

潘雪丽,郝春云,王 莉. iSchools 成员院校数据监护课程调查[J]. 中华医学图书情报杂志,2018,27(3):24-32.

DOI:10.3969/j.issn.1671-3982.2018.03.006

· 研究与探讨 ·

## iSchools 成员院校数据监护课程调查

潘雪丽,郝春云,王 莉

[摘要] 数据监护是科学管理数据并提供数据增值服务的重要手段,学院数据监护教育是培养专业数据监护人才的重要途径。调研了 66 所 iSchools 成员院校的数据监护课程,获得数据监护相关课程的课程内容形成课程内容标签,并结合科学数据管理生命周期,分析不同国家、不同课程等级以及科学数据管理生命周期不同阶段的课程主题,从课程内容角度进一步了解数据监护课程现状,对国内相关院校开设类似课程提供科学的依据。

[关键词] 数据监护;数据管理;专业教育;人才培养

[中图分类号] G203;G642

[文献标志码] A

[文章编号] 1671-3982(2018)03-0024-09

### Data curation courses in iSchools member colleges of universities

PAN Xue-li, HAO Chun-yun, WANG Li

(Institute of Scientific and Technological Information, Beijing 100038, China)

Corresponding author: WANG Li

[Abstract] Data curation plays an important role in the management of scientific data and provision of value-added data service. Data curation education is an important means to train the data curation professionals. The data curation courses offered in 66 iSchools member colleges and universities were investigated. The contents of data curation-related courses were used to form the tags of different courses. The theme of each course in different countries, level of different courses, and different stages of scientific data management lifecycle were analyzed according to the scientific data management lifecycle. Further understanding the current data curation courses in view of their contents can provide valuable reference for offering data curation course in domestic colleges and universities.

[Key words] Data curation; Data management; Professional education; Professional training

科学研究产生了大量的科学数据,精密仪器和大规模计算的应用使科学研究数据呈指数级增长态势。由于实验环境、实验设备、实验人员的限制,科学数据通常是不可复制和再现的,但又具有重要的现实价值和潜在价值,因此对科学数据的保存和管

理显得尤为重要。数据监护(Digital Curation or Data Curation, DC)是将科学数据的维护、保存和增值贯穿于科学数据生命周期每一环节的长期实践活动<sup>[1]</sup>。在数据生命周期整个过程中对数据的持续监管,不仅能够为学术、科研和教育提供便利,而且能够维护数据质量,提供数据增值服务和复用服务。

近年来,为培养数据监护的专业人才,满足数据馆员的职业技能需求,国内外许多大学及研究中心开展了 DC 相关项目或课程<sup>[2]</sup>。有学者调研了相关院校的 DC 课程<sup>[3]</sup>或 DC 认证项目<sup>[4]</sup>,从教学层次、师资力量、先修课程、指定教材、推荐阅读、作业形式、课程名称、课程目标等对 DC 相关课程或 DC 认证项目的信息进行了归纳整理。周霞等人通过调研

[作者单位] 中国科学技术信息研究所,北京 100038

[作者简介] 潘雪丽(1992-),女,广东清远人,在读硕士研究生,研究方向为信息检索与数据库建设。

[通讯作者] 王 莉(1974-),女,北京市人,硕士,研究员,硕士生导师,发表论文 20 余篇,参编国家标准 3 部,研究方向为数字图书馆技术、信息检索、知识组织。E-mail: wangli@istic.ac.cn

国内高校情报学专业的培养方案,从情报学硕士研究方向提炼课程内容,发现在数据监护方面,情报学硕士教育依然侧重于数据库技术及数据挖掘,鲜有涉及数据的管理与保存<sup>[5]</sup>。上述研究主要从学院网站调查结果入手,结合问卷调查等方式,归纳整理了相关课程的信息,揭示目前 DC 课程开设现状。

数据监护是贯穿于科学数据管理生命周期每一环节的长期实践活动。结合数据管理生命周期分析 DC 相关课程,对于梳理目前 DC 相关课程的内容主题及完善 DC 课程体系都具有重要的意义。本文将结合数据管理生命周期理论,分析数据监护课程的侧重点和发展方向,从内容主题角度进一步了解 DC 课程开设现状,为我国高校开设 DC 相关课程提供参考意见。

