中国科技资源导刊 ISSN 1674-1544 CHINA SCIENCE & TECHNOLOGY RESOURCES REVIEW
2009年7月 第41卷第4期 28-33 ISSN1674-1544 Vol 41 No 4 28-33,Jul 2009

科学数据产品的质量管理基础

Nicole M. Radziwill[1],苏 颖[2]
(1. 美国国家射电天文台,弗吉尼亚州,22901,美国; 2. 中国科学技术信息研究所,北京 100038)

摘 要:科学数据的质量越高,人们可以越快获得更准确的结论,社会更容易从中获益。要改进科学数据的质量,必须清楚地了解科学数据的性质及其产生过程。本文以科学决策过程为背景,提出数据产品和数据质量的定义,其中假设了两种典型的情景:收集观测数据及进行以文献为基础的研究。然后分析与全面质量管理(TQM)理念相关的两个延伸学科——全面信息质量管理(TIQM)和全面数据质量管理(TDQM),以确定科学数据质量的管理是否有别于其他数据和信息的管理。本文提出规划、评估/保障、控制和持续改进的建议,重点放在将质量设计到生产流程中去,而不是依赖大量的检验。

关键词:数据中心;数据管理;数据质量;科学数据;符号学;TDQM;TIQM;TQM
中图分类号: TG156 **文献标识码**: A DOI: 10.3772/j.issn.1674-1544.2009.04.006

1 引 言

在生成科学数据产品(Scientific Data Product, SDP)时,对质量进行管理尤为重要,因为数据的使用可能对组织的效率、未来的科学研究或者社会产生极其广泛的影响。

作为"气候的社会影响"项目的一部分,美国社会与环境研究所估算了由极端天气导致的经济损失每周高达3亿美元。准确的天气预报能使每台石油钻机每天节约25万美元。每避免一次航班取消能够节约大约4万美元,每避免一次航线改变可以节约大约15万美元。准确度欠佳的天气预报也会对航空公司造成经济损失。上述的各种情景都要求具备高质量的原始数据并结合准确而周密的方法以得出天气状况的合理模型预测结果。

国家科学基金会(NSF)的战略目标之一是实现科学思想的交叉孕育,其方法是允许某一领域的非专业人员研究可能需要其专业知识之外的另一领域的数据集。这些交叉领域的研究人员对于数据产品质量的微妙差异不那么敏感,因此,为了实现这种工作方法,需要对数据集进行更严格的管理。

类似流程行业,科学数据的收集和生成通常受环境条件的影响,其中一些环境条件是可以控制的。所有科学数据产品的生产设施都是通过提供高级组件的定制工程化而获得竞争优势的,而定制通常又会妨碍流程的协作和标准化,从而对质量管理实践产生自然抵触。产生的科学数据为其他的组装线或信息处理程序提供了原始数据[1]。

然而在实际过程中重视科学数据产品及其生产流程质量,忽视科学数据管理。科学数据中心的出现对科学数据产品提出了更高的要求,质量管理将成为此生产流程中重要的一个环节。要实现这一点的前提是必须先在此领域内建立理论基础,为数据质量管理提供切实可行的解决方案。当前研究的目标就是确定数据质量管理的基本流程。

第一作者简介:Nicole M. Radziwill(1966-),美国国家射电天文台(NRAO)软件开发部副主任,高级研究员,主要研究方向是源自信息技术的软件开发和流程改进。
基金项目:国家自然科学基金资助项目(70772021,70831003);中国博士后科学基金资助项目(20060400077)
收稿日期:2009年7月2日。

2 美国国家射电天文台的数据质量

作为一个小规模软件质量诊断项目的一部分，我们利用一年半的时间对美国国家射电天文台(NRAO)位于弗吉尼亚 Green Bank 的 Robert C. Byrd Green Bank 望远镜的科学数据和数据质量的性质进行了跟踪研究，以探查其质量的驱动因素。为了便于对比,本文作者利用位于南达科他州拉皮德城的大气科学研究所在气象数值预测方面的工作成果，以及位于科罗拉多州博尔德的美国海洋暨大气总署开发的气象观测系统的经验（该系统使用全球定位系统的信号对水蒸气进行遥感观测)。分析的焦点是 GBT(Green Bank Telescope)的数据生成,这也是此案例中最复杂之处。

GBT 是 100 米单碟射电望远镜。它具有独特的无遮挡孔径设计和完全可调节的表面面板。这两个特性可以减少对射电望远镜的光束图案产生不利影响的反射和衍射，并通过其碟面的调整可以达到近乎完美的抛物面形状而不受重达 1700 万磅望远镜主体取向的影响，使得 GBT 成为世界上最灵敏和最精确的单碟望远镜之一。无线电电波被碟面聚集到望远镜的接收器上，这些接收器的工作频率介于 200 MHz 到 50 GHz 之间。目前，将频率上限扩展到 115 GHz 的开发项目还在进行之中。收到信号后,经多次增强和混合,传送至一个称为后端的特殊的信息处理装置。使用 GBT 进行观测的天文学家选择一个或多个这种后端装置以完成其研究计划，并通过连接这些装置取得期望的目标。每次观测前都需要对数百个控制系统参数进行设置，在任何一个步骤中发生错误都会降低所产生的数据产品的质量。

