文章编号:1005 - 1538(2017)03 - 0097 - 07

·论 坛·

博物馆大数据运用初探
——以上海博物馆数据中心项目为例

刘 健
(上海博物馆,上海 200003)

摘要:对博物馆数字化建设来说,大数据的运用是一个无可避免的话题。为广泛发挥大数据在博物馆业务工作中的作用,从上海博物馆的数据中心项目入手,阐述了数据中心如何在加强收集和存储博物馆各项数据同时,紧紧围绕数据采用智能化的分析手段,挖掘更加有价值的信息。利用大数据对业务进行分析,加工形成有用的数据模型,进而为形成博物馆数字化运营管理体系打下了初步的基础。本研究介绍了数据挖掘技术的有效应用情况,并提出了博物馆大数据建设未来的发展思路。

关键词:博物馆;大数据;应用;发展
中图分类号:N28; N37 **文献标识码**:A

0 引 言

在人类从 IT(Information Technology,信息技术)时代走向 DT(Data Technology,数据处理技术)时代的过程中,大数据的运用是一个无可避免的话题。DT 的核心,是关于数据驱动的创新,也就是基于海量数据的价值挖掘为重心的创新体系及模式。对于博物馆来说,通过对博物馆所有的属于文化历史遗产一部分的藏品资源数据和公众的文化需求数据及行为数据进行收集、分析、挖掘和整合运用,为博物馆进行公众教育、文化传播、科学研究、征集收藏等任务提供新的平台、内容和形式的支持,并为博物馆的精准化管理提供数据支持。同时,从一定程度上增加博物馆本身的收藏研究方向以及展示角度,塑造基于大数据基础上的博物馆工作新模式,也应该是博物馆数字化建设中的应有之义。出于以上考虑,近期上海博物馆开始进行有关博物馆大数据应用的尝试,初步成果就是上海博物馆数据中心的建立。

1 博物馆大数据

“大数据”指的是一个体量特别大、数据类别特别多的数据集,并且这样的数据集已经无法用传统的数据库工具对其内容进行抓取、管理和处理。提到博物馆大数据,首先会想到的是藏品数据,或者更进一步想到的是观众的数据。如果以宜粗不宜细为原则,以博物馆的功能为导向,博物馆数据可以分为以藏品为核心产生的藏品数据(包括藏品本体的编目数据、检测数据、研究数据、保管使用数据等)、以博物馆业务行为需要产生的管理数据(包括博物馆日常管理流程所产生的数据、举办各类活动所汇聚的数据、与社会各方发生联系所形成的数据等)、以博物馆各类传播活动、数字化传播工具及其反馈机制所构成的传播数据和以观众行为为基础所累积的观众数据这四大类。在数据的力量日益受到重视的今天,数据能否成为博物馆工作的推动力,成为博物馆跃上新台阶的助推器,则在很大程度上取决于对博物馆大数据的运用。1989 年,美国管理学家罗素·艾可夫(Russell L Ackoff)在《从数据到智慧》(From Data to Wisdom)一文中,构建了著名的 DIKW 体系[1],清楚阐述了数据(Data)、信息(Information)、知识(Knowledge)及智慧(Wisdom)之间的相互关系(图 1)。其中,最底层的数据是基础的数值;越向上,数据的相关性就越强,价值也相应提升,而智慧的数据将拥有相当的语义判定和一定的逻辑推

收稿日期:2016 - 12 - 19;修回日期:2017 - 03 - 20
基金项目:国家科技支撑计划资助(2006BAK20B03)
作者简介:刘 健(1962—),男,上海大学文学院历史系考古与博物馆专业,本科,研究方向为博物馆数字化研究,E - mail: liujian@shanghai - museum. org

