广西科学院学报
Journal of Guangxi Academy of Sciences
2018,34(1):72～77
Vol. 34, No. 1 February 2018

网络优先数字出版时间:2018-02-23 DOI:10.13657/j.cnki.gxkxyxb.20180223.002
网络优先数字出版地址:http://kns.cnki.net/kcms/detail/45.1075.N.20180223.0926.004.html

蒙古语媒体资产编目问题研究*

Study of Mongolian Media Asset Construction and Cataloging Management

娜仁图雅[1],白双成[2**]
Narentuya[1], BAI Shuangcheng[2]

(1.内蒙古广播电视台媒体资产部,内蒙古呼和浩特 010058;2.内蒙古社会科学院,蒙古语信息技术研发中心,内蒙古呼和浩特 010020)

(1. Inner Mongolia Radio and Television Station, Hohhot, Inner Mongolia, 010058, China; 2. Mongolian Information Technology Center, Inner Mongolia Academy of Social Science, Hohhot, Inner Mongolia, 010020, China)

摘要:通过调查汇总内蒙古广播电视台蒙古语媒体资产编目现状,分析了现行的蒙古语媒体资产通过汉译,用汉文编目方法中存在的问题,提出直接用蒙古文编目的可能性和需要重点解决的问题。蒙古文标准编码环境逐步成熟,尤其 Windows 平台上的编辑输入和显示问题容易解决,重点在于蒙古文字符编码与字形之间的多对多复杂转换关系及录入不规范等众多原因,容易导致编目文本存在严重的拼写多样化现象和字形拼写错误。为此本文提出使用智能输入法避免错误录入、使用智能校对纠错、使用智能搜索模糊匹配等多手段结合的解决方案。随着这些技术的成熟,相信蒙古文媒体资产编目问题一定能得到解决并投入实际应用。此项工作的研究对其他蒙古语文资源建设及大数据建设和应用具有广泛的参考价值。

关键词:媒体资产 编目建库 智能输入法 校对纠错 智能搜索
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1002-7378(2018)01-0072-06

Abstract: By summarizing the status quo of Mongolian media cataloguing assets of Inner Mongolia Radio and TV Station, this paper analyzes the problems existing in the current Mongolian media assets through Chinese translation and using Chinese-language cataloguing methods, and proposes the possibility of directly cataloging with the Mongolian and the key problems that need to be solved. The Mongolian standard coding environment gradually matures, especially the editing input and display problems on the Windows platform are easy to be solved. The focus is on the many-to-many complex transformation relationship between the character encoding and the font of Mongolian, and the non-standard input conversion and other reasons, which can easily lead to serious spelling diversifications and font spelling errors in the catalog text. For this reason, the solution of combining multiple methods that use intelligent input method to avoid error input, use intelligent proofreading to error spelling correction, and use intelligent search to fuzzy search is proposed in this paper. With the maturity of these technologies, it is believed that Mongolian media asset catalog can be perfectly solved and put into practical application. The research of this work will be of great reference value for the construction of other Mongolian language resources and application of large data.

Key words: media asset, cataloging, intelligent input method, spelling correction, intelligent search