此研究主要关注的是提供优质的原始天文数据和派生数据产品(包括图像、光谱以及称作数据立方体的多维数据集合)，同时也对其他符合结构和质量管理模型的数据产品进行了研究。它们包括从气象观测系统收集的数据、用作天气预测模型输入的同化数据以及 NWP 模型输出数据。其他具有类似特征的数据集包括医学光谱、成像结果以及工程仿真的输入/输出(如有限元模型),不是本研究要专门解决的问题。

3 数据和数据产品的质量管理

低劣的数据质量将限制其有效性。在科学研究中，数据产品是导向终极目的的手段，其目的就是发现、公布和传播科学成果。然而,对于为研究人员提供数据的供应商而言，主要提供生产数据产品的生产设施。研究人员最感兴趣的是具有分析功能的数据产品。自 20 世纪 90 年代早期开始，就已经有了一些针对数据质量、信息质量和将信息作为产品（其是生产流程的成果）生产的研究。研究探讨了与数据质量诊断和管理相关的问题。然而,时至今日,还没有任何专门针对科学数据产品的生产和质量管理中所产生的问题进行研究。

Wang 和 Strong(1996) 将数据质量的 15 个属性分成 4 类：内在数据质量、环境数据质量、表示数据质量以及可访问性质量 [2]。这些属性在文献〔3〕中有更深入的研究，并且在 Enterprise Knowledge Management: The Data Quality Approach(Loshin 2001)中再次进行了精炼。表 1 归纳了 Loshin 有关适用于数据的质量属性的观点[4]。

到目前为止，大多数实用的数据质量研究都

表 1 Loshin 关于数据质量的维度描述

名称	指标	解释
数据模型	定义的清晰度	对对象、属性和关系进行明确的标识以及进行独特且具有代表性的命名。
数据价值	准确	已存储数据对“正确”信息的已接受来源的认同程度。
信息域	企业使用协议	以大家都遵守的术语进行交流。
	管理职责	确保指派了维护属性的完整性和通用性的职责。
	普遍性	鼓励分享数据资源和在不同应用当中标准化数据的使用。
数据表示	适当性	数据的格式和内容满足用户需求的程度。
	正确的解读	所提供信息的全面程度，由此用户可以据其做出准确的推论。
信息政策	可访问性	对于信息的易访问程度以及访问广度。
	元数据	确保不仅对元数据进行了定义，并且还满足了数据质量要求的维度。
	隐私和安全	确保已经设计好一种方法，它能选择性地显示信息的方法，并且还能保护数据和元数据

集中针对大型公司数据库或企业信息存储库。对于它们而言，主要的数据挖掘目标是为生成高质量的、用于商业用途的决策信息。为了使商业决策有合理的基础，数据必须达到某种质量水平。在这些信息存储库中，修正有缺陷的数据，以提高未来决策的效用。影响这些数据存储库的典型问题是：不当的重复、存在不准确的信息以及相同的参考实体存在差异。在这些情况下确保数据质量的主要目的是实现数据仓库对有效决策的支持。许多此类研究都是由主要的咨询公司完成的。实例参见文献[5]和文献[6]。这些研究都没有考虑适用于信息产品库（像科学数据档案）的信息质量问题。

4 科学数据产品的产生

要理解科学数据产品的定义，首先要了解其所生产的科学数据类型，每种类型的质量目标以及这些类型如何组合成产品。在一个复杂的仪器系统中，收集的数据形成数据集。这些数据集经过组合、预处理、聚合就成了“原始数据”。当排除仪器和环境因素后，原始数据就变成了校正数据。当运用算法对校正数据进行转换和组合后，生成了数据产品，再根据自身的要求对数据产品进行分析，以产生科学结果，形成“派生数据”。对于派生数据，还可以自动或人为地剔除主观确定的不良数据，并可修正这些数据，从而产生了“同化数据”。在一些学科中，这些同化数据可以用作复杂模型的输入，产生模型的输出数据。

表2给出了美国宇航局0到2级数据产品模型提取的数据分类概要以及美国宇航局地球观测系统（EOS）把此数据模型扩大至3级和4级后提取的数据描述。

那么，科学数据产品是由什么组成的呢？由于是用户进行最终的质量评估，所以无论数据是否适合其预期的用途，任何可以提供给用户作为中间数据产品或用作科学分析的数据集都可以视为科学数据产品。科学数据产品的最终质量取决于在其生产流程中的所有组成部分的质量，这些组成部分可能来自与其相同的级别，也可能来自更高的级别。表3概括了科学数据产品区别于其他数据类型或信息产品的特征。

