

略论中华寻根网文献资源构架

王昭

对数字资源系统而言，资源构架是资源的总体设计，即资源系统中包含哪些资源、资源的类型、粒度、各类资源间的关系以及资源生产、资源积累、资源消耗、资源更新等。中华寻根网作为以家谱文献资源为核心的数字资源系统，其文献资源构架模拟一个文献解构的过程，将家谱文献分解为卷、篇、叶，再分解为语义块、信息点、字组、单个文字。通过上述解构过程，产生不同粒度的数据，提供不同的检索、显示等应用方式。换言之，读者使用家谱文献时，大多数情况下也只对家谱中的一部分内容感兴趣，而真正有用的信息可能更少。家谱数字资源能够提供多种类型的检索和显示方式，提高家谱文献使用的效率。

1. 中华寻根网概述

为推进中华谱牒文化的发展，继承、弘扬传统文化，促进中华民族的团结，国家图书馆与澳门基金会合作建设“中华寻根网”（以下简称“寻根网”）。寻根网的项目策划始于2008年，经过项目设计、资源规划、软件平台设计、总体设计论证、专家论证等，2009年底完成原型系统，通过专家评审后，开始系统内测和大规模数字资源建设。2010年10月，寻根网开始线上测试，并于2011年3月5日正式开通上线，提供服务。

1.1 项目宗旨

寻根网项目是建立在广泛合作基础上的全球家谱数字化服务、教育和研究项目，以保存人类文明的共同记忆为最终目标[1]。作为一个开放型项目，希望各国各地的图书馆、档案馆、学术机构、宗乡会、家谱编委会等组织和个人加入，使之成为全球华人寻根问祖的家园，团结全球华人的文化平台，宣扬中华民族悠久历史文化的窗口，提供珍贵家谱馆藏的服务通道，专业家谱信息咨询的必备参考，家族文化寻根的规划助手。通过五到十年的努力，最终将寻根网建成一个全球化的华人寻根中心、家谱资料中心、家谱服务中心和家谱研究中心。

1.2 项目现状

寻根网项目建设已经取得了阶段性成果，建立了寻根网的系统模型、数据模型和软件系统。系统建设分解为八个模块，包括家谱目录系统、家谱全文系统、寻根百科系统、寻根导航系统、用户咨询系统、用户互动系统和后台管理系统。每个模块能够实现一个或多个功能，各个模块既相对独立，又相互联系，共同构成寻根网。

软件系统的总体架构设计采用了 MVC 模式与 J2EE 相结合的 STRUCTS 框架，辅之以缓存系统、数据持久层、AJAX 页面交互、VELOCITY 模板生成等技术框架。系统前台的业务展示功能组件分为门户引擎和扩展接口，后台的应用支撑组件和系统管理组件分为安全认证体系、平台监控及日志管理、资源体系管理、系统管理、社区空间管理和应用组件引擎。这样的系统架构是一个多层次、分布式的应用模型，具有强大的伸缩性、开放性和安全性。

系统检索模块支持每秒并发 2000 以上的访问能力，响应时间平均少于 3 秒，最长不超过 5 秒；对象文件调阅和全文数据版式还原模块支持每秒并发 1000 以上的访问能力，响应时间平均少于 3 秒，最长不超过 6 秒；数据分析模块支持每秒并发 500 以上的访问能力，响应时间平均少于 5 秒，最长不超过 8 秒。

目前，寻根网已上线姓氏数据 500 余条、家谱书目数据 30000 余条，家谱图像数据 2300 余种约 250 万叶家谱的扫描，同时发布与家谱相关的文献 6000 余种约 300 万叶[2]。

2. 寻根网文献资源构架

寻根网文献资源构架由目录、索引、图像、文本等构成，目录数据用于描述家谱文献的外部特征，并对文献内容进行整体性的揭示。索引数据用于信息点的揭示，提高图像数据的可用性。图像数据用于反映家谱文献的原貌，配合目录和索引数据使用。文本数据用于提供全文检索。

2.1 目录

目录数据作为元数据的重要组成部分，应用最为广泛，发展也比较成熟。国内早期的目录数据主要采用 CNMARC 格式（China Machine—Readable Catalogue，中国机读目录），强调描述文献的外部特征，格式也较为复杂。以《新安苏氏族

谱》为例，该书的 MARC 格式数据是：

```

001      312001068895
005      20121123094702.0
010      $b 綫裝
100      $a 20021005f17361795km y0chiy50      ea
1010     $a chi
102      $a CN
106      $a z
2001     $a 新安蘇氏族譜
          $9 xin an su shi zu pu
          $b 普通古籍
          $e 十五卷
          $f (明)蘇大纂
205      $a 活字本
          $b 木活字
210      $c 忠孝堂
          $d 清乾隆間[1736—1795]
215      $a 2 冊
          $c 圖及像
300      $a 版心題蘇氏宗譜
305      $a 9 行 23 字白口四周單邊單魚尾
6100     $a 家譜
6100     $a 蘇氏
6100     $a 安徽休寧
696      $a 傳 778.09
          $2 pgl
701 0    $a 苏大
          $9 su da
          $c 明
          $4 纂
801 0    $a CN
          $b NLC
          $c 19990914
905      $a NLC
          $b 96158
          $s JP3060
          $q FGPG