收稿日期:2017-12-20
修回日期:2018-01-07

作者简介:娜仁图雅(1976—),女,高级工程师,硕士,主要从事媒体资产管理及翻译学研究。

*内蒙古蒙古文信息化专项 2017 年度专项(MW-2017-MGY-WXXH-001)资助。

**通信作者:白双成(1974—),男,研究员,博士,主要从事语言工程研究,E-mail:BaiShuangCheng@qq.com。

0 引言

媒体资产管理(Media Assets Management,MAM)概念一般分为两层。狭义上,MAM 是指对音/视频资料、文本资料、图片资料等各种媒体内容进行全面管理,以信息化、数字化、网络化的管理理念和技术手段,将媒体资源内容进行采集、上载、制作、翻译、转换、编目、标注、检索、存储、审核、播出、发布和商业化运作衔接等,甚至可以包括对媒体资源的深度数据挖掘(Deep Data Mining)。MAM 的目的在于提高广播电视资料管理效率,对媒体资源进行永久化存储、生产化加工、资源化管理和便利化利用[1]。广义上,MAM 是指通过技术的、行政的、市场的、资本的等各种运作和手段,实现媒体资产价值的优化。价值的优化又可以细分为对内价值优化和对外价值优化。本文 MAM 主要指狭义上的概念,案例只限定于广播电视行业,但会从广义概念角度分析总结和提出解决方案。

媒体资产的基本概念由两部分组成:媒体内容和元数据(Metadata)。媒体内容是包含数字媒体(视频、音频、文本、图片等)的文件。元数据是媒体的描述信息,元数据被定义为关于数据的数据(Data about Data)。元数据是媒体内容以外的和媒体本身相关的信息,或者更为通俗地说,就是关于媒体资源的说明数据,或者叫媒体资源的标注数据。对媒体资产而言,常用的原数据包括 URL、文件名称、栏目标签、作者、翻译人员、后期制作人员、编辑人员、审核人员、制作日期、发布日期、编目日期、版权信息、版本控制、数字水印、条件接收密钥等等信息[2]。元数据在整个 MAM 系统里移动,与一个或多个媒体内容复用或嵌入。它建立在完整的数据统计和对目前业务的详细分析基础上,通过编目管理建立。所以,MAM 中编目工作是非常关键的一个环节,编目数据就是对媒体资产进行数据标注的过程,数据标注的好坏直接关系到资源数据的质量及使用效率,进而影响媒体资源管理水平的高低和节目资源的开发与利用[3]。

1 蒙古语媒体资产及管理现状

1.1 蒙古语媒体资产现状

内蒙古广播电视台是集广播、电视、网络、新媒体等多种业务为一体的省级广播电视大型综合传媒机构。内蒙古广播电视台拥有国内首屈一指的蒙古语广播电视节目生产能力和一流的节目播出能力。半个多世纪以来,通过蒙古语广播电视几代人的艰辛努力,内蒙古广播电视台制作了海量的音/视频节目,这些节目内容涵盖政治、经济、历史、科学、文化等各个领域和历史阶段,已经成了不可或缺的、不可再生的、珍贵的蒙古语声像资料和精神文化遗产[4]。

内蒙古广播电视台从建台伊始就是蒙古族文化艺术音/视频节目生产重要基地。现库存蒙古族文化艺术音频节目约 25 000 h、视频节目约 17 000 h。这些作品已经成为内蒙古自治区不可再生的文化瑰宝和后人传承弘扬的经典,也是"一带一路"倡议下文化传播中的一块宝藏。因库存节目量很大,还有 2/3 的节目没能数字化,现迫切需要加大投入,实现这些优质资源的数字化和共享,否则随着时间的推移,一旦磁带完全失效,资料将全部遗失,后果不堪设想。

另外,随着信息化、网络化时代的到来,内蒙古广播电视台也建设了蒙古文网站和客户端。从蒙古族网民和社会大众的反馈来看,他们提出了回看(听)或分享这些经典怀旧或历史内容的迫切需求。因此,使这些节目数字化和网络化,不仅可以满足广大网民的需求,也切合了内蒙古广播电视台传统媒体与新型媒体融合发展的需求,同时也有利于内蒙古自治区民族文化大区建设和民族文化遗产保护大业。以需求为导向、以应用促发展,充分利用数字化、网络化技术,开发利用符合蒙古文信息技术国家标准的数字资源,进一步科学保护和整合利用蒙古语言文字信息资源是功在千秋的事业。

