

社会探险家：美国人口统计制图数字图书馆

□ 刘燕权 / 美国南康涅狄格州立大学 纽黑文市 06515

马凌云 / 上海师范大学图书馆 上海 200234

王群 / 北京师范大学信息管理系 北京 100875

摘要：作为美国国家科学基金（NSF）资助的数字图书馆项目之一，社会探险家是一个涵盖美国1790年至今的人口普查数据、拥有上万张主题地图和数百份报告的交互式制图数字图书馆。文章对该数字图书馆的建设、意义及现状进行了综合性评价，包括项目概述、资源组织、技术特征、服务特征以及作者的评估与建议。

关键词：人口统计，人口普查，社会探险家，数字图书馆，美国国家科学基金会（NSF）

DOI: 10.3772/j.issn.1673-2286.2010.06.008

1 概述

社会探险家（Social Explorer）是美国一个关于人口统计的制图数字图书馆。它由美国国家科学基金会（NSF）的SMETE数字图书馆资助，自2002年9月1日起，到2007年3月31日为止，受助金额为389,027美元。该项目英文名称为Collaborative Research: A Digital Library Collection for Visual Exploring United States Demographic and Social Change，其目的是提供“基于网络的可以描绘和探索美国自建国以来变化与发展的馆藏资料”。这个项目由纽约城市大学、皇后学院和加州大学洛杉矶分校共同完成。项目负责人是安德鲁·贝佛里奇（Andrew A. Beveridge）教授，他是皇后学院的社会学教授，致力于人口统计学的数字化研究，是利用地理信息系统（GIS）技术整合人口统计资料的专家。社会探险家包含1790年至2000年之间的美国人口普查数据以及对未来十年的预测参考，通过直观的交互式的电子地图和数据报告的形式，为用户提供多种地理层面的人口统计信息。它是目前世界上资料最翔实、形式最有趣的人口统计网站，也是最繁忙的美国人口统计地图网站。网址为<http://www.socialexplorer.com>，如图1所示。

社会探险家的赞助者有《纽约时报》（The New York Times）、美国国家科学基金会（National Science Foundation）、资助纽约市立大学的福特基金会、纽约市立大学软件与信息学院、纽约市立大学的计算机与



图1 社会探险家的主页

信息服务、皇后大学社会学系等。其中《纽约时报》与国家科学基金会是主要的赞助者。

（1）《纽约时报》。创刊于1851年的《纽约时报》在新闻报道方面将自己看作是一份“报纸纪录”，它的报道可靠性非常高。自1993年，该报是社会探险家数据的来源之一，它的一些记者在人口普查协作项目中与社会探险家有着长期的合作。此外贝佛里奇教授经常在《纽约时报》上发表文章，并不断致力于获得其资助。虽然我们不知道《纽约时报》具体提供了多少资助，也无法了解其资助的形式是捐赠还是投资，但鉴于社会探险家是一个独立网站，所以有一种推测是，《纽约时报》可能拥有其一定的份额。

（2）美国国家科学基金。NSF是社会探险家得以研制与开发的主要投资方，由其本科教育部门负

责, 主要通过一系列项目的形式提供资金方面的资助。自1999年开始相关的项目有三个: 第一个项目, 映射与探索纽约市变化 (Mapping and Exploring New York City Change, 1905-2000: A Set of Interactive Web Based Tools)。该项目由皇后学院发起, 得到美国国家科学基金 (NSF) 的“课程、课目体系和实验室的改进 (Course, Curriculum, and Laboratory Improvement/CCLI) ——教育资源发展部”78,960美元的资助, 自1999年7月至2000年12月。其目标是为学生或其他人开发一套基于网络的工具, 便于查阅与理解自20世纪初以来纽约市的人口变化情况。使用此工具可以了解纽约特定区域, 在一段时间内特定人群的变化情况, 尤其是某区域种族、民族、家族、家庭状况、房屋以及收入的变化等。第二个项目, 可视化和探索美国城市 and 乡村的社会变革 (Visualizing and Exploring United States Urban and Rural Social Change, 1790-2000—Interactive Multimedia and Web Based Tools)。该项目同样由皇后学院发起, 由皇后学院和加州大学洛杉矶分校共同完成, 也得到了NSF的CCLI教育资源发展部213,000美元资助, 自2001年5月至2004年4月。它在前一个项目开发的“基于网络的可视化和有效地图软件工具”的基础上分析洛杉矶的变化与发展, 目的是使学生能够查阅和理解这两个大都市 (洛杉矶与纽约) 的人口变化情况。第三个是数字图书馆项目, 就是将前面提到的可视化探索美国人口与社会变化的数字化资源组成一个图书馆 (A Digital Library Collection for Visual Exploring United States Demographic and Social Change)。与前两个项目相比, 它的用户范围更广, 数据覆盖面也更大, 历史时间跨度更长。它为学生、图书馆、媒体以及其他感兴趣的读者人群提供直观形象的1790年至2000年的美国人口普查信息情况。