## 1 数据与方法

成立于 2005 年的 iSchools 是致力于促进信息领域发展的全球信息学院联盟,联盟成员来自全球 LIS(Library and Information Science)相关学院。截至 2017 年 11 月,该联盟已拥有来自北美、欧洲、澳洲、亚洲等地区的 82 所成员学院。iSchools 成员学院是全球信息学院中的领军院系,其 DC 课程设置对于国内外其他信息学院开展数据监护教育具有重要的启示和借鉴作用。本文采用网络调查法,以 66 所 iSchools 成员学院作为样本进行调查。首先通过访问院校网站搜集院校 DC 课程信息,并对其课程标题、教学大纲、课程目标、课程描述等课程内容信

息进行统计。将获得的 DC 相关课程内容形成课程内容标签,分析不同国家与地区、不同课程等级以及科学数据管理生命周期不同阶段的课程主题,从课程内容进一步深入了解数据监护课程开设现状,同时结合数据管理生命周期,揭示目前 DC 课程的教学侧重点。基于调研结果,结合目前国内 DC 课程开设现状,对我国信息学院开设 DC 相关课程提出建议。

为了便于研究,笔者对 DC 课程做了如下界定:课程名称、课程描述、课程目标或课程大纲中出现 Data Curation 或 Digital Curation 一词,课程的主要内容涉及 DC 相关主题。

## 2 结果与分析

### 2.1 不同国家的 DC 课程主题内容分析

本文共调研了 66 所 iSchools 成员学院。由于其中 7 所(中国大陆 4 所,德国 1 所,西班牙 1 所,丹麦 1 所)未在其网站上公布课程目录及课程信息,因此共获得 59 所院校的有效课程数据。59 所 iSchools 成员学院中有 2 所位于中国台湾,但并未开设 DC 相关课程,其余 57 所 iSchools 成员学院的国家分布及开设 DC 课程的 iSchools 成员学院数量如表 1 所示。共有 32 所成员院校开设了数据监护课程,共开设了 58 门 DC 课程。分析这些学院 DC 课程的主题内容信息,即对所有课程的课程介绍和教学大纲进行人工标注,得到 DC 课程内容标签云图(图 1),整体把握目前 DC 课程教育的发展

表 1 57 所 iSchools 成员院校的国家分布及开设 DC 课程的成员院校数量

国家	调研的 iSchools 院校数量	开设 DC 课程的 iSchools 院校数量	开设 DC 课程的 iSchools 院校比例/%
美国	36	23	63.89
英国	9	3	33.33
澳大利亚	3	3	100.00
加拿大	3	2	66.67
荷兰	1	1	100.00
芬兰	1	0	0.00
挪威	1	0	0.00
瑞典	1	0	0.00
日本	1	0	0.00
菲律宾	1	0	0.00



