DOI：10.16353/j.cnki.1000-7490.2016.07.024

● 徐　扬[1,2]，王申罡[1]
（1. 北京大学　信息管理系，北京　100871；2. 北京大学　一带一路研究中心，北京　100871）

数据起源研究进展*

摘　要：随着数据科学的兴起，数据已成为一种重要战略资源。在大数据的环境下，数据正以前所未有的速度不断地增长、累积、流动、复制和分发，由于数据处于动态演化的过程，验证原始出处变得更为困难，更容易导致数据质量问题。因此，加强数据起源的研究势在必行，文章通过对数据起源相关概念、含义、模型、方法、技术、系统和应用等方面的系统梳理，对其进行综述研究，并指出在大数据环境下对于移动端与物联网中的数据起源问题的研究是一个极具挑战与重大现实意义的科学问题。

关键词：数据起源；研究进展；大数据

Abstract: With the fast development of data science, data becomes a key strategic resource. In the environment of big data, data is increasing, accumulating, sharing, duplicating and distributing faster than ever. Since data is always in a dynamic process, how to determine the origin of data becomes more difficult, which is easily leads to data quality problems. As a result, it is necessary to carry on the research on data provenance. Based on the analysis of concept, meaning, model, method, technology, system and application of data provenance, this paper carries on the literature review of data provenance, and points out that the study of data provenance issues in mobile terminal and internet of things under the big data era is a scientific problem of high challenges and significance.

Keywords: data provenance; research progress; big data

随着数据科学的兴起，人们开始越来越多地通过各种途径获取自己所需要的数据，数据已成为一种重要战略资源，对数据的利用成为核心竞争力，对情报学的发展具有重要意义[1]。在大数据的环境下，数据呈现出不同于以往的新特点：数据正以前所未有的速度不断地增长和累积，数据流动加速，数据的复制与分发变得容易[2]。在这样的背景下，数据处于动态演化的过程，验证原始出处较为困难，容易导致数据质量参差不齐，甚至出现数据造假[3]。在大数据时代，各个领域数据起源的重要性日益凸显，影响用户对所获取数据的信任程度，特别是在开放存取（Open Access）背景之下，数据起源在科学研究中的重要性体现得尤为明显[4]。同时，e-Science 是科学研究在信息时代出现的新模式，使得全球的科研资源连接为一个互通的网络，科学家不受时间和地域限制地使用其资源[5]。开放存取和 e-Science 等信息技术的发展极大程度上解决了数据获取的问题，但是如果数据质量、可靠性、可信度无法保障，极有可能会发生 GIGO（Garbage In，Garbage Out）现象[6]。因此，为了保证数据的可靠性和质量，科研人员需要知道相关数据的起源信息。

数据起源在科学研究与生产生活的各个方面都具有重要意义，其主要作用包括评估数据质量与可信度；查询数据来源，必要时可进行数据来源的审计跟踪；定位数据派生过程中的错误，分析错误原因，确定责任人；实现数据生成过程的重演，重构数据或者实验过程，有利于数据共享和流程优化；管理数据的版权与知识产权。简而言之，数据起源的用途主要集中在数据质量评价、审计跟踪、权属证明、数据恢复、数据复用等方面[7]。

1　数据起源相关文献计量与分析

科学知识量的增长及其规律与科学文献的增长及其规律是紧密联系的，科学文献的数量变化直接反映了科学知识量的变化情况，因此科学文献的数量是衡量科学知识量的重要尺度之一[8]。本文从国内外权威科学文献数据库中对数据起源相关文献进行了检索，结果如表 1 所示。

检索结果表明，自 20 世纪 90 年代以来，数据起源相关研究在国际上已经兴起，并出现了 International Provenance and Annotation Workshop，Workshop on Data Derivation and Provenance，Workshop on the Theory and Practice of Provenance 等直接相关的学术会议。相较于国外，国内的

* 本文为国家社会科学基金项目“大规模个性化定制环境下的情报系统研究”的成果，项目编号：13CTQ022。

数据起源相关研究则起步较晚。

表1　数据起源相关文献检索结果

	中文数据库	外文数据库
统计来源	中国知网（CNKI）	Web of Knowledge 核心合集
时间范围	2014年（含）之前	1990—2014
检索条件（“篇名”或“关键词”中包含的词汇）	数据起源 数据溯源 数据世系	data lineage; data provenance; data pedigree; data derivation
文献数量	34（从检索到的37篇文献中删除其中不相关的3篇）	349