社会探险家首次把人口统计的各项数据放在一个网站上, 并且允许用户根据自己的选择创建新的数字地图。对于不同年龄组的学生和研究人员来说, 社会探险家是一个非常宝贵的资源库, 它是一个视觉媒介, 用直观易懂的方式向用户展示人口调查数据, 它不仅是有用的研究工具, 也是让学生学习了解地理信息系统的工具。作为一个新事物, 社会对社会探险家数字图书馆的好评表明它在社会学领域产生了重要影响。

2 数字资源及其组织

社会探险家的“馆藏资源”来源于美国人口普查局 (The Census Bureau) 公布的所有数据。从1790年第一次人口普查开始, 美国每十年推出一个全面的人口统计报告。为满足不同的研究需求, 人口普查局提供的数据是基于不同地理层面的, 比如国家、国会地区、州等人们熟知的层面, 也包括普查地段 (census tracts)、街区组 (block groups) 等人们不怎么了解的层面 (图2所示), 数据的内容也很广泛, 包括年龄、性别、种族、教育、收入、职业、房产等等。社会探险家将这些数据变得可视化, 运用交互式的主题地图, 通过颜色的不同深浅度呈现符合用户查询条件的某类人群的疏密程度。作为一个商业机构, 社会探险家提供两个版本的产品, 即免费的公众版与收费的捐赠版。目前他们的网站提供了1790-2000年的人口普查数据, 其中1940-2000年的普通数据是免费的, 而其他则是收费的。收费信息更加丰富、完整、全面, 信息量大约五倍于免费信息。



图2 人口普查中不同层次之间的关系

社会探险家为用户提供两种形式的可用资源: 地图和报告。公众版地图包括交互式人口普查地图 (Interactive Demography Maps) 和时间系列地图 (Time Series Maps)。前者包括1940-2000年的人口普查地图和1980-2000年的宗教地图。后者包含了1910-2000年的纽约种族地图, 1940-2000洛杉矶种族地图和皇后学院的纽约邻里地图 (neighborhood map)。报告包含了人口普查局的历史的与现代化的数据, 这一工具允许用户在已经有了明确感兴趣的数据时, 绕过地图工具而直接生成报告。其中只有2000年的报告是全部免费使用的, 1940-1990年的报告是有限使用, 1870-1930年的报告需要收费。这些报告未经过编目, 只是以年代为

序罗列出来, 因此只能浏览式查找, 没有提供搜索功能。1980年以前的报告仅有对历史文献扫描的PDF格式, 目前已有计划要将这些信息全部数字化, 成为机读格式, 比如xml格式, 但是这个计划尚未提上日程。1980年以后的报告既有PDF格式的, 也可以链接到人口普查局网站查看机读格式的数据。以现有的地图与报告为基础, 社会探险家可以根据用户对人口统计一些因素的限定与选择, 动态生成符合用户需求的个性化的地图或者报告。

在资源组织方面, 国家历史地理信息系统(The National Historical Geographic Information System, NHGIS)是社会探险家的主要合作伙伴。它的数据组织方法影响着社会探险家的资源组织方法。人口普查数据虽然来源于一个机构, 但是由于年代久远, 很多数据被分散保存。覆盖1790-2000年美国人口普查摘要数据大约有670GB, 它们分散在几十个档案馆中, 这些数据以相互不兼容的格式被保存在磁盘、CD-ROM或纸张等多种载体上。NHGIS使用了其中可机读的那一部分数据, 把它们转化成同一种格式, 采用DDI标准(Data Documentation Initiative Standard), 社会探险家即依据这种元数据标准组织描述所藏资源。