理的能力。由此来观察目前传统博物馆体系里常见的数据模式应用,则会发现很少有处于 DIKW 体系的第三层和第四层的数据应用。如何通过数据价值的提升,从而发现新的可供利用的视角和方向,正是努力的方向。通过 10 多年的数字化建设工作,上海博物馆目前已积聚了并依然在不断聚集着数量相当可观的数据。因此,如何将目前上海博物馆已有的数据整合、挖掘并予以展示,是上海博物馆数字化建设工作所应面对的。上海博物馆数据中心项目就是希望在将数据集中存储的同时构建出信息资源的体系,再按照一定的方式和规则对资源数据进行归并、处理、筛选,将数字资源汇集后管理利用并进行初步的挖掘分析,然后采用新颖的多媒体交互展示方式对数据进行展示,最终达到利用数据资源、数据分析、数据展示产生新的博物馆效益的目的。此项目的主要难点,就内容来说,一是博物馆大数据的收集整理和结构化,二是博物馆不同数据之间的智能关联、挖掘和主题的推理;从技术上来说,则是在于智能无线数据采集技术、大数据分析与智能推理技术以及大数据可视化这几个方面。从项目的完成情况看,这几个难题在项目设计和实施过程中得到了较好地解决。

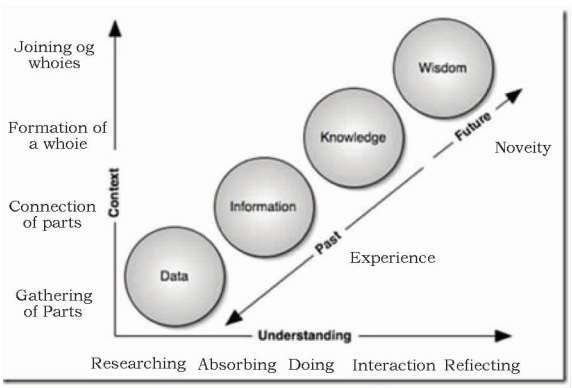

图 1 DIKW 体系模型[1]

Fig. 1 Model of DIKW pyramid

2 项目建设情况和创新点

项目要求以博物馆观众服务和专业研究为导向,结合上海博物馆运行现状,探索建立统一的数字资源管理和展示平台的可能性。根据国家一级博物馆运行评估指标体系,不断优化数字中心运行评估指标模型,模型涵盖展馆、展览、藏品、观众等核心指标,描述博物馆信息资源及其载体,构建、挖掘、分析呈现信息资源及核心指标之间的相互联系,及时准确在以上几个方面反映出上海博物馆实时运行状况。为上海博物馆今后的精准化管理、大数据挖掘以及可视化展示工作的进一步发展打下基础。系统要求具有与上海博物馆现有应用系统的信息共享与管理,具有高度的集成性及可视化的信息图展示功能,同时留有一定冗余,能满足今后系统不断更新升级的需要。由此,从数据的视角去重构博物馆的数字运用和管理是这一项工作的重点:系统之间的畅通联接,各类价值数据的清洗和挖掘,采集观众行为数据、建立量化的评估数据模型的研究,这些都是本次项目工作的核心。

数据中心是一个综合性的管理展示平台,需要将上海博物馆现有内外网的系统整合到统一的平台上,因此在设计之初就必须要充分考虑到各个系统间直接的接口调用与整合问题。在设计过程中充分考虑各连接系统接口所涉及的应用扩展情况,给各个接口定义 web service 接口规范,并采用目前业内比较广泛使用的 XML、EXCEL 等格式数据作为主要的数据传输载体来进行数据交换,让需要展示的系统都有统一的传输格式传送到显示大屏上来。同时也注重数据传输过程中数据的中断和反馈机制,以保证了数据的稳定性。另外,在整体页面展示的设计上也是完全依照屏幕的大小和尺寸来量身定做,使设计的页面风格、布局与大屏的整体相协调。

这一项目包含了多个系统的功能,功能设计见图 2。具体的功能包括:

图 2 数据中心首页

Fig. 2 Homepage for the data center

1）观众流量展示。实时显示到馆的总人次数、每个入口人数、每个出口人数。2）网站访问状况的展示。包括实时的点击量、访问来源分布等。3）藏品数据的展示。包括博物馆藏品总体情况，以及藏品类型、年代、库房等分类统计；显示博物馆藏品出入库的一般情况，如藏品修复、保护等。同时涵盖物联监测展示，显示本馆内物联网监测各项运行状态。4）机房设备运行展示。显示网络、主机、安全等设备运行状态。5）明清家具馆 AP 接入人数展示。实时展示当前明清家具馆内的观众数量情况和他们的移动轨迹，以及在家具馆停留 15min 和 30min 以上的人数，以及具体热门展品前人数累计和时间累计的情况等（图 3）。6）具体某一特展的展示。除了相关特展的基本情况显示外，还通过比较有特展和无特展时候人员的变化状况来试图探索其中的规律性。另外，还能关联了解特展期间网站、APP、微信等传播系统上的关注人员的变化情况。7）单一文物的展示。在基于藏品基本信息的介绍的前提下，将具体文物的地理信息、功能诠释、修复的数据进行了有效的整合，使单件藏品的介绍在时间和空间上都得到了有效的延伸；并通过数据挖掘，将附着于藏品上的显性信息和隐性信息以多媒体形式展示出来。该项目的创新点主要表现在以下 4 个方面。

图 3　明清家具馆实时数据显示

Fig. 3　Chinese Ming and Qing Furniture Gallery's real – time data display

2.1　基于网状可自我量化的数据模型

结合博物馆运行现状设计了螺旋式上升评估指标模型方案，整个项目的数据涵盖展馆、展览、藏品、观众、活动、微环境等核心指标，描述博物馆信息资源及其载体，构建、挖掘、分析呈现信息资源及核心指标之间的相互联系，及时准确、全面综合地反映了上海博物馆实时运行状况[2]（图 4）。同时利用信息可视化手段，提供个性化的观众分析报告、藏品统计报告、设备运行报告等新形式的服务，包括观众流量可视化展示、网站访问可视化展示、藏品数据可视化展示、物联监测可视化展示、设备运行可视化展示。另外还提供观众基础属性分析、观众信息反馈分析、展区观众行为分析、综合评估的设计与开发。

2.2　基于时空数据的服务创新

大数据之所以强大，还因为它提供了新的观察角度和新的研究方向。例如，就观众数据而言，首先，观众服务是一款典型双向的数据应用，用户既是数据消费者（使用导览等信息），又是数据生产者（产生用户行为信息），两者同等重要。其次，这是

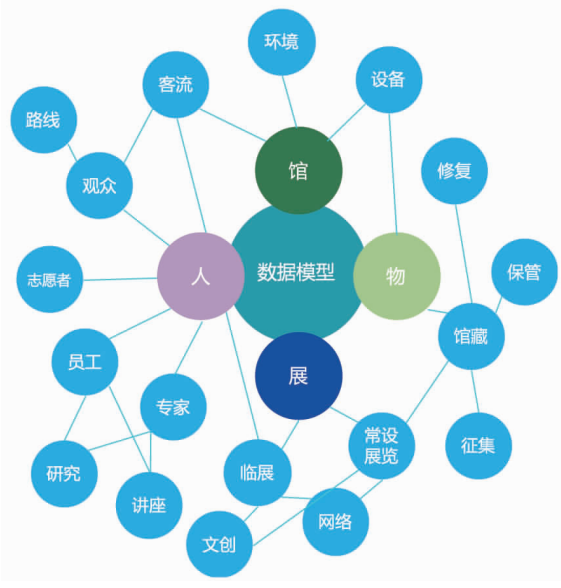

图 4　博物馆数据整体模型框架（第一级）

Fig. 4　Museum's data model framework （First – Class）

一项高频度的数据应用，按博物馆的人流量及用户游览时间，会在每天的参观周期内产生大量的数据，

这个数据既会有宏观面的大数据，也会有体现个性化的小数据。第三，用户是分层的，不同的用户感兴趣的信息数据是存在差异化的。基于数据的复杂性，开发的系统使用空间信息为主线对数据进行组织构造，以图层的模式将数据的接入或表现进行划分：底图即为博物馆物理的楼层图；不同的图层代表了不同的信息视图逻辑，可接入不同的数据源或数据类型，也可能会面向不同的用户数据需求提供完整的功能逻辑；POI（point of interest，信息点）以矢量形式标注详细数据，也可用于动态数据的标注，如实时流动的参观者位置。图层式的模式不仅使数据视图更直观，用户的使用操作更便捷，还易于数据扩展和数据切分，今后还可通过图层增加团队视图等（图 5）。

