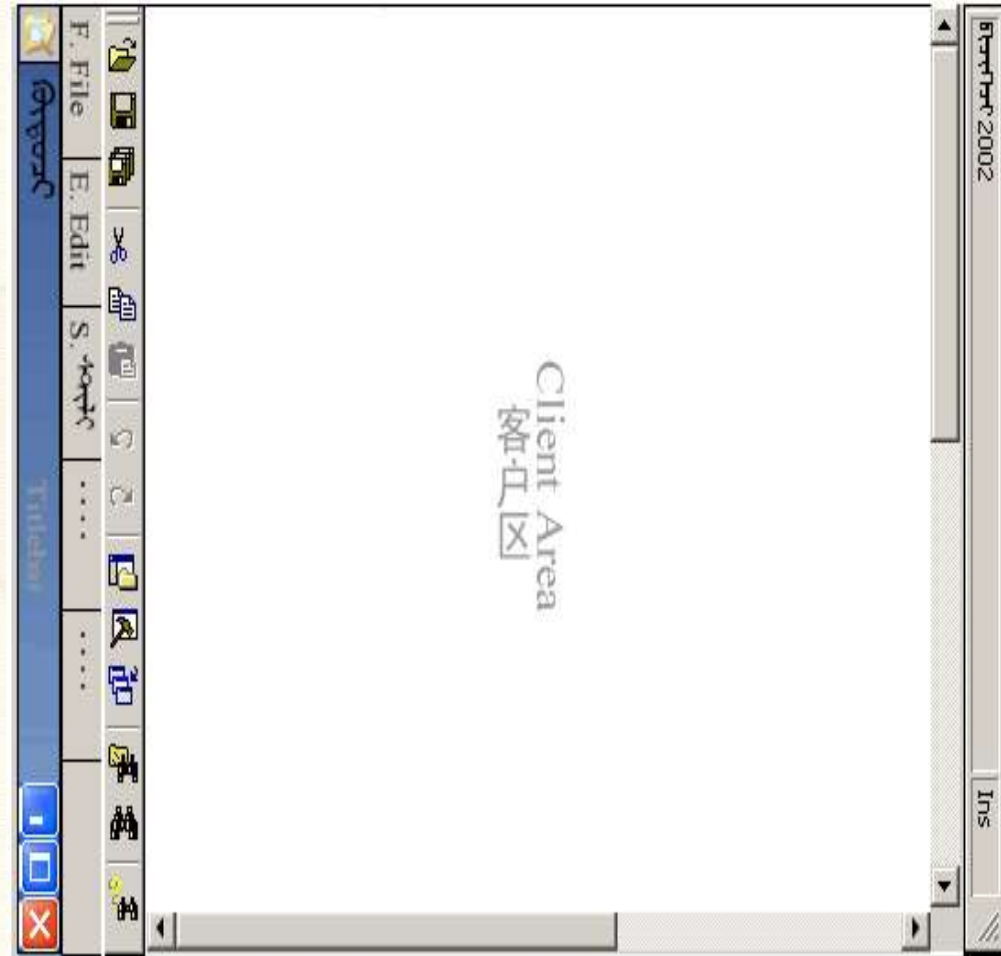
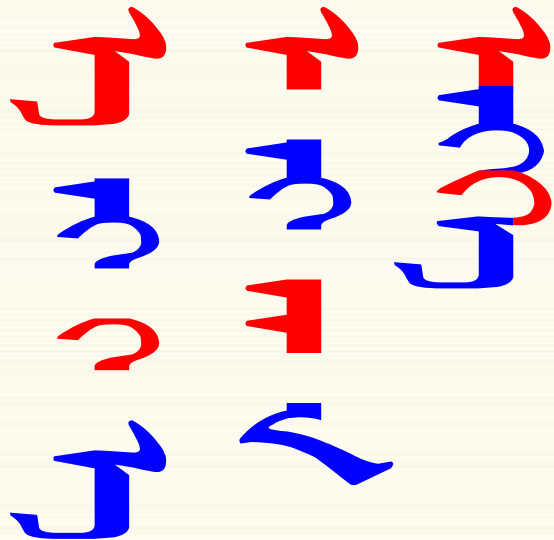
ب	ۇ	م	ا	ش	ى	ان	1	.	2	3	4	这	是	汽	车
---	---	---	---	---	---	----	---	---	---	---	---	---	---	---	---