1.2 蒙古语 MAM 现状

内蒙古广播电视台一直重视蒙古文信息建设,通过批量采购、场地授权等方式购入蒙古文办公软件,先后重点建设了"蒙古文文稿管理系统""传统蒙古文与斯拉夫蒙古文网站建设""蒙古文非线性编辑软件升级改造"等工程[5],目前所有编辑工作基本达到无纸化办公。

源于多种原因,蒙古语媒体资产编目目前采取汉文编目方法。具体操作是将蒙古语节目的所有编目信息都用汉文标注[6]。以文艺类节目编目为例,对歌名、作曲、作词、演唱者等所有信息都要翻译为汉文,再用汉文编目,而对于歌词等批量文本无法翻译,所以只能忽略。这种方法的好处:(1)不用额外投入研发,直接使用现有汉文 MAM 系统即可;(2)蒙古语节目和汉语节目同时使用汉文标注编目,便于统一管理。但这种以汉文标注蒙古语媒体资产的编目方法并非最佳选择。

2 蒙古语 MAM 编目存在问题

2.1 翻译质量得不到保障

目前投入蒙汉翻译和编目工作的人员都是媒体资产部蒙汉语兼通技术人员,没有专业翻译人员。从翻译学(Translation Studies)理论而言,其研究领域牵涉到与翻译有关的各个领域,包括文学翻译、非文学翻译、各种形式的口译以及电影字幕翻译。翻译学至少可以在对比语言学、比较文学和文化研究3个层面上得到研究,但是翻译学不单单隶属于上述任何一门学科门类。

蒙古语和汉语隶属于两大语系,差异较大,蒙古文(语)翻译为汉文(语)属于语际翻译(Interlingual Translation)。在蒙古语节目编目时同时要兼顾蒙古文字幕的汉译和蒙古语节目内容的边听边译,既有笔译(Translation),也有口译(Interpretation)。就翻译处理方式而言,根据实际需要,翻译人员自行判断并摘取原文的中心内容或个别章节(或段落)进行翻译,摘译的内容一般是原作的核心部分或内容梗概,是一种典型的摘译方式。摘译过程属于特别领域,翻译质量完全取决于翻译人员对字幕原文、音/视频原语的精确理解和汉文表述能力。另一方面,电视节目所涉及知识面特别广泛,如果想保证摘译质量,应该对翻译人员按其需翻译资料进行分工。翻译是一门非常严谨的学科,不是懂两种语言的人就能胜任这种专业性很强的翻译工作。如果不顾这些规律而行,只能导致一种结果:翻译质量得不到保障。

2.2 搜索利用困难

翻译的主流理论都建立在对等原理(Principle of Equivalence)的基础上,把对等、等值、等效作为终极目标。除了日常用语和个别简单规范词外,很难达到这种目的。蒙古语的一个词 W_m,翻译为汉文词 W_h,而 W_h 翻译回蒙古文不一定保持为 W_m。例如,常用词"[illegible]"翻译为汉文时最大可能是"牛粪",再翻译回来时最大可能就变成了"[illegible] [illegible] [illegible]",即使翻译人员非常仔细,汉译为"干牛粪",再蒙译时也可能是"[illegible] [illegible] [illegible] [illegible]"。

蒙古语媒体资源编目的目的和价值在于搜索利用。很显然,目前这种完全汉文标注蒙古语媒体资源的方法,除了为系统管理员进行统一管理和不懂蒙古文的领导监管时提供便利,以及对类似"天气预报"等资源种类确定、名称翻译固定的资源影响较小之外,对多数资源、多数用户和后续各种应用都将带来诸多不便。