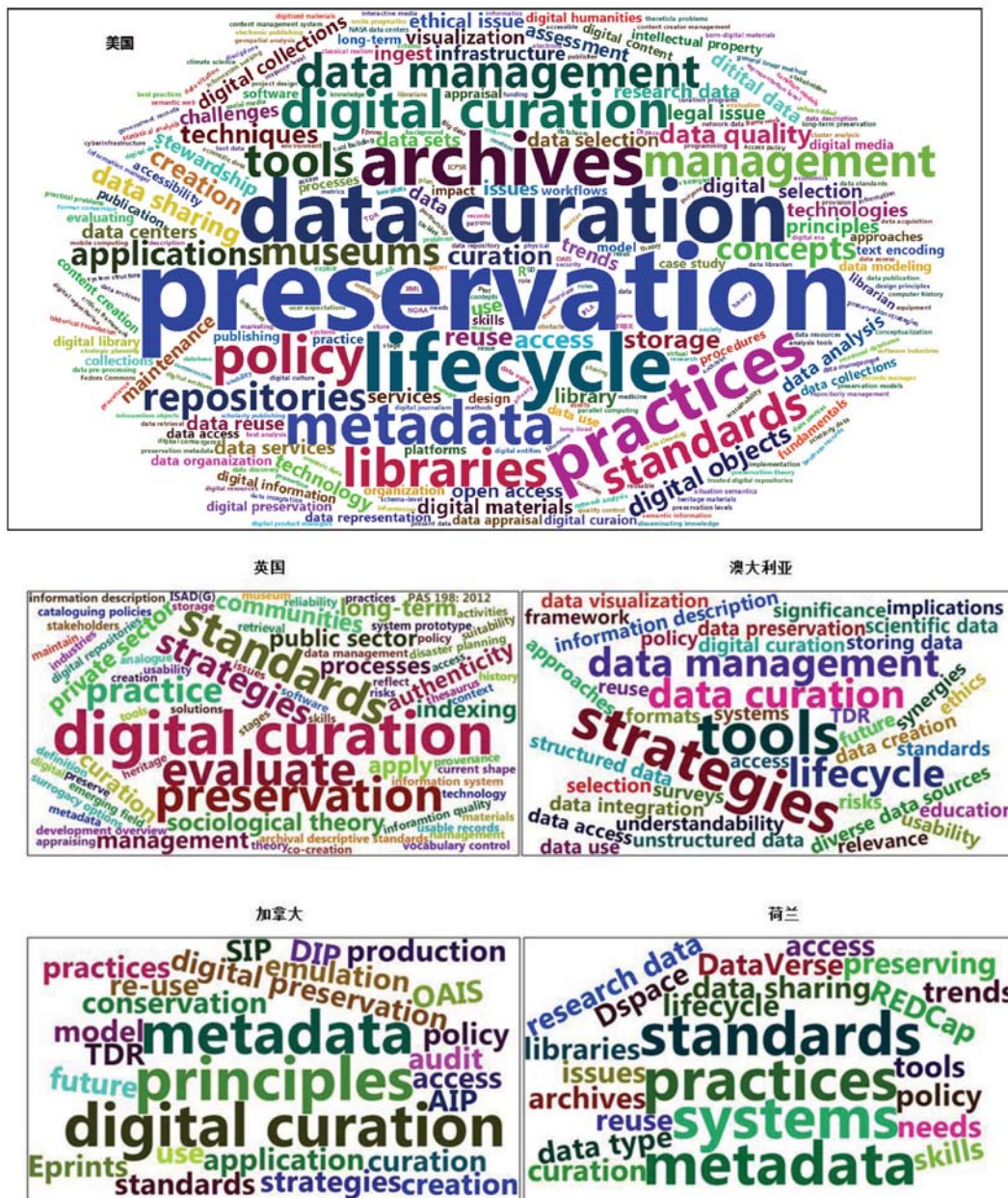


图 2 不同国家 iSchools 院校 DC 课程内容标签云图

的数据监护生命周期管理。墨尔本大学的 E-Science 课程<sup>[9]</sup>强调科学数据管理的生命周期理论以及科学数据管理和处理的方法、工具及面临的挑战,深入讲授数据的特性、结构化和非结构化数据的处理、数据分析、存储、获取以及数据可视化等科学数据的有效管理过程。

加拿大 iSchools 院校的 DC 课程十分重视重要模型,如 OAIS 模型以及数据管理生命周期的学习和运用。多伦多大学的课程 Digital Preservation and Curation<sup>[10]</sup>重点关注数据从预涉入到传播的工作流。麦吉尔大学的课程 Digital Curation<sup>[11]</sup>明确将

OAIS 模型和 DDC 的科学数据管理生命周期作为学生指定的阅读材料,并在课程中设计了一个部分专门讲授 OAIS 模型、AIP、SIP、DIP 和数据管理生命周期。机构库 (institutional repository) 和数字仓储 (digital repository) 作为数字资源保存和利用的重要平台,也是加拿大 iSchools 院校 DC 课程关注的重点。多伦多大学关注数字仓储的特点以及可信赖数字仓储的审计与认证问题。除此之外,麦吉尔大学的 Digital Curation 课程还详细讲授数字仓储的实施与评估问题,讨论常见的 Eprints, DSpace 和 Fedora 等机构库软件工具。Digital Curation 课程中仅有的