这是汽车 1.234 بۇ ماشىنا

显现顺序 ←

民族文字处理的繁难

- 民文处理的特点
 - 蒙古文
 - 自上向下、从左到右黏着书写，字母位置不同，写法不同，而且变化更复杂(阴、阳...)



民族文字处理的繁难

- 民族文字集标准不充分

- 蒙古文、藏文、维吾尔文都收入到ISO/IEC 10646:1-

- 2000

- 编码原则：按抽象字形或字母编码，而不是按字义、字音、字型

- 藏文只收入基本字符

- 维吾尔文字符认定尚有争议

- 蒙古文只收入名义字符(抽象字符)，未收入显现字符

- 处理时，必须进行名义字符到显现字符的转换

ل ←

ل

0644 0645 062C

لم ←

لم

0644 0645 062C

لمج ←

لمج

0644 0645 062C

民族文字处理的繁难

- 典型的复杂文字
 - 传统的TrueType字库无法满足需要
 - OpenType字体技术
 - 在字库内部嵌入部分变换规则
 - 与字体引擎密切相关
- 民族文字应用需求
 - 多文种应用需求对软件处理提出了更高的要求
 - 在大陆，民族文字应用需要与汉字应用并存
 - 民文软件必须具有处理汉字、英文信息的能力
 - 不同文字基线不同，在处理时将会带来很多问题

བཞེས་པ་

中国

English

民族文字处理的繁难

- 一些系统(如, Linux)国际化与本地化体系结构
 - 对于支持双语环境支持较好
 - 基本上不具有充分的支持处理多文种的能力
 - 民文系统要求支持多种文字
 - 民(如蒙、藏、维)、汉、英文
-

民族语言文字信息处理技术现状

• 标准

- ISO/IEC 10646-1:2000在基本多文种平面收入了藏、维、哈、柯、彝及蒙文编码字符集
- 八思巴文字符集国际标准已完成提案
- 西双版纳傣文编码字符集国际标准已经完成
 - 提交ISO/IEC JTC1/WG2表决
- 制订藏文字符集扩充集、点阵字型、键盘标准
 - 在国际标准框架下
 - 以藏文垂直组合字符为编码单元
- 新疆维吾尔自治区政府启动《维哈克文标准化》项目
 - 定义ISO/IEC 10646.1:2000中的阿拉伯文字符与维吾尔、哈萨克、柯尔克孜文字符的对应关系
 - 向ISO/IEC JTC1/SC2/WG2提出对ISO/IEC 10646的补充提案
- 制订并将修订蒙、维、哈、柯、彝、傣文等标准

民族语言文字信息处理技术现状

• 民文基础软件

- 操作系统

- 一批基于DOS和Windows的民文系统
 - 维、哈、柯文DOS/Windows系列(新疆大学等)
 - 藏文DOS/Windows98/2000/NT(西北民大、青海师大)
 - 蒙文类DOS和Windows系列(内蒙古大学、蒙科立公司等)
 - 彝文DOS和Windows系列(西南民大)
 - 傣文Windows系列(青鸟华光、云南)
 - 纳西象形文字(大连民院)
 -

• Linux系统

- 中国科学院软件研究所牵头，民族地区研发机构参与
- 支持蒙古、藏、维吾尔等民族文字
 - 符合ISO/IEC 10646/Unicode标准
 - 支持复杂文本处理技术和OpenType字体