表2列出了2005—2014年数据起源领域研究论文的详细数量。

表2　2005—2014年数据起源领域论文收录情况

		2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
中文数据库	A	1	0	1	1	0	9	7	6	6	3
	B	1	1	2	3	3	12	19	25	31	34
外文数据库	A	18	15	17	25	33	44	26	45	53	23
	B	18	33	50	75	108	152	178	223	276	299

注：A表示当年论文数，B代表论文累积量。

表2显示，中文数据库与外文数据库收录的数据起源领域研究论文的数量基本呈增长趋势，这表明数据起源研究领域已吸引了越来越多国内外学者的关注，而在论文收录数量上，外文数据库明显多于中文数据库。据此绘制的论文增长趋势比较图如图1所示。可以看出，该领域知识增长较快。

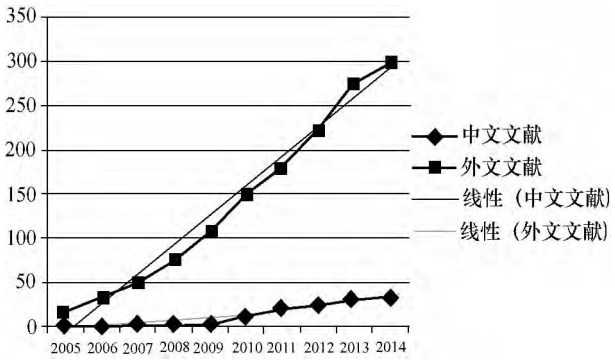

图1　2005—2014年数据起源领域论文收录情况趋势图

数据起源研究涉及诸多学科与研究领域，这也反映了其学科交叉的研究特点。外文数据库中数据起源相关文献主要研究领域分布如表3所示。

由表3可知，在外文数据库的数据起源相关文献中，研究领域以计算机科学与工程学为主，同时涉及通信、遥感、生物、数学、天文与天体物理学等多个研究方向。与之相比，中文相关文献研究方向较为单一，主要集中在计算机科学，而其他研究方向几无涉及，这也反映出中文文献在研究深度和广度上与外文文献的差距。

表3　外文数据库中数据起源相关文献主要研究领域分布

研究领域	数量	比例(%)
计算机科学（Computer Science）	238	68.19
工程（Engineering）	73	20.92
电信（Telecommunication）	18	5.16
遥感（Remote Sensing）	15	4.30
生物化学分子生物学（Biochemistry Molecular Biology）	14	4.01
图像科学摄影技术（Imaging Science Photographic Technology）	14	4.01
数学计算生物学（Mathematical Computational Biology）	13	3.72
数学（Mathematics）	12	3.44
天文学（Astronomy Astrophysics）	11	3.15
地质学（Geology）	10	2.87

2　数据起源的概念及相关研究

数据起源由“Data Provenance”翻译而来，同义表述有“Data Lineage”“Data Derivation”“Data Pedigree”等。除“数据起源”外，国内学者在相关研究论文中也使用到“数据世系”“数据溯源”等其他译名。截至目前，关于数据起源尚未达成统一的定义，不同学者基于各自的研究视角赋予其不同的含义。Buneman等将其定义为跟踪并记录数据来源及其在数据库间的移动轨迹，并指出数据起源的两重含义：why-provenance（数据因哪些源数据而产生）与where-provenance（数据从哪些源数据复制而来）[9-10]。Frew和Bose基于数据起源开发了数据管理基础架构，用于存储和发布数据产品[11]。Glavic和Dittrich重点关注电子科学和数据仓储中的数据起源问题，其构造的分类模型可以区分不同的数据出处类型[12]。Braun等将数据起源表示为含有标注的因果图，以节点为实体，连线为实体之间的因果关系[13]。Kondylakis等利用一个简单的映射架构构造本体和关系数据源，并给出了3种算法用以检索起源信息[14]。Ram和Liu认为，起源信息需要被追踪、捕捉才能使其在未来得到充分的开发与利用，然而由于在其语义或含义上达成的共识有限，导致起源信息本身的不完整性造成起源信息无法被跨应用或系统分享，针对这一问题，他们开创性地提出了一个阐明数据起源语义的本体模型，W7模型，其将起源定义为数据的n元组的集合：$p(D)=\{<What, When, Where, How, Who, Which, Why>\}$[15]。从数据起源的概念及相关定义的发展历程来看，数据起源已不局限于单一领域，需全面、多视角、以发展