32 所 iSchools 学院中,仅有 3 所院校开设了 DC 特殊专题课程。从课程内容标签看,特殊专题主要介绍目前常见的数据监护模型 OAIS 以及现有的科学数据管理的最佳实践项目,如美国国家海洋大气管理局(NOAA)、美国国家大气研究中心(NCAR)、美国校际社会科学数据共享联盟(ICPSR)。

32 所 iSchools 学院中,有 26 所学院开设 DC 常规课程,占 81.25%,而且这些学院的地域分布覆盖了所有开设 DC 课程的国家:美国、英国、澳大利亚、加拿大和荷兰。从课程内容标签看,常规课程出现最多的标签是 preservation,其次是 archives, standards, metadata, management, practices, data curation, libraries, policy, tool 等。可见,科学数据的保存问题是常规课程的重点教学内容。具体来说,常规课程涉及以下内容:概述数字资源或科学数据从创建、选择、组织、保存、存储、获取、使用到复用整个生命周期过程中的原理、标准、技术与方法,如田纳西大学的 INSC562 Digital Curation 课程<sup>[13]</sup>、查尔斯特大学的 INF462 Data Curation 课程<sup>[8]</sup>、伦敦大学学院的 INSTG064 Introduction to Digital Curation 课程<sup>[14]</sup>。该课程重点讲述了数字资源或科学数据长期保存涉及的保存元数据、保存技术和保存挑战问题,同时分析了目前最佳保存实践项目-机构库(Digital Repository)和可信赖数字仓储(Trusted Digital Repository, TDR),从最佳实践的案例中获取数字资源长期保存的实践经验。少数课程还探讨了数字环境下科研团队、数据中心、图书馆与存档馆等不同利益相关方如何管理数据和数字资源的版权、数据安全以及数据隐私、数据道德问题,如加州大学的 262A Data Management and Practice 课程<sup>[15]</sup>、德雷塞尔大学的 INFO591 Data and Digital Stewardship 课程<sup>[16]</sup>、阿姆斯特丹大学的 ARST 556K Research Data Management for Information Professionals 课程<sup>[12]</sup>。

有 3 所院校开设了 DC 认证项目或 DC 培养方向:北卡罗来纳大学的 Digital Curation 认证项目<sup>[17]</sup>、北德克萨斯大学 Digital Curation and Data Management 认证项目<sup>[18]</sup>、罗伯特高登大学 Digital Curation 培养方向<sup>[7]</sup>。从课程内容标签看,与常规课程不同的是,系列课程中出现最多的标签是 life-cycle,突出了生命周期在数据监护中的重要性。系

列课程是包括了 DC 理论课、DC 技术课、DC 应用课 3 类课程在内的具有连贯性和承接性的 DC 课程体系。DC 理论课不仅涵盖了常规 DC 课程中涉及到的数据监护的基本概念和理论,还深入探索了共创环境下多源数据融合问题、数据监护研究前沿问题;技术课主要涉及以下主题:底层数据的描述与表示,如不同学科领域元数据的设计与标准选取与元数据抽取技术、知识组织原理与技术,信息系统设计与开发包含了数据存储、数据检索、数据获取甚至是数据可视化呈现功能的信息检索系统的开发与设计,数据长期保存系统开发工具与应用;应用课主要是作为认证项目的结业课程,学生除了需要写 DC 相关主题的毕业论文外,还需完成 1 个 DC 实践项目,将所学的 DC 理论知识和技术知识应用到实际项目开发与项目管理中,在实践中深化对原理和技术的理解和运用。

### 2.3 科学数据管理生命周期不同阶段的 DC 课程主题分析

数据管理生命周期模型从不同角度描述了数据从产生、收集、描述、存储、发现、分析到再利用的整个生命周期<sup>[19]</sup>。本文结合数据监护课程特点,将数据管理生命周期划分为数据计划、数据处理、数据保存和数据利用 4 个阶段。将课程标签投射到数据管理生命周期理论的每一阶段,获得涉及数据管理生命周期 4 个不同阶段的课程数量(表 3)。