- 办公套件Office Suite

- 蒙、藏、维文OpenOffice(中科院软件所)
- 蒙文WPS(金山公司、蒙科立公司)
- 维吾尔、哈萨克、柯尔克孜文(无锡永中公司、新疆)

民族语言文字信息处理技术现状

• 电子出版系统

- 北大方正和北大青鸟华光

• 蒙古、藏、维、哈、柯、彝、傣电子出版系统

• 数据库与语料库建设

- 一批民族语言文字基础资源库

• 中国少数民族语言文字多媒体数据库(部分)

• 《500万词级现代蒙古语文数据库》和蒙古文语料库建设和词类词形标注与统计

• 中世纪蒙古语文数据库与《现代蒙古语词频统计》，
- 整理出《现代蒙古语频率词典》

• 藏文词频统计的研究与藏文电子词典开发

• 维吾尔文中小学多媒体大型教育信息资源数据库及其检索、索引技术，

- 创建了几万词的语料库、智能库、维汉英电子辞典等

• 满文档案数据库

•

民族语言文字信息处理技术现状

- 自然语言处理技术
 - 机器翻译与辅助翻译技术研究
 - 汉-蒙古公文、英-蒙古、满-汉、汉-藏、汉-维，等等
 - 面向互联网检索与分类
 - 印刷体识别
 - 蒙、藏、维文
 - 智能输入技术
 - 蒙古文整词智能化输入输出系统
 - 基于蒙古语整词知识库
- 语音处理技术
 - 统一的民族语言语音声学参数数据库
 - 藏语拉萨话语音系统
 - 维吾尔语音识别的研究：建立了语音库和知识库
 - 民族语言文语转换系统

软件所民族语言处理研发

- 主要工作重点

- 基础软件平台的民文处理

- Linux操作系统与OpenOffice办公套件

- 基于开放源码研发，支持多种语言
- 支持蒙、藏、维文处理的系统采
- 用Unicode编码，扩充Linux国际化/本地化机制
- 采用OpenType和复杂文本(Complex Text)处理技术
- 对qt库和KDE用户界面进行垂直布局的扩充
- 办公套件支持Linux和Windows

- 数据库多语言支持

- 基于PostgreSQL
- 支持藏文处理的原型系统

- 多语言网站结构和面向多语言处理的浏览器引擎

- 基于多文种Linux
- 多语言网站样本：支持汉、英、藏、维、彝、朝等文字

软件所民族语言处理研发

• 主要工作重点

- 民文自然语言处理技术

• 汉-民辅助翻译系统

- 面向特定领域(法律、法规和政府公文)的汉语与多种民族语言间辅助翻译关键技术

- 基于平行语料库的智能辅助翻译技术
- 辅助翻译平台的多民族语言支持技术
- 特定领域多术语库的自动抽取技术
- 各民族语言翻译特性的支持
- 术语一致性自动检查与更正
- 翻译项目的组织与管理, 等等