搜索利用蒙古语节目的用户可分为两大类:(1)内部用户,如编辑人员参考、应用历史资料,他们希望能快速、准确定位所需资源;(2)外部用户,如听众搜索欣赏感兴趣的节目。不管哪一种利用,其受众肯定是懂蒙语的人,对他们来说,原本要查找蒙古文资源,现在需要汉译关键字才能搜索,只有当查找的人和编目的人翻译相同时才有可能找到,而这种可能性太小[7]。还是以文艺节目为例,"美丽的草原我的家""十五的月亮""嘎达梅林""雕花的马鞍""鸿雁"等脍炙人口的、蒙汉双语都很流行的歌曲,不管是对编目人员还是搜索应用人员都还可能容易一一对应。但对于一些蒙古语歌曲,例如新疆蒙古族民歌"[illegible] [illegible]",经歌手斯琴格日乐改编后被人们所熟知,在专辑"我自己"中将其翻译为"两只小山羊",而就字义而言,很难将两者对等起来。所以,准备搜索这首歌曲的用户除非知道这首歌曲的这种汉文翻译名称,否则基本无法利用编目信息,这是一种极大的浪费和遗憾。更何况,还有很多东西只能音译或意译,例如,波如(茹)来(民歌)、阿萨(斯)尔(宫廷乐)、乌云山(珊)丹等,到底用哪个汉字,完全取决于译者喜好。

2.3 后期应用面临二次标注

模拟信号的媒体资源转换为数字化资源保存并编目,无非是要增加其价值,而增加价值最直接的方法是将历史珍贵资源放在云平台,通过网页、APP等形式让广大用户利用[4]。对于蒙古语媒体资源,虽然还有汉文及其他文种使用需求,但毕竟蒙古文使用者是最主要的用户群。以蒙古文音乐 APP 为例,至少需要歌曲名称、歌手、作词、作曲等信息,很显然这种汉译形式的编目资源无助于媒体资源的利用开发,完全面临二次标注和编目,是一种极大的浪费。

3 解决方案

鉴于以上分析,蒙古语媒体资产应主要用蒙古文编目,而名称等少量信息用汉文编目即可。在现有汉文 MAM 上支持蒙古文,需要解决蒙古文的显示、编辑、搜索3大问题。由于蒙古文标准编码环境逐步成熟,显示和编辑问题基本没有技术障碍。尤其对于 MAM 编目来说,工作环境相对封闭单一、投入工作人员有限、服务器和终端设备稳定不变。相对于通用软件来说,MAM 技术实现时只需考虑现有服务器、终端设备和操作系统即可,其技术实现

更具稳定性。

由于蒙古文字符编码与字形之间的多对多复杂转换关系及录入不规范等众多原因,容易导致编目文本存在严重的拼写多样化现象和字形拼写错误[8]。对此,主张使用智能输入法提前避免错别字录入、使用智能校对后处理,纠正拼写错误、使用智能搜索实现模糊搜索等多手段结合的解决方案。

3.1 使用智能输入法避免误录

推广普及录入规范,尤其是培训编目人员尽快改变不良录入习惯,坚决按照蒙古文正字法的地方法规(目前可遵循的规范有《蒙古文正字法词典》(第二版)[9]和信息处理国家标准——《信息技术 传统蒙古文单词词形规范 基本集》(第二版)[10]等)的规定进行正确拼写。单从编目角度考虑,要想让长期习惯于只顾及字形录入的工作人员按规范录入,具有一定难度。兼顾字形和读音理解会额外耗费一些精力,这明显有悖于其绩效考核。那么如何使他们的付出受益(不限于经济利益),让他们意识到这样做的好处并愿意付出这份努力也许是个不简单的系统工程问题。将媒体资源作为商品授权给研究机构及商业公司,也许能起到一点辅助作用,但也不是长效机制。

在标准编码实现中,一般都会附带一个键盘映射(Keyboard Mapping)形式的输入法。因字体中实现了字符映射关系,输入法本身一般都较为简单,不做过多处理。因蒙古文是个同形异音字符较多的文字,与标准编码配合的这种键盘映射输入法无法避免同形字符输入错误,从字形又不易察觉到错误,为检索、统计等后续处理埋下隐患。综上所述,推出智能化程度较高的输入法尽量避免和预防错误输入是一种有效途径。鼓励使用完全符合规范和标准的智能输入法,从录入源头避免错误。智能化输入法确保录入字形和读音正确的同时给用户简单易用的体验,让用户不再感觉遵循规范是个负担。此处所述输入法不局限于全键盘录入,也包括智能终端的虚拟全键盘、数字键盘及光学文字识别(OCR)识别录入、语音识别录入等所有输入方式。这些输入方式中必须加入监督机制,尽量避免用户录入错误。