的眼光认知其内涵。虽然目前业界尚未形成严格的、统一的数据起源的概念或定义，但上述研究成果极大程度上加强了人们对数据起源的认识。

3 数据起源相关模型

3.1 开放起源模型 OPM

数据起源模型研究一直是学者关注的重点，尤其是系统性能、起源方法表示的等内容，起源挑战（Provenance Challenge，PC）问题应运而生[16]。2007 年，Moreau 等对业界成果进行了整理并加以形式化，提出了开放起源模型（Open Provenance Model，OPM），并介绍了基于 OPM 的起源模型与抽象语法表示[17]。OPM 设计的目标是为不同系统提供可交换的起源信息，并允许开发人员创建并共享操作该模型的工具。

本质上，OPM 是一个描述事物间关系的图，包括 Artifact（实体）、Process（过程）和 Agent（代理）3 类节点。Artifact 代表一个状态，该状态对应一个物理实体对象或者计算机系统中的数据对象；Process 是一个或多个施加在数据对象上的行为或操作，进而产生新版本的数据对象，同时产生对应的状态节点；Agent 是发起或控制过程的个人或机构。三类节点之间存在 5 种关系，分别是 wasGeneratedBy（生成）、used（使用）、wasControlledBy（控制）、wasDerivedFrom（获得）、wasTriggeredBy（触发）。将 OPM 应用到实践中的包括 OPM Toolbox，Tupelo，Taverna，Provenance JS，VisTrails，Swift，eBioFlow，Karma，ourSpaces，OPMProv，Kepler Provenance Listener，PLIER（Provenance Layer Infrastructure for e-Science Resources），Living Knowledge，WebN + 1，SPADE 等众多系统[18]。

3.2 基于本体的起源模型

这类数据起源模型主要包括如下几种:

1）Provenir 模型。它是一种基于本体的方法描述起源信息的模型，定义了顶层 Data（数据）、Process（过程）、Agent（代理）三大基本类及相应五大子类，并且定义了它们之间的关系[19]。这些类与类之间的关系构成了模型的框架，对模型进行逻辑描述。该模型在生命科学、海洋、传感器和卫生保健等领域得到了广泛应用。

2）CRMdig 模型。它是一个以事件为中心的本体模型，是对本体 CIDOC-CRM 的扩展，能捕获 e-Science 环境中和数字对象起源相关的建模和查询需求，提供了丰富的术语以描述科学数据产生过程的相关物理环境[20]。

3）安全起源模型。Park 等指出，系统中数据起源信息至少引发了两个与安全相关的问题，一是如何利用数据起源信息加强系统的安全层级，二是如何保护那些可能比数据还要敏感的起源信息以防止其被破坏。他们提出了一个基于数据起源的存取控制模型，实现了利用起源信息保障系统安全的目的[21]。

4）基于起源的数据质量评估模型。依靠数据做决策，数据必须具有良好的数据质量和可信度。数据质量相关研究由来已久，其与数据起源的研究具有很大的相关性，基于起源的视角，不少有效的数据质量评估模型或方法被提出。Dai 等提出了一个数据起源信任模型，设计了信任分值计算算法并验证了其有效性[22]。Hartig 与 Zhao 提出了一种网络数据起源模型和利用网络数据的起源信息评价数据质量和可信度的方法，可用于给定数据质量指标（如及时性、准确性等）的评价[23]。

5）流起源信息模型。Vijayakumar 和 Plale 对复杂事件处理系统中的流起源问题进行了研究，并构造了一个由若干实体构成的流起源信息模型，实体间紧密联系，依据不规则事件的起源历史时间构成起源图，进而推断数据起源[24]。

6）时间—值中心起源模型。该模型将基于标注的起源模型与基于过程的起源模型优点进行了有效结合，使其具有良好的描述能力，同时降低了存储与处理的成本[25]。

7）四维起源模型。在 Simmhan 等介绍的四维起源模型中，数据起源被看成一系列离散的活动集，这些活动发生在整个工作流生命周期中，并由活动执行层次的层次维、工作流组件所在位置的空间维、活动发生的时间维、活动过程中数据产品产生和消费的数据流维 4 个维度组成[26]。