- 辅助翻译平台的研发

- 跨平台的汉藏、汉蒙、汉维等语言的辅助翻译系统

• 汉语及多民族语言平行语料库及术语库建设

• 民文全文检索系统

- 应用支持英、汉、日文跨语言检索技术
- 建立了民文跨语言全文检索模型
- 初步实现支持藏、汉、英跨语言检索系统

- 可以以藏、汉、英任意一种文字的关键字检索三种语言网页

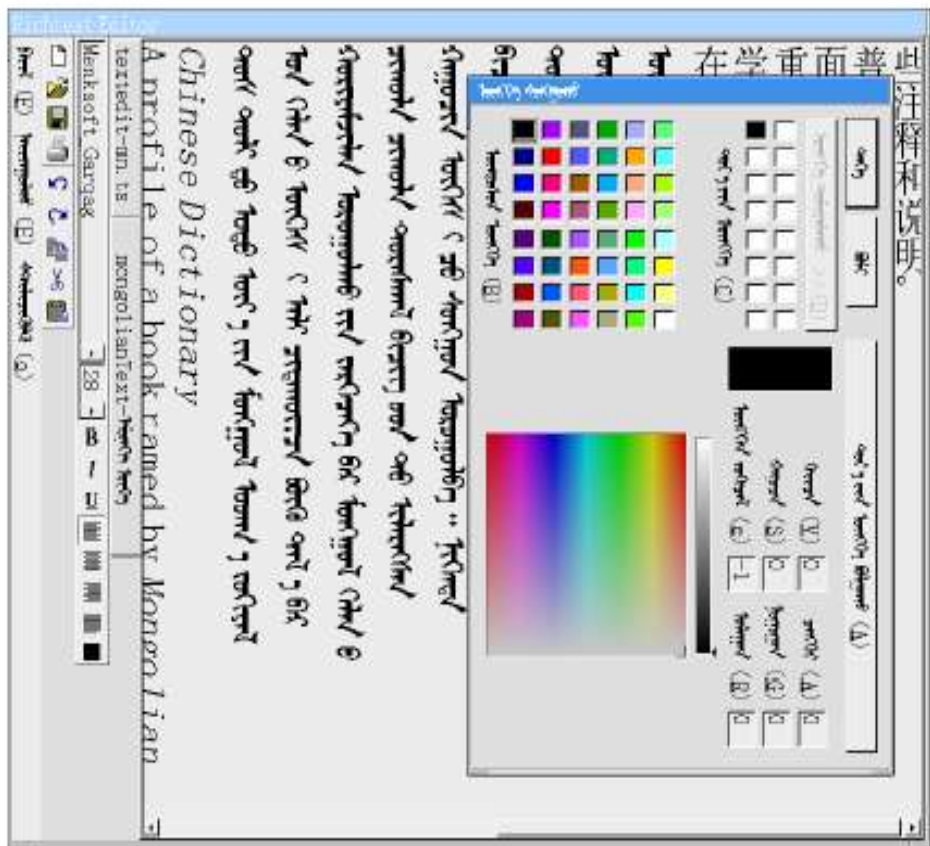
实现的技术路线

- 民族文字处理技术存在大量共通之处
 - 有大量基础性、共同性的技术问题需要解决
 - 基本系统平台多民族语言信息处理体系
 - 多语言图形用户界面
 - 复杂文档处理
 - 语言模型，等等
- 民文系统研发整体技术路线
 - 首先建立支持多民族语言信息处理体系和系统框架
 - 在统一的系统框架内实现支持不同的民族文字
 - 集中解决民文处理中基础性、共同性的技术问题。
 - 在此基础上，根据各民族文字与文化特点研发针对不同民族用户的软件系统