```

2005 年底发布的《家谱元数据规范》(CDLS—S05—015) [3]和赵亮、苏品红主编的《国家图书馆家谱元数据规范与著录规则》[4]都是由上海图书馆主持完成，基于 DC (Dublin Core, 都柏林核心) 元数据格式，包括核心元素 14 个 (题名、主要责任者、主题及关键词、附注、出版者、其他责任者、日期、资源

类型、格式、资源标识、来源、语种、相关资源、权限管理)、古文献核心元素 4 个(版本类别、载体形态、收藏历史、馆藏信息)和个别元素 3 个(先祖及重要族人、谱籍、堂号)。与 MARC 格式相比,《家谱元数据规范》做了一定程度的简化,但是还是留有 MARC 格式的影子。

寻根网的目录数据格式强调简单实用,在目录、索引、图像、文本的数据环境中,只著录必要的信息点,同时考虑家谱文献的特殊性。仍以《新安苏氏族谱》为例:

表 1 寻根网目录数据样例

题名	新安蘇氏族譜
卷数	十五卷
修撰者	(明)蘇大纂
版本年	明成化三年刻清乾隆元年重印
版本	刻本
谱籍地	安徽休寧
堂号	忠孝堂
始祖	(唐)蘇頌
始迁祖	(宋)蘇壽
收藏地	國圖

2.2 索引

索引数据配合图像数据使用,揭示家谱文献中的信息点。寻根网的索引数据分为人物索引和篇目索引两类。人物索引标引谱系中的每个人物,篇目索引标引家谱中的序跋、凡例、源流、家规家训、派字、像赞、祠墓图、世系(支派)、传记、墓志、碑铭、艺文等。例如:

```
<?xml version="1.0" encoding="UTF—16" standalone="true"?>
<tag source_id="SZJP00005—10002410—00001.jpg">
  <persontag>
    <person sex="男" surname="吴" region="2694,633,2940,1788" id="1">
      <firstname>八府君</firstname>
      <hui>疎</hui>
      <zi/>
```

```

    <hao>敬文</hao>
    <alias/>
    <shi/>
    <birthday/>
    <death_date/>
    <shidai>四代</shidai>
    <branch>堂陰世系圖</branch>
  </person>
</persontag>
<titletag>
  <title region="3459,793,3585,1125" id="1">
    <title>殷太公行略<reverse variant="1" type="reverse"/></title>
    <author>殷天正</author>
    <dynasty>清</dynasty>
    <keyword>傳記</keyword>
    <keyword2>殷太公 曹氏</keyword2>
  </title>
</titletag>
</tag>

```

人物索引的范围包括家谱谱系图、谱系图表、谱系描述文字中的人物，人物索引包括人物编号、在图像中的位置、姓、名、讳、字、号、别称、谥号、性别、生、卒、世代、支派等字段。按照支派、世代的先后顺序，同一支派的同一世代按照从右向左的顺序依次索引。篇目索引的范围包括家谱中的全部篇目、标题、插图等，篇目索引包括篇目编号、在图像中的位置、题名、著者、著者朝代、关键词等字段。按照篇目标目出现的顺序依次索引。索引数据要求完整、准确，文字错误率低于 0.3‰，支持 unicode5.0 的字符集，支持检索和显示。

2.3 图像

目录数据和索引数据属于元数据，而图像数据和文本数据属于对象数据。图像数据能够直接反映家谱文献的原貌，也是文献数字化的基本形式之一。

寻根网图像数据采用彩色扫描方式，扫描分辨率为 300ppi，颜色模式为 RGB 模式，位深为 24 位真彩色，存储为 TIFF、JP2 和 PDF 格式，压缩因子为 150。要求扫描图像真实反映原书，以原书的上边沿为基准，以中缝折线为中心线，保持原书的天头、地脚的尺寸不变，左右两边的尺寸基本不变，扫描后的影像文件要叶码连续，没有重叶、缺叶，错叶、折叶等情况（原书缺叶、错叶除外）。补扫缺叶图像要与正本图像大小一致，颜色接近。原书有大面积的残缺时，要酌情

添加衬纸（与原书的颜色接近），避免透光、透字。扫描后进行必要的去图像黑边、图像拼接等处理，图像拼接后不得有明显的拼接痕迹。图像偏斜不超过 1° ，扫描后每叶影像尺寸大小与原书一致，误差小于 1‰。

2.4 文本

文本数据包括家谱文献中的文字信息和必要的版式信息。谱系是家谱文献的主要内容,通常采用谱系图、谱系表、描述文字等形式。如果将谱系部分直接转换成文本数据,只能借助全文检索来查找人物信息,却不能体现家族内部人物之间的关系。因此,寻根网对家谱文献的非谱系部分采用全文版式还原方式进行数字化,与其他文献相同;谱系部分(包括谱系图、谱系表等)都按照“谱系树”方式数字化。