8）生物学嫁接模型。将生物学领域中一些恰当的研究方法与模型应用到数据科学中，往往会取得令人满意效果。奚建清等参照生物进化论的观点进行了数据起源建模，定义了数据基因、基因序列、数据基因组等基本概念及相互间的关系，并通过数据基因的遗传和变异来记录数据的特征及与其他数据的关系[27]。陈颖借鉴 DNA 双螺旋结构进行数据起源建模，运用 DNA 的半保留复制、DNA 修复等知识对模型原理、特点及意义进行说明，为数据起源的研究提供了一种新的思路与方法[28]。

从宏观上看，上述几种模型都与数据起源相关，但就其侧重点而言，却存在差异。本文将数据起源模型大致分为以下 4 个层次:

第一层次，以 OPM 与基于本体的起源模型为代表的底层与框架模型。

第二层次，面向起源信息本身的安全防护模型，也是安全起源模型的重要组成部分。

第三层次，以基于起源的数据质量评估模型、时间—值模型等为代表的起源功能实现模型。

第四层次，包括生物学嫁接模型的其他起源模型。

在数据起源模型的4个层次之中，第一层次作为基础型研究，其研究意义与影响力最为重大，如OPM，Provenir等模型在业界均有很大的影响力和很强的认可度。第二层次模型主要为功能实现提供有效保障。第三层次以前两个层次为基础，脱离了第一层次的支持会导致其缺乏理论依据，若忽视第二层次，会直接影响其功能实现的程度。第四层次可看作是学科交叉融合背景下对起源模型的发展与补充，可以拓展研究思路、丰富研究成果。

4 数据起源方法与技术

Simmhan等根据起源技术应用的数据质量、审计追踪等5个不同方面，从概念层面对技术进行了分类，并简单地比较了各自的优缺点[29]。基于信息生命周期机制，则可将数据起源方法与技术分为应用于收集、存储、验证（计算与分析）与转移等不同阶段的方法与技术，其中关于存储与计算这两阶段的方法与技术的讨论较多。目前，用于实现数据起源的追踪与溯源的方法与技术有标注法和逆查询法。

标注法是一种简单有效的数据起源方法，常用来记录注释、声明等关于数据的辅助信息，以供使用者共享[30]。标注用于数据起源，即在标注中记录关于数据出处、作者、时间及其演变历史等重要信息，并使标注可与数据一起传播，用户可直接通过查看目标数据的标注来获得数据的起源信息。

逆查询是另一种数据起源的应用方法，通过逆向查询或构造逆向函数对查询求逆，由结果追溯到源数据，一般在需要用到数据起源时才进行计算[31]。该方法在早期将数据起源用于视图维护与更新问题时被提出，如Cui等人研究了SPJ（Select-Project-Join，选择—投影—连接）视图、ASPJ（Aggregate-Project-Select-Join，聚合—选择—投影—连接）视图和更复杂的带有集合操作ASPJ视图等3个不同类型的视图的计算起源的方法[32]。表4对标注法和逆查询法进行了比较。

表4 标注法与逆查询法比较

	优点	缺点
标注法	易于实现、易于管理	适合小型系统，而难以为大型系统的细粒度数据提供详细的数据起源信息；需要较大的存储开销和查询语言的支持
逆查询法	追踪简单，不需要存储额外的标注信息，只需少量的元数据就能实现对数据起源的追踪	实现相对复杂，用户需要提供逆置函数和相对应的验证函数，计算也会比较耗时