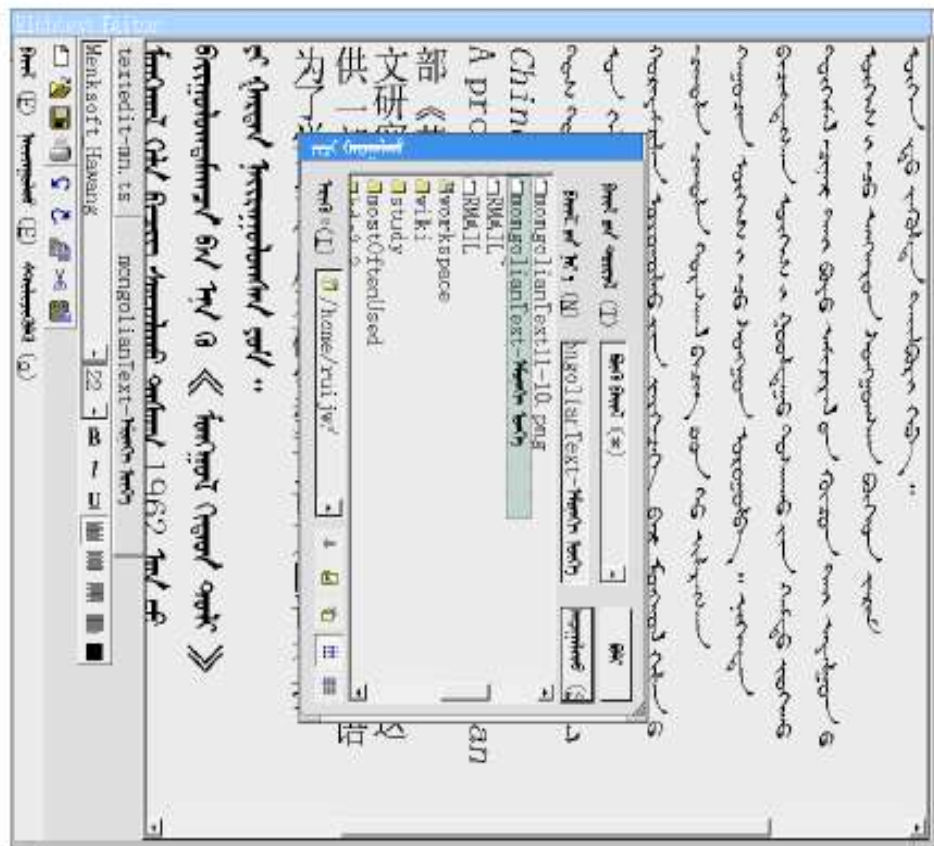
民文操作系统的设计与实现

- 在统一的系统框架下，完成民文操作系统研发
 - 基于Linux，支持传统蒙古文、藏文、维吾尔文
 - 传统蒙古文和维吾尔文
 - 采用ISO/IEC 10646/Unicode标准字符集
 - 采用OpenType字库和复杂文本处理技术
 - 实现了传统蒙古文和维吾尔文的变形显现
 - 实现了维吾尔文从右向左书写方式和用户界面
 - 实现了传统蒙古文从上向下、从左向右的书写方式和用户界面
 - 藏文
 - 支持国家标准《藏文编码字符集扩充集A》
 - 初步实现基于ISO/IEC 10646/Unicode标准基本字符集动态叠加显现
 - 采用OpenType字库和复杂文本处理技术
 - 完成了实用的传统蒙古文、藏文、维吾尔文Linux操作系统