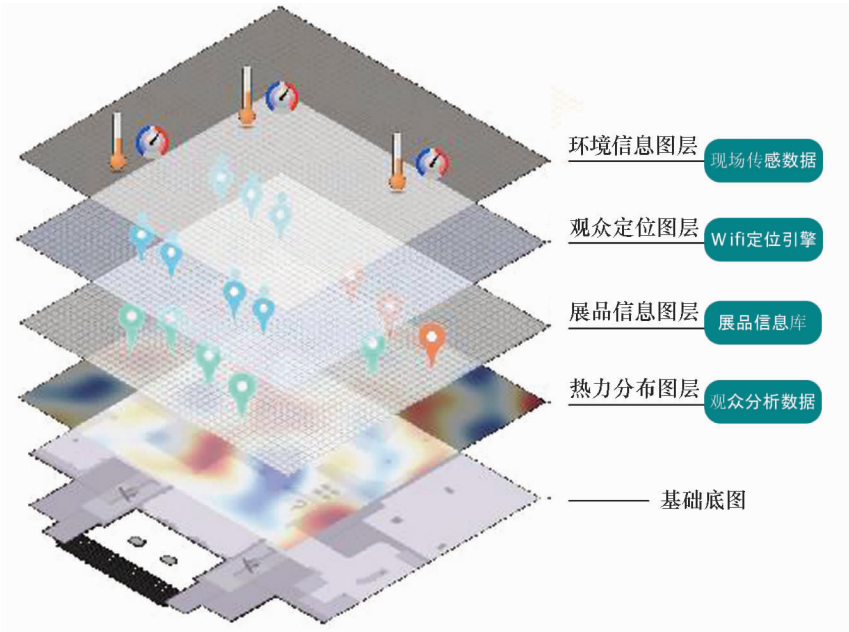

图 5 博物馆观众导览系统的主要图层

Fig. 5 The main layers of the museum's audience guiding system

2.3 基于多元价值的数据挖掘

面对庞杂的原始数据，系统采用了 HDFS（The Hadoop Distributed File System，分布式文件系统）、MapReduce（做大数据处理软件框架）、Hive（数据仓库工具）等中间件构建了一套完善的大数据系统，通过对定位系统、App、环境传感（物联网）、互联网的数据进行清洗、融合和转换处理，形成了 HDFS 的数据集，提供多样化的大数据服务，处理过程见图 6。