谱系树是将每个人物作为一个节点,每个节点包含该人物的主要信息和父子、兄弟等亲属关系,同时将像赞、荣恩、家传、艺文等与该人物进行关联,依据每个节点的直系亲属关系可以将各个节点串联起来,形成树状拓扑结构[5]。“谱系树”中的每一个人物具有唯一的 ID 号,以结构化数据的方式记录每一个人物的姓、名、字、号等信息和父子、兄弟等关系。如图 1 所示。

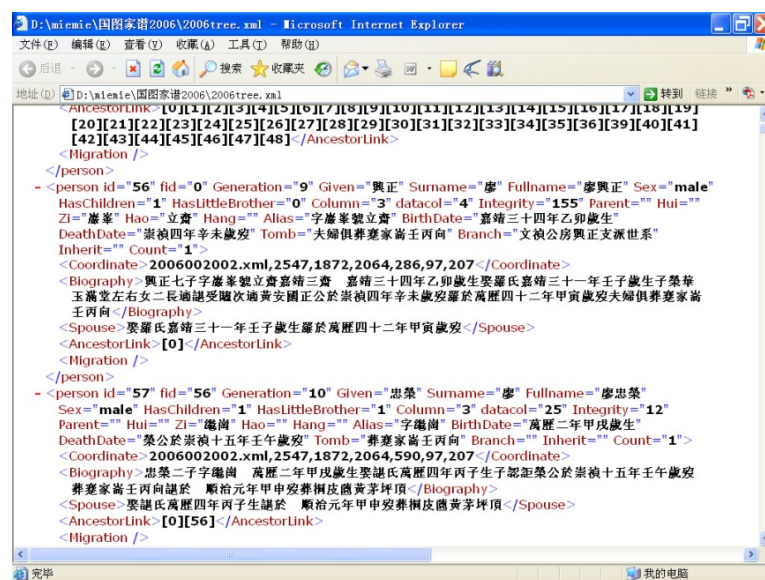


图 1 谱系树样例

以“谱系树”为基础,系统可以生成比较规整的谱系图,用文字来描述人物的信息,用线段来表示人物之间的关系,如图 2 所示。同时依据用户的需求,可以采用多种方式显示“谱系树”。

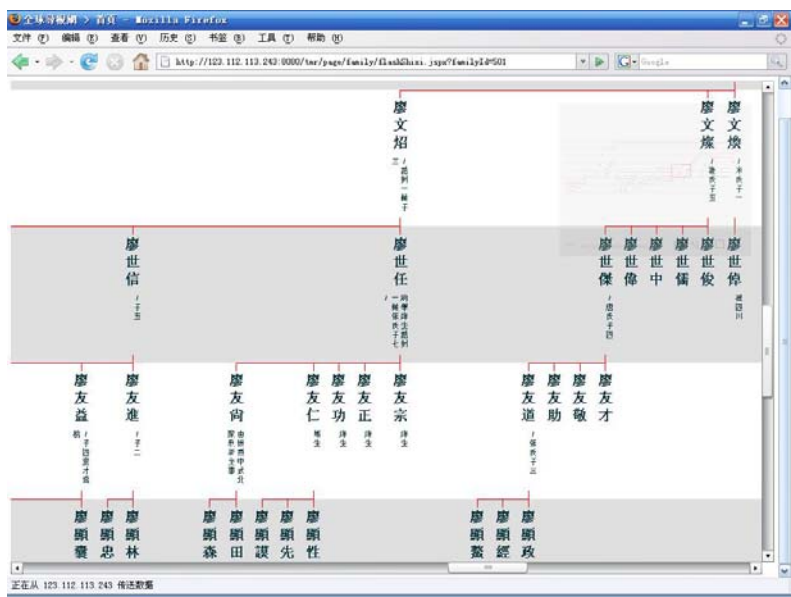


图 2 谱系树显示效果

3.小结

寻根网的后续建设计划包括馆藏家谱数字化（第二期）、网站导航库、家族名人库、家族著述库等，还将提供开放目录、在线捐赠、家谱离线编辑工具等。同时，我们也将尝试引入 HGIS（History Geographic Information System，历史地理信息系统）。无论是数据量增加或数据种类增加，还是系统功能扩展，文献资源构架都是寻根网的核心与基础，保持相对的稳定性。

参考文献

[1] 关于寻根网 [OL]. [2015—11—2]. <http://ouroots.nlc.cn/about.jsp>.
[2] 中华寻根网 [OL]. [2015—11—2]. <http://ouroots.nlc.cn/index.jsp>.
[3] 赵亮，苏品红. 国家图书馆家谱元数据规范与著录规则 [M]. 北京：国家图书馆出版社，2014.
[4] 家谱元数据规范 [OL]. [2013—9—24]. <http://ir.las.ac.cn/bitstream/12502/4878/1/CDLS—S05—015%e5%ae%b6%e8%b0%b1%e5%85%83%e6%95%b0%e6%8d%ae%e8%a7%84%e8%8c%83.pdf>.
[5] 肖禹. Lib2.0 环境下谱牒的收集与整理——以全球中华寻根网项目为例 [J]. 科技情报开发与经济，2010(29): 35—39.