传统蒙古文Linux垂直风格用户界面



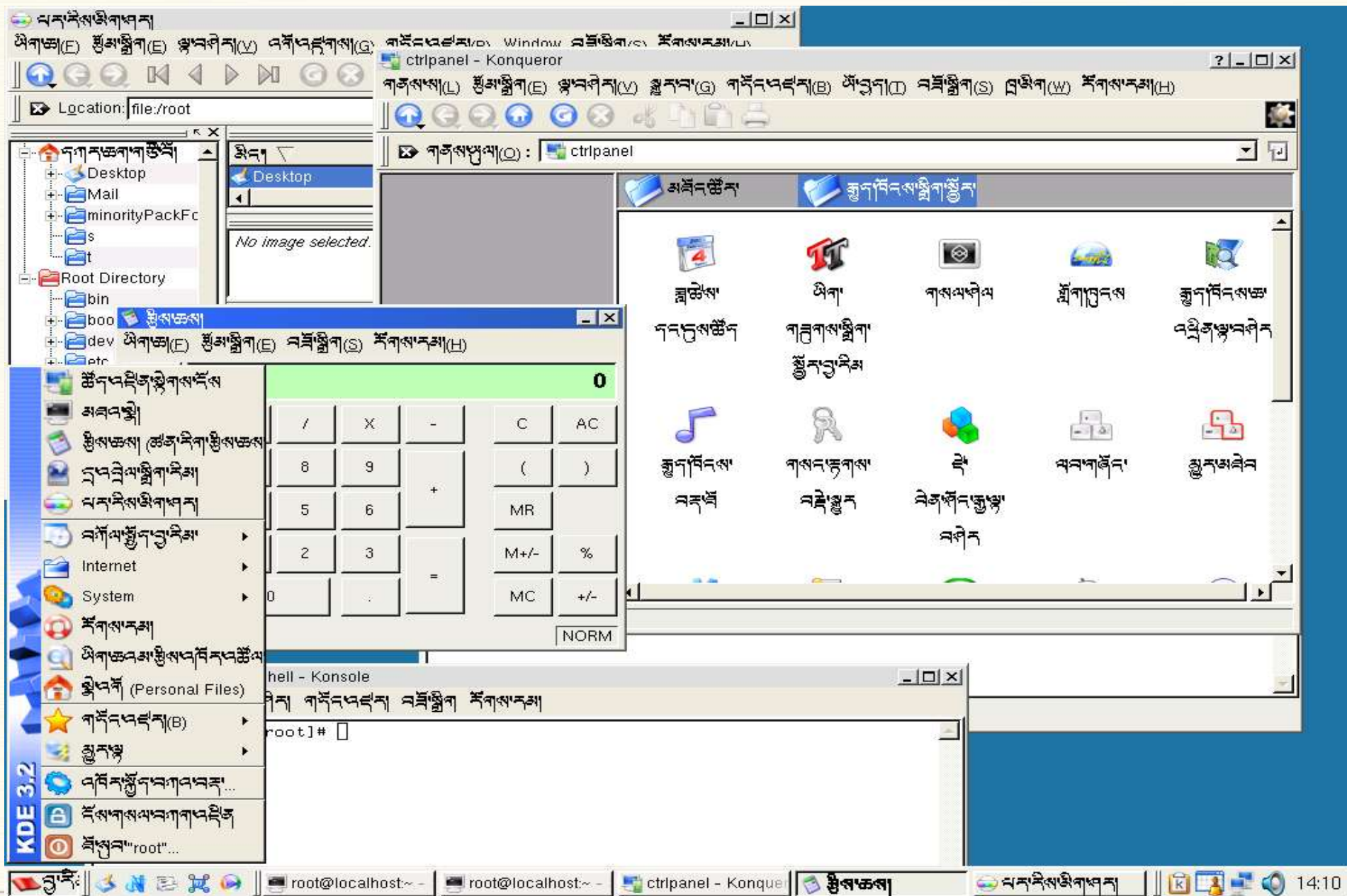
(a) 保存文件对话框



(b) 颜色编辑对话框

藏文Linux用户界面

HTTP://SONATA.ISCAS.AC.CN



多文种办公套件的设计与实现

- 本地环境(Locale)
 - 定义符合民文用户的使用习惯描述信息
 - 货币、数字、日期格式、历法、排序查找、文本的断行等
- 复杂文本的处理
 - 根据民族文字变形组合的规则解决复杂文本布局
 - 文本的显现，光标在组合字形间的定位和移动，文本尺寸的计算，文本的拷贝、粘贴、选择、删除等等
- 排版
 - 不同的民族文字在排版时也表现出不同的特性
 - 蒙古文从上向下，从左到右的书写方式
 - 维吾尔文从右到左书写，混合的数字、英文从左到右书写
 - 藏文文本断行后要用音节点“”补全一行中没有被文本填充尽的物理空间
- 界面的本地化(Localization)