此外，双向指针追踪法、利用图论思想和专用查询语言追踪法以及以位向量存储定位等方法也较为常见。

5 数据起源在系统中的应用与实现

基于相关模型、方法和技术，数据起源已在诸多系统中有所应用与实现，典型系统如表5所示。

表5 数据起源典型系统

系统名称	系统描述
DBNote[33]	通过标注方式存储与管理数据并支持标注传播
Trio[34]	基于传统的关系数据库，将数据的不确定性与起源信息与数据本身作为顶层概念
Panda[35]	整合基于数据与基于过程的两种类型的数据起源，实现一个通用的集数据起源获取、存储、操作与查询于一体的开源系统
Kepler[36]	为科学家提供了一个方便易用的工作平台，通过记录工作流执行状态重现整个操作的全过程，实现工作流的创建、运行和共享一体化，通过跟踪数据项以及数据集合的历史记录，将结果反馈给用户
Galaxy[37]	支持生命科学领域中易获取的、可复制的、透明的计算研究，自动追踪、管理数据起源，为捕捉情境与计算方法的使用意图提供支持
CMCS[38]	管理异构数据流与跨领域学科的元数据，为多层次科学研究提供协作和基于元数据的数据管理功能
Perm[39]	能够实现起源信息计算、存储与查询等功能，其中起源计算通过使用查询复写技术标注带有起源信息的元组实现
myGrid[40]	提供一套网格中间件，其服务包括资源发现、工作流设置、元数据与数据起源管理，可以实现信息集成，帮助解决语义的复杂性
Taverna[41]	由可用服务面板、工作流图面板和高级模型浏览器三个主要部分构成，在底层实时获取工作流执行的信息，而这些原始的起源信息进一步通过结构性注释后存入关系数据库
PASS[42]	面向文件或文件系统并在统一环境下自动采集、存储、管理与查询起源信息的存储系统

上述分析显示，数据起源系统主要实现了在数据库中的应用和工作流中的应用。同时，虽然数据起源系统涉及学科较多，但主要集中在生物、化学和计算机等对数据处理要求较高的自然科学。随着大数据思想和技术不断深入人们生产生活的各个方面，数据起源系统将在更大的范围内得到广泛应用。

6 结束语

兴起于20世纪90年代的数据起源研究经过20余年的发展，以其重要的理论价值与现实意义得到了学者们的充分重视，尤其是在大数据广泛应用的背景下，其研究显得尤为关键。

本文从相关概念、含义、模型、方法、技术、系统和应用等方面对数据起源的研究现状及其突出成果进行了系

统的梳理。通过综述研究，发现数据起源虽然已经取得了较为丰硕的研究成果，但仍有一些问题有待进一步解决，包括统一的业界标准仍需不断发展与完善、可视化技术在数据起源系统中的应用有待进一步加强。

尤其值得关注的是，作为新一代信息技术的重要组成部分，移动互联网技术与物联网技术使得越来越多的数据（如个人数据、用户数据等）产生于移动设备（如智能手环、智能手表等），因此在大数据环境下对于移动端与物联网中的数据起源问题的研究是一个极具挑战与重大现实意义的科学问题。□

参考文献

[1] 化柏林，李广建. 大数据环境下的多源融合型竞争情报研究 [J]. 情报理论与实践，2015，38（4）：1-5.

[2] 孟小峰，慈祥. 大数据管理：概念、技术与挑战 [J]. 计算机研究与发展，2013（1）：146-169.

[3] HASAN R，SION R，WINSLETT M. Preventing history forgery with secure provenance [J]. ACM Transactions on Storage，2009，5（4）：1-14.

[4] 范亚芳，高中华. 开放存取资源整合服务模式研究 [J]. 情报理论与实践，2009，32（1）：66-69.

[5] 秦长江. E-Science（科研信息化）对现代科学的影响 [J]. 科技进步与对策，2008（8）：143-145.

[6] SIMMHAN Y L，PLALE B，GANNON D. A survey of data provenance in e-science [J]. ACM SIGMOD Record，2005，34（3）：31-36.

[7] LI J，CHEN X，HUANG Q，WONG D. Digital provenance：Enabling secure data forensics in cloud computing [J]. Future Generation Computer Systems，2014，37：259-266.

[8] 邱均平，苏金燕，熊尊妍. 基于文献计量的国内外信息资源管理研究比较分析 [J]. 中国图书馆学报，2008（5）：37-45.

[9] BUNEMAN P，KHANNA S，TAN W C. Data provenance：some basic issues [J]. Lecture Notes in Computer Science，2000，1974：87-93.

[10] BUNEMAN P，KHANNA S，TAN W C. Why and where：a characterization of data provenance [J]. Lecture Notes in Computer Science，2001，1973：316-330.

[11] FREW J，BOSE R. Earth system science workbench：a data management infrastructure for earth science products [C] //Proceedings of the 13th International Conference on Scientific and Statistical Database Management，2001：180-189.

[12] GLAVIC B，DITTRICH K. Data provenance：a categorization of existing approaches [C] //Proceedings of Datenbanksysteme in Business，Technologie und Web，2007：227-241.