3.2 使用校对纠错工具后纠正

目前"词典+规则"是实现蒙古文文本校对的常用方法[11]。不管是使用不确定有限状态自动机(NFSA)数据结构获取较高计算效率、使用词干词缀和生成规则来节省存储空间,还是使用最一般的字符串匹配的库结构,其本质无非都是依赖词库,词库中有的词认为是正确词,词库中没有的词(OOV)就认为是错别词。校对和纠错效果基本取决于词汇量。再进一步用搭配库[12]或规则对部分同形异音词(例如  的录入为的)进行甄别。从公开资料来看,未登录词、同形多音词处理还不够成熟,句法和语义层面错误基本未能投入应用。字形拼写错误的纠错(例如  一词多一个字牙或少一个字牙时,除提示拼写错误外还应提供正确拼写建议)是个需要不断完善提高的研究课题[13],所以无法单纯依赖校对和纠错。

3.3 使用搜索引擎实现模糊搜索

因为蒙古文编码的特殊性,通用搜索引擎对蒙古文搜索支持一直不太理想。即使到目前为止,baidu.com 刚刚将蒙古文编码范围(18 区)纳入可搜索字符范围(之前干脆不认),仅支持单个字符为单位的内码匹配搜索(将蒙古文字符当成互不关联的独立字符),bing.com 虽说情况稍微好些,但还是没有针对语言的支持,更无法解决拼写形式多样化问题,只有拼写完全一样的才能按内码匹配搜索到,无法满足日常应用所需。如图 1 所示,蒙科立搜索引擎(http://hai.menksoft.com)是目前为止可找到的、唯一一个在线服务的蒙古文搜索引擎,该引擎有针对性地爬取蒙古文网站并兼容解决了蒙科立编码和标准编码,可以按蒙古文字形模糊搜索。这一系统从侧面证明了编目中蒙古文搜索完全可以解决。

图 1 百度、Bing 及蒙古文搜索引擎搜索蒙古文

Fig. 1 Search Mongolian in Baidu, Bing and Mongolian search engine

4 结束语

本文主要分析内蒙古广播电视台目前在蒙古语媒体资源管理中采用的蒙译汉，用汉文进行编目方式中存在的问题，提出了使用智能输入法预防错误录入、使用校对系统纠错、使用搜索引擎实现模糊录入等 3 种策略相结合的解决方法。这一方法的好处在于，在现有汉文 MAM 基础上，尽量不对其进行重新架构的前提下，用最低廉的工程成本实现蒙古文编目目的。很明显，这 3 种方法单独使用，都有不足之处，最好能综合利用。就目前研究状况而言，现有部分产品已经满足了以上要求，希望这些基础研发和产品能更加完善，进一步减轻工作人员的负担，提高工作效率。这 3 项研究和产品恰恰是蒙古文信息化中最基本和最核心的应用之一，各项工作的开展势必依赖语料库建设、知识库建设及相应大数据、机器学习等方面的突破[14]。相信在相关专家和企业努力下，这些工作必将取得进一步突破。

参考文献：

[1] 罗蕴军，黄瑞卿. 编目子系统在媒体资产管理系统中的应用[J]. 数字通信世界，2013(9)：74-77.
LUO Y J，HUANG R Q. Application of cataloging subsystem in the media assets management system [J]. Digital Communication World，2013(9)：74-77.

[2] 龚亦炜. 音像资料编目在数字化媒体资产管理中的地位与前景[J]. 西部广播电视，2015(4)：69-70，80.
GONG Y W. The position and prospect of audio and video data cataloging in the management of digital media assets[J]. West China Broadcasting TV，2015(4)：69-70，80.