DDI是一个通过国际合作的、为描述社会科学数据的技术文档而建立的元数据标准。按照社会科学研究资源的共性, DDI把对每个资源集合的描述分成5个部分, 每个部分分别从不同的角度描述资源对象。DDI的最顶层元素是CodeBook, 该元素有四个属性: ID(文档中每个元素对应唯一的ID)、xml:lang(语种属性, 建议采用ISO的语言代码)、Source(来源属性)和Version(DDI的版本)。在顶层元素CodeBook下有5个部分: (1) 文档描述(Document Description); (2) 研究描述(Study Description); (3) 数据文件描述(File Description); (4) 变量描述(Variable Description); (5) 其他相关信息(Other Material)。

DDI概念模型的设计目标有以下几个: (1) 在特定生产和技术环境下, 从数据集的创建过程中获取信息, 并且在数据进出特定环境时对其能进行转换; (2) 该模型下的子模块能够记录数据和数据生成过程中产生的信息, 包含结构信息以及数据的发现、提取和操作过程中必要的关联信息; (3) 必须在各种技术环境下都能通用; (4) 为特殊的数据类型和存储提供特殊的模块, 以便一致地使用DDI中的所有元素; (5) 独立地使用各个模块。DDI3.0概念模型如图3所示。

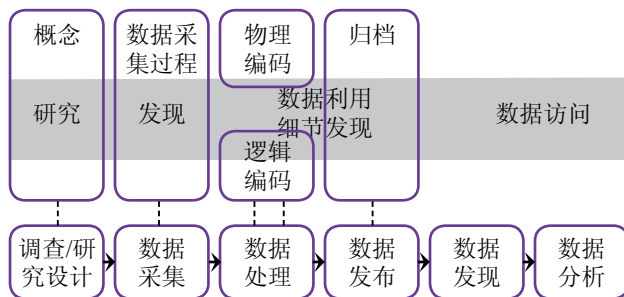


图3 DDI3.0的概念模型（草案）

3 技术特征

3.1 数据采集、存储和输出

社会探险家的数据采集得益于The Great American History Machine和NHGIS等项目。前者是一个历史文献数字化的项目, 提供数字化方面的支持。后者是社会探险家的重要合作伙伴, NHGIS使用的是行业标准、开源技术, 包括Java、Tomcat、Apache、MYSQL、LINUX操作系统, 因此社会探险家也采用了这些技术。

有关社会探险家的数据如何整合, 交互式地图如何处理, 人口普查信息如何描述等方面的资料很难获取, 只能从其网站及其他行业文献中略知一二。给人口普查数据绘制地图的技术不是现成的, 社会探险家选择GeoMicro公司提供的AltaMap软件作为平台来绘制主题地图, 配以它自己的MapServer软件在其网站上提供所有的地图产品。AltaMap是基于互联网的GIS软件, 是由美国海军、通用电气和MapQuest共同创造的交互式地图软件, 它具有线性可测量性的特点, 并且允许大量用户同时使用。此外它还使用了ArcGIS、SAS等其他的软件。

社会探险家的网站运用aspx文件格式提供信息, 与xml兼容。它提供的产品具有多种输出格式。地图可以打印、保存, 多个图片可以制成幻灯片, 可在Microsoft PowerPoint中播放。用户生成的报告是Microsoft Excel文件形式, 可以支持Excel 2007、2003、2000等版本(1997以前的版本每页只能显示255列65535行)。报告可以下载, 包含两种格式: CSV格式与PDF格式。CSV全称是Comma Separated Values, 这是种高度便携式格式, 可以使用多种软件打开阅读, 包括许多统计软件包、数据库、电子表格和GIS程序。也可以直接用Excel打开, 或者导入Access中打开。PDF格式的报告比