[13] BRAUN U，SHINNAR A，SELTZER M. Securing provenance [C] //Proceedings of the 3rd Conference on Hot Topics in Security，2008，Article 4.

[14] KONDYLAKIS H，DOERR M，PLEXOUSAKISl D. Empowering provenance in data integration [J]. Lecture Notes in Computer Science，2009，5739：270-285.

[15] RAM S，LIU J. A new perspective on semantics of data provenance [C] //Proceedings of the 1st International Workshop on the role of Semantic Web in Provenance Management，collocated with the 8th International Semantic Web Conference，2009.

[16] SIMMHAN Y，GROTH P，MOREAU L. Special section：the third provenance challenge on using the open provenance model for interoperability [J]. Future Generation Computer Systems，2011，27（6）：737-742.

[17] MOREAU L，CLIFFORD B，FREIRE J，et al. The open provenance model core specification [J]. Future Generation Computer Systems，2011，27（6）：743-756.

[18] The OPM Provenance Model（OPM）[EB/OL]. [2015-10-28]. http：//openprovenance. org/.

[19] SAHOO S S，BARGA R S，GOLDSTEIN J，SHETH A. Provenance algebra and materialized view-based provenance management [C] //Proceedings of the 2nd International Provenance and Annotation Workshop，2008：531-540.

[20] DOERR M，THEODORIDOU M. CRM_{dig}：a generic digital provenance model for scientific observation [EB/OL]. [2015-10-01]. http：//cidoc-crm. org/docs/CRMdig-TAPP11. pdf.

[21] PARK J，NGUYEN D，SANDHU R. A provenance-based access control model [C]. IEEE Proceedings of the 10th Annual International Conference on Privacy，Security and Trust（PST），2012：137-144.

[22] DAI C Y，LIN D，BERTINO E，KANTARCIOGLU M. An approach to evaluate data trustworthiness based on data provenance [J]. Lecture Notes in Computer Science，2008，5159：82-98.

[23] HARTIG O，et al. Using web data provenance for quality assessment [C] //Proceedings of the International Workshop on the Role of Semantic Web in Provenance Management，2009.

[24] VIJAYAKUMAR N N，PLALE B. Tracking stream provenance in complex event processing systems for workflow-driven computing [C]. Second International Workshop on Event-driven Architecture，Processing，and Systems，2007.

[25] WANG M，BLOUNT M，DAVIS J，MISRA A，et al. A time-and-value centric provenance model and architecture for medical event streams [C] //Proceedings of the 1st ACM SIGMOBILE International Workshop on Systems and Networking Support for Healthcare and Assisted Living Environments，2007.

（下转第 135 页）

[8] 周军杰，左美云．线上线下互动、群体分化与知识共享的关系研究——基于虚拟社区的实证分析 [J]．中国管理科学，2012，20（6）：186-192.
[9] 宁连举，刘自慧，冯鑫．虚拟社区零售中信息搜索数量的影响机制研究 [J]．北京工商大学学报：社会科学版，2013，28（5）：55-61.
[10] 钱坤，孙锐．用户参与虚拟社区中产品创新的影响因素研究——扎根理论研究方法的运用 [J]．科技管理研究，2014（6）：6-10.
[11] 熊回香，张晨，李玉波．基于 Web 3.0 的个人知识管理平台建设研究 [J]．图书情报工作，2010，54（18）：95-99.
[12] 胡海波．Web 3.0 环境下基于用户兴趣的信息集聚服务 [J]．情报理论与实践，2014，37（8）：117-121.
[13] 吴一平．基于 Web3.0 思想的图书馆 3.0 服务新模式 [J]．情报杂志，2010，29（1）：244-247.
[14] 马振萍，杨姗媛．基于 Web3.0 的网络信息交流模式 [J]．情报资料工作，2011（1）：61-64.
[15] 吴胜，苏琴．Web3.0 数据整合流程研究 [J]．图书情报工作，2011，55（24）：112-115.
[16] 刘萍，郑凯伦，邹德安．基于 LDA 模型的科研合作推荐研究 [J]．情报理论与实践，2015，38（9）：79-85.
[17] 周永红，宫春梅，陈思．企业联盟知识共享网络层次及其演化 [J]．情报理论与实践，2015，38（9）：60-63.
[18] 刘琼，任树怀．论 Web3.0 下的信息共享空间 [J]．图书馆，2011（2）：83-85.
[19] 杨巧云，姚乐野．基于协调理论的应急情报部门跨组织工作流程研究 [J]．情报理论与实践，2015，38（8）：75-84.
[20] MALONE T W，CROWSTON K G. The interdisciplinary study of coordination [J]. ACM Computing Surveys，1994，26（1）：87-119.
[21] 熊莉君．虚拟社区中信息交流的引导机制研究 [J]．图书馆学研究，2011（9）：45-47.
[22] 王众托．无处不在的网络社会中的知识网络 [J]．信息系统学报，2007（1）：1-7.
[23] 陈浩义，孙丽艳，王文彦．产业集群中技术创新信息流动模式及进化机理研究 [J]．情报理论与实践，2015，38（5）：47-50.
[24] 熊回香，陈姗，许颖颖．基于 Web 3.0 的个性化信息聚合技术研究 [J]．情报理论与实践，2011，34（8）：95-99.