数据管理计划阶段作为科学数据管理生命周期中概念性的规划设计环节,对于后续的科学数据管理具有重要的指导作用。目前只有 1 个 iSchools 成员院校(英国罗伯特戈登大学 Project Management for Digital Curation)针对数据管理计划开设了独立的 DC 课程<sup>[7]</sup>,课程内容主要涉及数据管理生命周期每个阶段所包含活动的界定及项目管理政策。

数据处理阶段包括从数据创建、数据清洗与选择、数据描述与组织到数据分析的一系列数据处理过程。共有 12 门 DC 课程涉及数据处理阶段,其中涉及数据处理阶段最多的课程主题是数据的选择与描述。并不是所有的数据都需要进行保存和监管,只有具备真实性、准确性和完整性的有价值的数据才是长期保存和管理的对象。不同类型的数字资源具有不同的元数据描述标准,如何选择合适的元数

表 3 DC 课程内容手工标签-数据管理生命周期阶段映射表

DC 课程内容手工标签	数据管理生命周期阶段	对应 DC 课程数量
management strategy, key stages, curation lifecycle	数据计划阶段 ( data planning stage)	1
data creation, data selection, data appraisal, analysis, metrics, data provenance, assessment, data quality, quality control, data value, data cleaning, data pre-processing, XML, schema, data representation, organization, FRBR, ISAD(G), ontology, data acquisition, digital convergence, metadata standards, data modeling, semantic web, evaluation, visualization, statistical analysis, cluster analysis, R, data sources, policy	数据处理阶段 ( data processing stage)	12
long-term preservation, repositories, data retrieval, workflows, systems, policy Cyberinfrastructure, metadata, preservation, ICPSR, NOAA, NASA data centers, NCAR, storage, ingest, pre-ingest, data access, open access, accessibility, Dspace, Fedora Commons, Eprints, sustainability, conservation, availability, dissemination, OAI, audit, certification, hardware preservation, emulation, systems, maintenance, security, REDCap, DataVerse, Archivematica, privacy, intellectual property, PD 5454: 2012, PAS 198: 2012	数据保存阶段 ( data preserving stage)	24
data citation, data sharing, data publishing, data discovery, data use, reuse, scholarly publishing	数据利用阶段 ( data using stage)	19

据标准、数据格式、知识表示和知识组织技术以支持后续的数据的获取和利用,是 DC 课程关注的重点。

从表 3 可看出,涉及数据保存阶段的 DC 课程最多(24 门)。数据监护的初衷是保证有价值的科学数据在较长一段时间内能够得以保存,以便以后的研究人员能够复用这些数据进行科学研究。因此,数据的长期保存是数据监护最重要的环节,DC 课程的设置也说明了数据长期保存的重要性。通过分析这 24 门数据保存主题的 DC 课程内容标签,发现这些课程主要涉及了以下主题内容:数字资源长期保存面临的管理、技术、社会、经济上的挑战以及长期保存策略的制定;数字仓储和可信赖数字仓储的建立与认证,包括 Eprints, DSpace, Fedora, REDCap, DataVerse, Archivematica 等相应软件工具的介绍;选择合适的元数据类型与标准以支持数据的获取、管理和保存;保存数据的选择与评估;长期保存技术如数据更新、数据迁移、数据仿真中数据完整性和准确性的维持;数据长期保存中涉及的版权和数据隐私问题。

涉及数据利用阶段的 DC 课程数量排在第二位,共 19 门。数据利用包括数据访问与获取、数据复用以及数据出版。目前所有 iSchools 院校都不将涉及数据利用主题的 DC 课程列为一门独立课程,

而是与数据保存一起纳入同一门 DC 课程当中。数据利用主题的 DC 课程主要涉及以下主题内容:数据的访问与获取范围、条件、方式和流程等,数字出版或学术出版中涉及的元数据标准、数据格式 (XML)、出版工具。

### 3 结论

对科学数据的持续、高效监管为当前和未来的学术、科研和教育提供了有力支撑。LIS 学院作为长期以来培养信息管理和知识管理专业人员的摇篮,在培养数据管理和数据处理专业人才方面应该承担更多的责任。