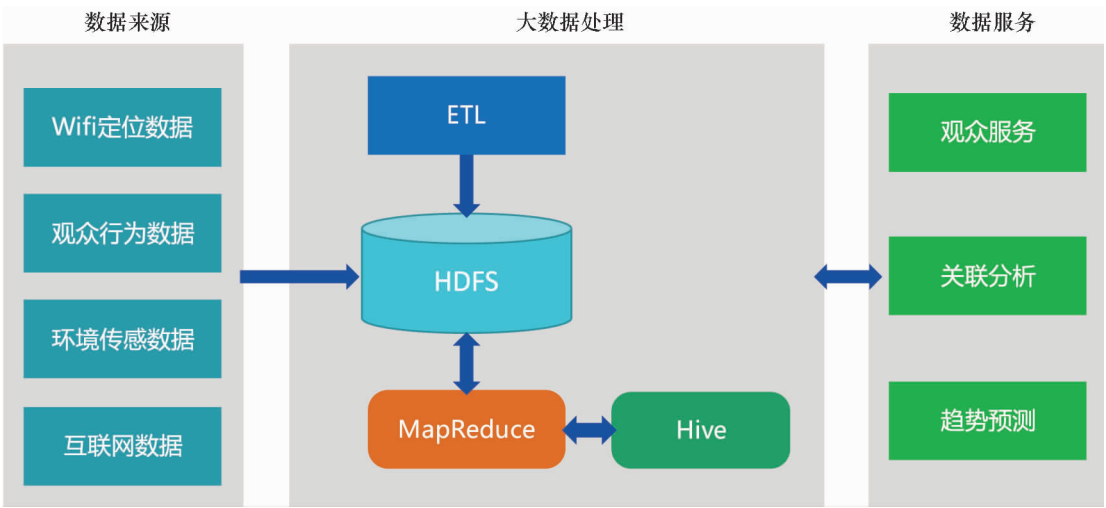

图 6 源数据分析处理图

Fig. 6 Source data analysis and processing chart

目前,系统已经在观众行为大数据集的基础上实现对观众反馈、参观热点分析、停留时间、关注内容、社交行为进行综合分析。此外还能提供:观众分布及轨迹分析、展馆区域热图、重点展位详细分析、现场驻留及关注度分析等;举办展览的线下参观数据(平日数、实时人数、学生数等)、线上数据(网站浏览量、网站预约量、微信浏览量、微信关注量、APP下载量等)等。通过大数据分析可以发现其中的规律、关联,甚至可以推理用户的行为模式。

2.4　基于国家标准的评估指标体系

为了进一步发挥数据的作用,利用大数据分析及其可视化工具有效地从庞杂的业务数据中提炼出数据内涵,发掘内在关联。根据国家制订的《一级博物馆运行评估指标体系》,结合博物馆运行情况提供不同的评估指标模型设计方案。比如根据评估指标中的陈列、科研、文化交流、数字化、文创、教育活动及媒体关注这样几个大的指标项,为单个藏品做了一个来自社会维度的藏品社会利用评估体系的模型,根据藏品利用中不同指标的权重、属性、数据给出了评价的量化数值,尝试采用从量化数据角度来展示藏品的社会效益,虽然未必完全准确和有效,但是通过不断的摸索和校正,相信数据的价值会逐步显现。总之,本项目通过大数据分析与可视化技术的运用,使之成为博物馆深化应用、提升应用层次、强化管理能力的有效手段。数据中心在加强收集和存储博物馆各项数据同时,还需紧紧围绕数据采用智能化的分析手段,挖掘更加有价值的信息。并利用大数据对业务进行分析,加工形成有用的数据,进而为形成博物馆数字化运营管理体系打下了初步的基础(图 7)。

图 7　单个藏品的可视化数据和评估模型

Fig. 7　Visualizing data and evaluation models for individual collection

3　博物馆大数据建设未来的发展

大数据技术必将给博物馆带来深刻的影响,大数据技术及思维也将改变传统博物馆的存在模式和工作理念,改变博物馆业务的思路和流程,突破博物馆对资源的垄断。未来的博物馆大数据之路还很漫长,这条路该如何走,以下 3 点是未来发展必须考虑的,也将是进行数据中心后续建设中所应该有所作为的。

3.1　数据结构知识化

当博物馆数字化建设开始发生从以系统为中心到以用户为中心的转变时,数据的知识化问题就被自然地提了出来。所谓知识化首先就是将杂乱的数据改变成结构化的实体知识,甚至可以顺着知识图谱探索建构更深入、广泛和完整的知识体系,进而激发用户发现意料之外的知识。换而言之,真正的大数据应用应该体现在数据挖掘的深度。正如宋新潮先生所说[3]:“把过去、现在以及未来的大量文物数据,整合为有机知识体系,提炼为可高效利用的知识。从而更好地实现博物馆的教育和研究功能,使博物馆真正成为‘虚拟世界的真实性源泉’和数字时代知识的创造、生产机构。”如果说藏品的本体数据也就是它的基本信息还构不成大数据研究的条件,但若能与公众大数据以及其他社会类数据,比如像各类的学术性平台数据进行智能关联,有效整合,形成关联数据,就能够形成资源挖掘和主题推理的条件,从而能够使博物馆专业人员在大数据的条件下便捷、有效地利用这些资源。这里的关键在于藏品知识库的建立。什么是藏品知识库,它与现在一般所用的藏品管理系统还有所不同。最大的特点在