作者简介：汪传雷，男，1970 年生，博士后，教授。研究方向：信息资源管理，物流管理与工程。
朱绍平，男，1992 年生，硕士生。研究方向：物流信息管理。
万一荻，女，1994 年生，硕士生。研究方向：物流技术经济及管理。
蒋孝成，男，1990 年生，硕士生。研究方向：物流技术经济及管理。

收稿日期：2016－01－24

（上接第 140 页）

[26] SIMMHAN Y L，PLALE B，GANNON D. A framework for collecting provenance in data-centric scientific workflows [C] // Proceedings of the IEEE International Conference on Web Services，2006：427-436.
[27] 奚建清，郭玉彬，汤德佑．数据基因：数据的进化过程管理模型 [J]．计算机科学，2007（1）：12-16.
[28] 陈颖．一种基于 DNA 双螺旋结构的数据起源模型 [J]．现代图书情报技术，2008（10）：11-15.
[29] SIMMHAN Y L，PLALE B，GANNON D. A survey of data provenance techniques [EB/OL]．[2015-10-08]．http：//www.cs.indiana.edu/pub/techreports/TR618.pdf.
[30] FREIRE J，KOOP D，et al. Provenance and annotation of data and processes [M]．Springer，Heidelberg，2008.
[31] 刘喜平，万常选．数据起源研究综述 [J]．科技广场，2005（1）：47-52.
[32] CUI Y，WIDOM J，WIENER J L. Tracing the lineage of view data in a warehousing environment [J]．ACM Transactions on Database Systems，2000，25（2）：179-227.
[33] CHITICARIU L，TAN W C，VIJAYVARGIYA G. DBNotes：a post-it system for relational databases based on provenance [C] // Proceedings of the ACM SIGMOD International Conference on Management of Data，Baltimore，Maryland，USA，2005：942-944.
[34] WIDOM J. Trio：a system for integrated management of data，accuracy，and lineage [EB/OL]．[2015-10-10]．http：//www-db.cs.wisc.edu/cidr/cidr2005/papers/P22.pdf.
[35] PandaWhale [EB/OL]．[2015-10-10]．http://pandawhale.com/post/8327/stanford-panda-project.
[36] Kepler [EB/OL]．[2015-10-12]．https：//kepler-project.org/.
[37] Galaxy [EB/OL]．[2015-10-12]．https：//usegalaxy.org/.
[38] MYERS J D，ALLISON T C，BITTNER S，et al. A collaborative informatics infrastructure for multi-scale science [J]．Cluster Computing，2005，8（4）：243-253.
[39] GLAVIC B，ALONSO G. The perm provenance management system in action [C] // Proceedings of the ACM SIGMOD International Conference on Management of Data，Providence，Rhode Island，USA，2009：1055-1058.
[40] STEVENS R D，ROBINSON A J，GOBLE C A. Grid：personalised bioinformatics on the information grid [J]．Bioinformatics，2003，19（suppl 1）：i302-i304.
[41] OINN T，ADDIS M，FERRIS J，et al. Taverna：a tool for the composition and enactment of bioinformatics workflows [J]．Bioinformatics，2004，20（17）：3045-3054.
[42] PASS [EB/OL]．[2015-10-16]．http://www.eecs.harvard.edu/syrah/pass.

作者简介：徐扬，男，1981 年生，博士，副教授，研究员。研究方向：情报分析。
王申罡，男，1992 年生，硕士生。

收稿日期：2015－12－24