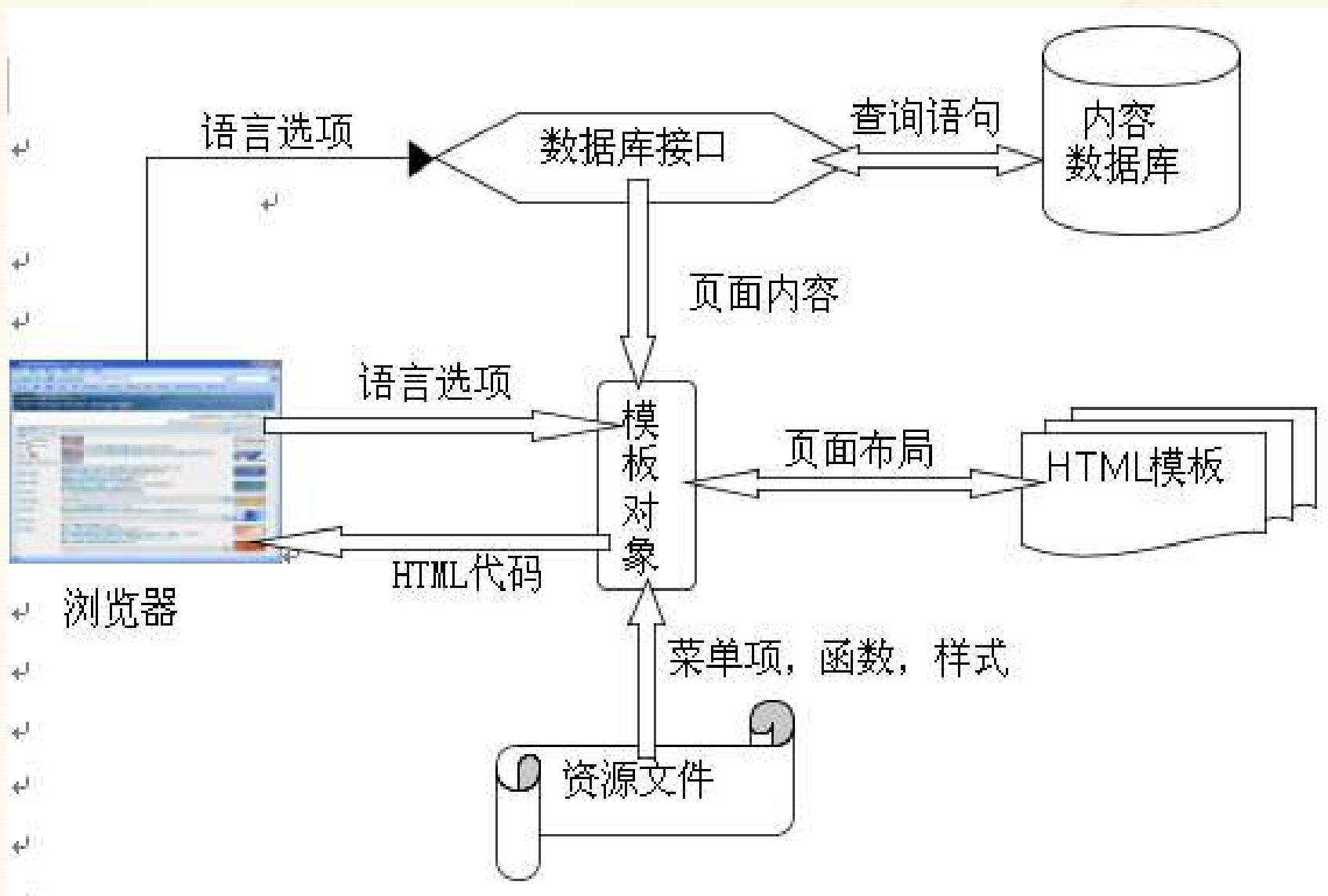
多文种办公套件的设计与实现

- 支持多种文字混合排版
 - 传统蒙古文、藏文、维吾尔文、汉文、英文
- 具有跨平台功能
 - 支持Windows和Linux
- 兼容MS Office
 - 可直接存取不同版本的Word(.doc)、Excel(.xls)、PowerPoint(.ppt)等微软Office套件生成的文档
- 具有二次开发能力
 - 可以进一步开发办公自动化系统