于它能为一件藏品建立比较完整的知识体系,更注重关联性而非仅是结果。也就是说,将基于数据的应用发展到基于知识的应用。当然,构建并非孤立进行。在知识库的构建过程中,需要实现多数据源的知识融合。比如通过藏品数据的原始积累,再将文物藏品的本体数据与该文物的海量的、多源的、异构的数据(如考古、地理、环境、测试、文献、学术研究),通过自然语言处理、大数据分析以及计量学,软件科学等组织起来,同时也可与不同格式、不同结构的数据间建立关联。如文物藏品的文字信息、二维与三维信息、影像甚至声音之间的相互联系、附属关系等;并在不同的应用模型间相互映射和自由索引切换,最终形成能够支撑文物藏品研究的知识呈现体系。这种联系和模型叠加,既是博物馆数据资源整合的基础,也是博物馆数据应用的重要支撑。国际上也已经有一些这方面的尝试,也有了不少有效的工具。其重要的一个网络基础就是语义网的应用。比如大英博物馆的探索空间(research space)就是基于语义网的一个成功尝试。而这一切,都离不开关联数据和数据的开放。

3.2 数据资源开放化

目前社会上对博物馆的资源开放的呼声很高,尤其是藏品资源。但在实体资源还不具备开放的条件下,加快开放数字资源是缓解这一需求的有效方法之一,也是博物馆进行数字化建设的本质所在。当然,即使是数字资源的开放也不可能一蹴而就,还是会有一个过程。但在学术应用领域,这一过程实在有加快的必要。比如前面所提到的数据关联。要发布关联数据,按照互联网的发明人——Berners – Lee T[4]的说法,就应遵循4个原则:1)使用URI作为任何事物的标识名称;2)使用HTTP URI(网络上的统一资源标识符)使任何人都可以访问这些标识名称;3)当有人访问某个标识名称时,提供有用的信息;4)尽可能提供相关的URI,以使人们可以发现更多的事物。其实质就是开放数据,如果没有数据的开放,所谓的有机知识体系的形成就是空中楼阁。如果说数据采集是博物馆大数据应用的物质基础,那么数据开放就是它的行为准则,而这更多地取决于博物馆人思维的转变。早在2002年,图书情报界就达成了布达佩斯开放存取先导计划(BOAI,Budapest Open Access Initiative),要求实现对学术期刊资源的开放存取。由此计划逐渐衍生出了开放数据的概念。与开放存取一样,开放数据的目的也是消除共享障碍,赋予用户的使用权利。即如开放数据手册[5]所言:"开放数据是一类可以被任何人免费使用、再利用、再分发的数据——在其限制上,顶多是要求署名和使用类似的协议再分发"。但与一般的共享概念不同的是开放数据更注重对数据的"再利用、再分发",即通过对数据的聚集、整合,从中实现数据价值的"增值":一是数据通过开放完成了二次生产,从而使应用价值增加;二是从获取者的角度来看,有更多的人能够得到这些开放数据及其传递出的价值,由此产生信息共享空间的增值。当然,更为重要的是,在这一理念的引导下,博物馆可以突破本馆和本行业的限制,实现馆藏资源与外部资源的互联,从而使信息流的采集、整理、发布与使用者的利用、收藏、互动之间能够形成一个闭合的数据资源链,且不同的结点都可以进行资源的共享。在博物馆界,开放数据在欧美也有所使用,如上文已介绍的大英博物馆,还有欧盟的Europeana项目等。但在国内还很少见到有这方面的应用。