较多,其中1990至2000年的统计数据还可以链接到人口普查局的网站,查看详细内容。

3.2 硬件和软件要求

任何与因特网联接并支持网络浏览器的计算机都应可以访问社会探险家的网站。官方支持的浏览器包括Microsoft IE5.0以上版本、Netscape6.2以上版本、Mozilla1.4以上版本、Firefox0.8以上版本、Safari1.2.1以上版本。使用语音动画帮助功能,需要安装6.0以上版本的Adobe Flash软件,用户必须打开JavaScript和cookies,确保弹出阻止器失效,以便生成弹出窗口。

3.3 界面设计

社会探险家的主页纵向分为三个部分,分别是解决方案与产品、新闻与公告、捐赠信息,其主要的产品“地图”、“报告”则放置于左侧醒目的位置。比较有特色的是其地图界面,简洁清晰。用户选择2000年的普查地段后,默认的界面是用不同颜色表示人口密度的整幅美国地图,如图4所示。



图4 社会探险家地图功能页面

地图的上端是包含文件、后退、前进、初始视图、打印、查找、放大、缩小、浏览以及生成报告等工具按钮,下端是生成及播放幻灯片的工具,右侧是可供选择的人口普查的年代、各种项目(包括年龄、收入、教育以及结果显示方式)的下拉式菜单。该地图可以由用户选择放大或缩小,也可以利用查找工具直接定位某一城市或者街道,并用虚拟图钉标志出来。根据用户选择的年代、区域以及有关人口的因素,可以生成地图与报告或者制成幻灯片输出。

报告界面主要利用下拉式菜单,供用户选择地理层面的类型,选择地区,选择数据库,选择数据类型等,然后生成报告。

社会探险家提供完美的在线帮助信息,包括概况、地图与报告工具、理解人口普查三个部分。它的帮助信息非常清晰详尽,利用图示一步一步解释使用方法,并且提供配以解说的flash动画来演示地图与报告的使用方法。此外所有帮助信息还提供可以下载的PDF格式的文件。帮助页面的下端有“常见问题”的链接,这为用户提供有关支持浏览和系统需求的信息。

4 服务特征

4.1 目标用户

社会探险家提供非常重要的人口普查数据资源,但是它毕竟是一个商业性网站,虽然它的一些服务免费提供给公众,但是大约80%的数据资源仅限于付费用户。它的主要服务对象是大学图书馆、媒体以及其他感兴趣的机构。根据访问服务器数量的不同,会员每年的费用从300美元到2000美元不等。

4.2 服务方式与维持

站点服务的亮点是交互式的地图工具,它允许用户根据自己的选择来创建人口统计地图,也允许用户从国家到单独街道等不同层面来查看数据特点。地图包含地区信息检索功能,使用户可以找到某个具体地点,其幻灯片功能为用户创建数张地图来展现某区域几十年来人口变化的趋势。缺点是没有搜索引擎,用户或会员必须逐个搜寻文本,由于人口调查文件非常庞大,所以查找起来很不方便;没有实时帮助,如果用户遇到的问题超出帮助信息的范畴,只能通过电子邮件来寻求并等待答复。

在项目创建期结束后,社会探险家目前由明尼苏达州大学的人口中心保证支持继续运作,计划是十年,确保其数据库长期可获取。该中心联络了政治与社会研究学院制定长期保存与运作的目标。由于可持续发展需要获得稳定的经费保证,该数字图书馆的前途尚不稳定,未来仍需获得必要的资金保障。

4.3 版权管理

社会探险家严格遵守美国版权法律规定,对其资源拥有所有的版权。它的资源包含免费与收费两部分,它允许用户使用、展示、发行、复制其提供的所有免费数据,包括文本、图形、幻灯片及其他资料,但必须注明来源以及符合使用惯例。收费部分的资源只对付费用户或者捐赠者开放,利用这部分资源必须严格遵守使用条款及细则,只能在授权用户之间或者在学术和教育方面的讲座、研讨会或者类似的活动上使用,并且只能采用显示和打印的方式。无论是单独地还是并入其他材料,社会探险家的数据都不得用于其他公开的数据库或网站中,也不得以电子的或其他方式提供给无授权的个人或机构使用。