#### 3.1 课程既有共同关注的主题内容,又有各自的侧重点

目前 iSchools 成员院校的 DC 课程普遍关注数据保存相关问题、数据管理生命周期、数据监护相关的标准和元数据问题,并且将现存的最佳实践作为案例研究,探索实际数据监护中面临的问题及挑战。不同国家的 iSchool 学院有各自偏好的课程主题。英国 iSchool 院校的 DC 课程强调对数字资源的质量管理以及适用于数字资源长期保存与长期可获取的信息系统的开发与设计;澳大利亚 iSchools 院校的 DC 课程重视对数据管理生命周期每个阶段的任务和活动的梳理,并且突出了大数据时代对科学数

据的管理而非宽泛的对数字资源的管理;加拿大 iSchools 院校的 DC 课程不仅十分重视 OAIS 模型及数据管理生命周期的学习和运用,而且对作为数字资源保存和利用的重要平台的机构库也给予充分关注;荷兰则更强调不同学科领域科学数据管理的需求及实践。

3.2 课程等级以常规课程为主,专业方向课程体系有待发展完善

目前常规 DC 课程所占比重较大,系列课程或者认证课程还处于起步阶段,开设的院校较少,反映出大部分 iSchools 学院的数据监护教育依然以基础概念和理论的梳理、最佳实践项目的介绍为主,只有少数学院尝试将数据监护设为一个新的专业方向。

3.3 课程内容涉及数据管理生命周期的每个阶段,重点应进行数据的长期保存

目前的 DC 课程对科学数据生命周期每个阶段的理论、标准、技术、方法和工具都有涉及,说明当前数据监护课程教学内容主题的范围较广。其中,数据的长期保存是数据监护最重要的环节。从课程数量来看,DC 课程的重点落在数据长期保存阶段的相关问题上,具体主题内容包括长期保存目前面临的挑战、元数据的选择、保存数据的选择与评估、长期保存技术与工具以及相关版权问题等。

## 4 讨论

国内武汉大学、南京大学、北京大学等高校已经意识到数据监护专业教育迫在眉睫,并在图书情报学院、信息资源管理学院等增设了与数据监护相关的课程,但目前还处于起步阶段<sup>[20]</sup>。国外 iSchools 院校数据监护教育的蓬勃开展以及对数据监护课程内容主题的探索,对于我国相关高校开设数据监护课程具有重要的启示和借鉴意义。

### 4.1 积极开展数据监护专业教育

2018 年 1 月,国家标准《科学数据引用》(GB/T 35294-2017)正式发布,标志着科学数据可以像学术论文一样被学术同行标准化引用,这将在一定程度上促进数据拥有者开放共享其数据。越来越多的开放科学数据需要专业的数据监护人员进行科学有效的管理,数据监护人才需求迫切。虽然从 2014 年起,中国图书馆学会专业图书馆分会联合中国科学院文献情报中心已经举办了 5 期科学数据管理研修

班,为各界培训具有科学数据素养的专业人员,但是研修班的规模小且学习时间短,许多内容都是浅尝辄止,培养的数据监护人才不足以满足社会的需求。因此,加强高等院校的数据监护专业教育迫在眉睫。相关院校也需要做好充分准备,积极开展数据监护专业教育。

### 4.2 注重基础理论学习与最佳实践研究的融合

目前,数据监护领域已经形成了一些较为成熟的理论、模型和元数据标准以及较大规模的科学数据管理与共享平台,如 GBIF (Global Biodiversity Information Facility), DataOne, Dryad。相关院校在开设数据监护课程时,既要注重基础理论的学习,也要将现存的最佳实践作为案例研究,探索实际工作中面临的问题、挑战及其解决方法。

### 4.3 充分利用 LIS 学院的传统优势课程,完善数据监护专业课程体系

信息组织与信息描述是 LIS 学院的传统优势课程,可将其与数据监护领域中涉及的数据组织、数据描述进行无缝接轨,纳入数据监护专业课程体系之中。

### 4.4 重视机构库建库软件的学习和利用

机构库和可信赖数字仓储 (TDR) 是数据监护领域的最佳保存实践项目。机构库和 TDR 的开发依赖于目前流行的开源软件,包括 Eprints, DSpace, Fedora, REDCap, DataVerse, Archivematica 等。不同的开源软件各有其侧重点和优缺点,只有对主流开源软件有了充分的了解,在实际工作中才能够根据自身需求和实际条件,选择合适的软件构建机构库。