3.3 数据研究平台化

在博物馆数字化建设的过程中,数字化研究始终是处于被遗忘的角落。基于网络在一个开放的学术圈内进行协作研究,是数字化时代知识创新的一个重要机制和运作形式。实际上,在博物馆累积了庞大的数字资源和能量以后,如果在具备了前面所述知识化和开放化的前提下,博物馆应该考虑尝试以数字资源为主要对象的数字化研究工作,建立起以协作研究为主的科研数字平台。即根据博物馆科研活动的特点,基于网络化管理,以数字资源的整合、共享、科研工具的共有为前提,以数字化研究手段的普遍应用为基础而形成的一个开放式的研究平台。平台将建立藏品研究数据库(知识库)及相应的查询接口,按照数字化的格式编制成相互关联的数据集合,其搜索模块将直接检索到与研究相关的知识单元,如概念、表格、数据、事件、多媒体等。未来的科研数字平台将集科研、资源、管理于一体。其中,博物馆科研人员无疑是系统运行的主导,通过角色划分或权限分配来规范不同人员的操作内容和范围;以科研人员的研究特点来思考系统功能架构的构成,系统将支持各类数字化的研究手段和策略,同时构建出规范、开放、安全、基于服务的新型网络化科学研究环境,并运用网络技术提供了一种崭新的科研协作模式。这样一个平台的建立,它的好处是显而易见的:首先它可以打破目前博物馆研究中普遍存在的学术孤岛现象,为外脑的引入与社会性协作建立了条件。其次是可以发挥网络的互联特点,进行最大限度的数字资源的整合。其三是能起到弭平博物馆中久存的文理鸿沟的效果。很多人都担心

人文学科对数字化的不适应。事实上,在人文学科中,“数字人文(Digital Humanities)”已经日益流行,即“借助数字科技方能进行的人文研究”。2016 年美国新媒体联盟的《地平线报告》(博物馆教育篇)中,就在一年内会采用的技术里首次明确提出了数字人文技术的概念。数字人文研究目前常见采用的技术方法,如历史地理的可视化、采用历史文献的文本挖掘与词频分析及考古学方面的图像解析、色彩还原和数字重建等等,也在国外博物馆的各领域中逐渐开始得到应用。总之,数据研究平台化的形成,将给博物馆带来符合社会发展趋势的改变及随势增长的契机。同时这也将是博物馆大数据建设的一个必然结果。

4　结论

上海博物馆数据中心项目的建设使博物馆的数字化建设工作迈上了一个新的台阶。首先,此次建设初步实现了各类不同来源、格式、类型数据的融合,建立起了统一的数字资源聚合管理平台;其二,通过多媒体展示屏,以多屏组合方式实现显示内容的自由布局组合,将复杂的数据以图形、图像、多媒体等各种形式做了效果良好的可视化呈现,形成了统一的数字资源展示平台;第三,是对藏品数据、展览数据、观众数据进行尝试性挖掘,并建立起了初步的量化评价模型,成为博物馆大数据挖掘平台的一个雏形。上海博物馆的这一探索将对国内博物馆大数据的应用和发展起到积极的推进作用。

参考文献:

[1] Russell L Ackoff. From data to wisdom[J]. J Appl Syst Anal,1989(16):3-9.

[2] 童　茵. 博物馆数据创新公共服务体系的研究[J]. 软件产业与工程, 2016(6):54-56.
TONG Yin. The research of innovation public services architecture of museum data[J]. Software Ind Eng,2016(6):54-56.

[3] 博物馆智慧化之路——数据知识化与呈现方式[EB/OL]. [2017-03-10] http://huadong.artron.net/20160103/n807071.html.

[4] Berners-Lee T. Linked data[DB/OL]. [2016-10-17]. http://www.w3.org/DesignIssues/LinkedData.html.

[5] Open data handbook[R/OL]. [2016-07-03]. http://opendatahand-book.org/zh_CN/what-is-open-data/index.html.

On the use of big data in museums ——Shanghai Museum Data Center project as an example

LIU Jian

(*Shanghai Museum*, *Shanghai* 200003, *China*)

Abstract: The use of “big data” is an inevitable consequence of museum digitization efforts. Starting with the Data Center project of the Shanghai Museum, this paper sets forth how Data Center reinforces the collection and storage of museum's data as well as employs intelligent analytical methods based on the data to dig more deeply into valuable information with the purpose of playing an extensive role in the museum's operation. Big data methods analyze the museum's daily operation to form a useful data model and to lay a preliminary foundation for the digitization operation framework of the museum. This work also introduces the effective data mining tools and proposes future big data development strategies for museum use.

Key words: Museum; Big data; Application; Development

(责任编辑　马江丽)