多文种网站信息发布技术

- 采用统一的结构，支持多种文字
 - 将格式化信息、内容与导航信息分开
 - 区域无关性内容与区域相关性的内容完全分开
 - 不同文字的页面内容相互独立地保存在数据库中
 - 显示时通过模版将页面内容与包含页面布局的HTML模版进行组合，形成完整的HTML文件
 - 添加一种新文字，只需建立相关的翻译文件
- 提供同步机制，保证多个文字版本内容的同步更新
 - 一种文字的页面发生更新，将通知其他文种的编辑人员进行相应的更新
 - 电子邮件等方式
 - 只有该网页所有文字的新版本全部完成，该页面才会被发布

多文种网站结构



西藏语言文字网藏文版

བོད་སྐད་ཡིག་ལམ་ལུ་ཤིང་། - Mozilla Firefox

PageRank: unranked Alexa 文件(F) 编辑(E) 查看(V) 转到(G) 书签(B) 工具(T) 帮助(H)

http://159.226.5.81:8080/index.php?ml=tibetan

转到

西藏藏语言文字网

Adblock

ལོ་ལྔ་བུ་ལྔ། ལམ་ལུ་ཤིང་གི་ལྷན་ཁག་། སྤོང་རྒྱུད་ཡིག་ལམ་ལུ་ཤིང་། བཀའ་བཅད་ཚད་གཞི། རིག་གཞུང་སྤེལ་རིམ། བོད་ཡིག་ཤེལ་བུ། མེ་ལུ་རོ་རྒྱུད།

ལྷོ་བོད་ཡིག་ལམ་ལུ་ཤིང་།

- བོད་རང་སྤྱོད་ལྷོ་བོད་མེ་འཕེལ་ལྷན་ཁག་གི་ཚོགས་ཚུན་གྱི་སྤྱི་བསྟུན་གསལ།
- ལྷོ་བོད་ཡིག་ལམ་ལུ་ཤིང་།**
- ལྷོ་བོད་ཡིག་ལམ་ལུ་ཤིང་གི་ཚོགས་ཚུན་གྱི་སྤྱི་བསྟུན་གསལ།
- ལྷོ་བོད་ཡིག་ལམ་ལུ་ཤིང་གི་ཚོགས་ཚུན་གྱི་མེ་འཕེལ་ལྷན་ཁག་གི་ཚོགས་ཚུན་གྱི་སྤྱི་བསྟུན་གསལ།
- བོད་རང་སྤྱོད་ལྷོ་བོད་མེ་འཕེལ་ལྷན་ཁག་གི་ཚོགས་ཚུན་གྱི་སྤྱི་བསྟུན་གསལ།

བོད་རྒྱུད་ཡིག་ལམ་ལུ་ཤིང་།

上海市语委支援西藏藏语建设“藏语言文字”网站签约仪式 75年9月

- བོད་རྒྱུད་ཡིག་ལམ་ལུ་ཤིང་གི་ཚོགས་ཚུན་གྱི་སྤྱི་བསྟུན་གསལ། | 2005-07-15 |
- ༢༠༠༥ལོའི་དབྱིན་མཁའ་རུ་མཚན་དང་བོད་ཤི་ལོ་གསལ་ཤི་དགའ་ལོའི་ཚོགས་ཚུན་གྱི་སྤྱི་བསྟུན་གསལ། | 2005-07-13 |
- ལྷོ་བོད་ཡིག་ལམ་ལུ་ཤིང་གི་ཚོགས་ཚུན་གྱི་སྤྱི་བསྟུན་གསལ། | 2005-07-11 |

བཀའ་བཅད་ཚད་གཞི།

- 《ཆ་ལྷོ་བོད་ཡིག་ལམ་ལུ་ཤིང་གི་ཚོགས་ཚུན་གྱི་སྤྱི་བསྟུན་གསལ།》

ལྷོ་བོད་ཡིག་ལམ་ལུ་ཤིང་།

中国语言文字网

国家民族事务委员会

西藏大学

TIBET 中国西藏信息中心

中国藏学网

完成

Adblock Fri: 25° C

谢谢!