## 【参考文献】

- [1] 杨鹤林. 英国数据监护研究成果及其在高校图书馆的应用: DCC 建设回顾[J]. 图书馆杂志, 2014, 33(3): 84-90.
- [2] 高珊, 卢志国. 国外数据馆员的能力需求与职业教育研究[J]. 图书馆, 2015(2): 65-69.
- [3] Harris-Pierce RL, Liu YQ. Is data curation education at library and information science schools in North America adequate? [J]. New Library World, 2012, 113(11/12): 598-613.
- [4] 黄如花, 吉翠芳. 伊利诺伊香槟大学数据管理教育现状及启示[J]. 图书与情报, 2015(1): 61-65.
- [5] 周霞, 赵静. 情报学硕士课程设置研究: 我国情报学硕士企业招聘的反思[J]. 情报杂志, 2015, 34(8): 26-30.
- [6] University of Glasgow. Humanities Advanced Technology and Information Institute. Management, Curation & Preservation of Digital



- Materials ARTMED5021 [EB/OL]. [2018-02-05]. <https://www.gla.ac.uk/coursecatalogue/course/?code=ARTMED5021>.
- [7] Robert Gordon University. Digital Curation Modules & Overview [EB/OL]. [2018-02-05]. <http://www.rgu.ac.uk/file/digital-curation-module-overview>.
- [8] Charles Sturt University; School of Information Studies. INF462 Data Curation [EB/OL]. [2018-02-05]. <http://www.csu.edu.au/handbook/subjects/INF462.html>.
- [9] University of Melbourne; Melbourne School of Information. E-Science (SCIE90007) [EB/OL]. [2018-02-05]. <https://handbook.unimelb.edu.au/2017/subjects/scie90007>.
- [10] University of Toronto; Faculty of Information. Digital Preservation and Curation [EB/OL]. [2018-02-05]. <https://ischool.utoronto.ca/course/digital-preservation-and-curation/>.
- [11] McGill University, Montreal; School of Information Studies. Courses [EB/OL]. [2018-02-05]. <http://www.mcgill.ca/sis/courses>.
- [12] University of Amsterdam. RESEARCH DATA MANAGEMENT FOR INFORMATION PROFESSIONALS [EB/OL]. [2018-02-05]. <http://slais.ubc.ca/arst556k/>.
- [13] University of Tennessee. INSC562; DIGITAL CURATION [EB/OL]. [2018-02-05]. <http://www.sis.utk.edu/sites/default/files/syllabus/INSC562syl.pdf>.
- [14] University College London. INSTG064 - Introduction to Digital Curation [EB/OL]. [2018-02-05]. <http://www.ucl.ac.uk/dis/study/pg/INSTG064>.
- [15] University of California. Graduate Courses [EB/OL]. [2018-02-05]. <http://www.registrar.ucla.edu/Academics/Course-Descriptions/Course-Details?SA=INF+STD&funsel=3>.
- [16] Drexel University. Courses [EB/OL]. [2018-02-05]. <http://catalog.drexel.edu/coursedescriptions/quarter/grad/info/>.
- [17] University of North Carolina at Chapel Hill. Certificate in Digital Curation [EB/OL]. [2018-02-05]. [https://sils.unc.edu/programs/certificates/digital\\_curation](https://sils.unc.edu/programs/certificates/digital_curation).
- [18] University of North Texas. Digital Curation and Data Management [EB/OL]. [2018-02-05]. <http://informationscience.unt.edu/digital-curation-and-data-management>.
- [19] 杨林, 钱庆, 吴思竹. 科学数据管理生命周期模型比较 [J]. 中华医学图书情报杂志, 2016, 25(11): 1-6.
- [20] 湛爱容. 国外数据馆员培训实践及其启示 [J]. 大学图书馆学报, 2018, 26(1): 75-82.

[收稿日期: 2018-02-16]

[本文编辑: 刘娜]