所有资源的免费使用包括以下几种情况:1. 学术活动,包括文章、书籍和其他学术刊物;2. 课堂教学和相关课堂活动;3. 学生的功课;4. 作为非商业的学术或者教育方面的公开展示,比如教育、文化、学术性研讨会、讲座、会议、展览、工作组或者类似的非赢利性专业活动;5. 研究或论文,包括论文的复制品、图书馆的收藏等。所有这些使用都必须注明出处,并符合惯例。

5 评价和建议

社会探险家采用交互式地图形式为用户提供简便的方法利用人口普查数据,这些本来分散在各个机构、具有不同格式、由不同软硬件支持的人口普查数据被整合到一个网站中,这本身就是一个创举。社会探险家是一个有价值的社会学项目,随着其包含的普查记录越来越丰富,它有可能成为研究美国历史的一个不可或缺的工具。但是它也有一些需要改进的地方:(1)地图的显示:用颜色深浅来表示人口的密度,符合逻辑,很容易理解,但是在小点的处理上有缺陷。比如红色的小点代表一个小的族群,但是在显示一块区域时,那些小点可能会显得很多,造成一种该族群人数很多的假象,因此有必要再完善这些小点的表示。(2)数据的精确性:存在一些明显的数据错误需要修正。(3)搜索引擎:若能够对数据全面的编目,提供搜索引擎服务,将为社会探险家增色不少。(4)长期运营计划以及资金支持:缺乏未来计划以及固定资金支持,可能阻碍社会探险家信息的更新以及数据的长期存取。

参考文献

- [1] FALUDI J. Worldchanging: social explorer [EB/OL]. [2008-9-15]. <http://www.worldchanging.com/archives/002564.html>.
- [2] GeoMicro, Inc. [EB/OL]. [2008-9-15]. <http://geomicro.com/gis/>.
- [3] LeGATES R. Teaching spatially integrated research methods [EB/OL]. [2008-9-22]. <http://gis.esri.com/library/userconf/proc05/papers/pap1009.pdf>.
- [4] welcom to NHGIS. National Historical Geographic Information System [EB/OL]. [2008-9-22]. <http://www.nhgis.org/>.
- [5] NSF. Collaborative research: a digital library collection for visually exploring united states demographic and social change [EB/OL]. [2008-9-22]. <http://www.nsf.gov/awardsearch/showAward.do?AwardNumber=0226279>.
- [6] Social explorer [EB/OL]. [2008-9-06]. <http://socialexplorer.com/pub/home/home.aspx>.
- [7] Directions Magazine Press Release: Social Explorer Releases Interactive Maps and Reports 1940-2000 [EB/OL]. [2008-9-28]. <http://www.directionsmag.com/press.releases/?duty=Show&id=15581&trv=1>.
- [8] Data Documentation Initiative [EB/OL]. [2008-10-05]. <http://www.icpsr.umich.edu/DDI/index.html>.
- [9] 杨波,胡立耘. 用于社会科学信息组织的元数据标准——DDI[J]. 现代图书情报技术,2005(8).

作者简介

刘燕权,毕业于美国麦迪逊大学信息学院,现在南康涅克州立大学任教。曾在各类书刊杂志上发表80余篇专业论文。他近年的出版物主要集中在数字挖掘、数字图书馆、信息存储、数字化及多媒体技术、国家信息基础结构、图书馆统计及管理。通讯地址: School of Communication, Information & Library Science, Southern Connecticut State University, USA 06515. E-mail: liuscsu@gmail.com
马凌云,上海师范大学图书馆,馆员。通讯地址:上海桂林路100号 200234

Social Explorer: A Cartographic Digital Library

Yan Quan Liu / Southern Connecticut State University, New Haven, CT, USA

Ma Lingyun / Shanghai Normal University Library, Shanghai, 200234

Wang Qun / Beijing Normal University Department of Information Management, Beijing, 100875

Abstract: Social Explorer is a premier U.S. demographics digital library providing an Internet connection access to the American census data from 1790 to present. It features over 15,000 interactive data maps and reports, is a collection of interactive demographic maps that can be viewed, queried and manipulated. The article provides an extended review on the construction and status of the digital library, including project overview, resources organization, technological features, user services, as well as comments and suggestion made by authors.

Keywords: Demography, Census, Social Explorer, Digital library, National Science Foundation (NSF)

(收稿日期: 2010-04-02)