[3] 徐少勇. 宁波电视台媒资编目使用数据浅析[J]. 新媒体研究，2016，2(15)：62，165.
XU S Y. Data analysis of Ningbo TV MAM cataloguing use[J]. New Media Research，2016，2(15)：62，165.

[4] 何锋，高晓华，李凤英. 内蒙古电视台媒体资产建设编目管理与实践[J]. 现代电视技术，2014(11)：113-117.
HE F，GAO X H，LI F Y. Media asset construction，cataloging management and practice of Inner Mongolia

TV station[J]. Advanced Televition Engineering,2014(11):113-117.

[5] 娜仁图雅,白振东. 浅析大洋 3000 非编软件蒙古文字幕系统[J]. 内蒙古广播与电视技术,2008,25(1):39-40.
NARENTUYA,BAI Z D. Analysis of Mongolian subtitle system of Dayang 3000 non-linear editing software [J]. Inner Mongolia Radio & TV Broadcast Engineering,2008,25(1):39-40.

[6] 李澎涛,王力栋. 内蒙古广播电视台媒资编目生产管理与实践[J]. 内蒙古广播与电视技术,2014,31(3):20-21,17.
LI P T,WANG L D. Media cataloging management and practice of Inner Mongolia TV station[J]. Inner Mongolia Radio & TV Broadcast Engineering,2014,31(3):20-21,17.

[7] 曼宁,拉哈万,舒策. 信息检索导论[M]. 王斌译. 北京:人民邮电出版社,2010.
MANNING C D,RAGHAVAN P,SCHÜTZE H. Introduction to information retrieval[M]. WANG B (ed). Beijing:Post & Telecom Press,2010.

[8] 白双成. 蒙古文原始语料统计建模研究[J]. 中文信息学报,2017,31(1):118-125.
BAI S C. Study of Mongolian raw text modeling[J]. Journal of Chinese Information Processing, 2017, 31(1):118-125.

[9] 《蒙古文正字法词典》编委会. 蒙古文正字法词典[M]. 修订版. 呼和浩特:内蒙古人民出版社,2012.
Editorial Committee of "Mongolian orthography dictionary". Mongolian orthography dictionar[M]. Revised edition. Hohhot:Inner Mongolia Peoples Publishing House,2012.

[10] 六十三,金湖,浩特劳,等. 信息技术 传统蒙古文单词词形规范 基本集:GB/T 32912—2016[S]. 北京:中国标准出版社,2017.
LIU S S,JIN H,HAO T L,et al. Information technology—Specification of word form of traditional Mongolian words—Basic set:GB/T 32912—2016[S]. Beijing:China Standard Press,2017.

[11] 斯·劳格劳. 基于不确定有限自动机的蒙古文校对算法[J]. 中文信息学报,2009,23(6):110-115.
S·LOGLO. A proofreading algorithm of Mongolian text based on nodeterministic finite automata[J]. Journal of Chinese Information Processing, 2009, 23(6):110-115.

[12] 斯琴. 蒙古语普通名词词语搭配研究[D]. 呼和浩特:内蒙古大学,2009.
SI Q. A study of the collocation of Mongolian general nouns[D]. Hohhot:Inner Mongolia University,2009.

[13] 确精扎布. 确精扎布蒙古文信息处理专辑[M]. 呼和浩特:内蒙古教育出版社,2014.
QUEJINJAB. QueJinjab's Mongolian information processing album[M]. Hohhot:Inner Mongolia Education Press,2014.

[14] 龚明,高晨. 媒资管理系统中自动编目实现的研究[J]. 电视工程,2016(3):11-13.
GONG M,GAO C. Research on automatic cataloging in media management system[J]. Television Engineering,2016(3):11-13.

(责任编辑:陆 雁)