表2 科学数据类型的概括和描述

领域	数据类型	术语	描述
仪器	装置监控数据	N/A（1级）	由仪器生成，通常未经过预处理，并且通常未作为原始数据产品的一部分存储。它有助于以下两方面：准备趋势数据以侦测正在出现的仪器故障；为其他故障提供操作响应，进而动态改进生产优质数据产品的潜力。
仪器	原始数据	0级	由仪器生成。可能在固件（如自相关频谱计）中进行有限的预处理。
仪器	校正数据	1级	在去除仪器和环境的影响后得到的数据。EOS将其细分为1A级（附带有注解和校正信息的原始数据）和1B级（处理为校正数据后的原始数据）。
科学	派生数据	2级	根据流程、技术或算法，对校正数据和其他校正数据或其他派生数据进行合并而生成。科学分析可能出现在此级或更高的级别。
科学	同化后的数据	3级	通过对派生数据进行网格化、重取样和/或改变参考系而生成。
科学	模型数据	4级	通过对同化和派生数据产品应用一个或多个数学、物理或随机的模型而生成。

表3 科学数据产品(SDP)的特点

序号	特点
1	数据生产设施的可靠性是建立在数据本身的质量（或缺乏质量）之上。
2	原始、校正、派生以及同化数据类型的多重、嵌套以及相互依赖的多种组合。
3	产品要求运用知识和洞察力得出科学的结论。
4	通常取决于由软件、硬件以及在特定时间执行的算法构成的系统的性能和复杂的相互作用。
5	质量高度依赖于软件中运行的算法和生产流程的完整性。
6	产品质量不仅取决于执行观测操作的系统的状态，还取决于被观测的对象或系统的状态。
7	通常不会进行持续的提炼（例如：数据清理），因为数据产品通常代表的是某一时间点的观测情况。
8	数据清理适用范围有限，因为人们通常并不知道什么值才是正确的。
9	为重复值或者多个数据表示而对数据库进行修复不是通常所关注的事情。
10	可能对许多物理表示的数值范围都可用，并且在不同的数值范围之间可能存在交互（比如，作为NWP输入的派生GPS数据产品；作为天文干涉测量实验输入的单碟数据产品）。
11	需要元数据，以便用户可以对结果的有效性和适用性作为判断。
12	数据质量日益提高，因为人们对产品本身或生产过程阶段中构成质量的要素有了新的认识。

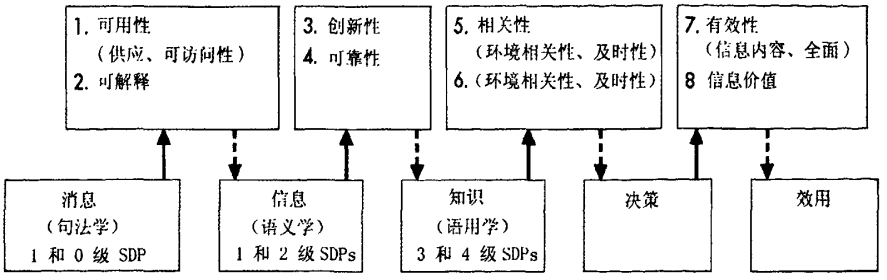

图 1　决策流程环境中科学数据产品的数据质量标准

5　科学数据的性质

科学数据产品的质量可以概括为其符合科学意图的适合程度,这是由 J. M. Juran 提出的"适用性"条件[7]变化而来。根据 TDQM[8]的规定,适用性也是评判数据质量的一个主要标准。为了使科学数据产品具备适用性,其需要满足在图 1 中描述的一些数据质量的属性,比如可访问性、适当性、通用性以及时效性。由于科学数据产品只是达到目的的一种手段,即产生和传播合理的科学结论,其质量评估必须在决策流程的环境中进行。

由于科学数据产品的质量高度依赖于生产流程的完整性,并且生产流程遵循符号学分支的发展,所以 Graefe 关于决策流程中的信息质量标准能够与各种级别的科学数据产品相联系,如图 1 所示。

图 1 中概括了 Loshin 数据质量维度的分类,保证了科学数据产品的质量标准假设数据模型的数据质量,其中包括在"可解释"维度中数据值的数据质量,以及在"可靠性"维度中数据表示的数据质量,因为这些要素是产生可靠性所必需的。为了由决策流程产生效用,必须将信息领域的质量和信息政策的质量的整合纳入质量管理方法。

6　质量管理模型的考虑因素

质量低劣的数据是多种因素综合作用的结果,包括在实验设计过程中的人为错误、不一致或不准确的数据生成过程或者仪器故障。然而,对于由生成和处理科学信息产品的专业设施生产的科学数据产品,即使是由仪器得出的优质数据,可能也不能用于未来不同的实验中,由此就成了质量低劣的数据。因此,即使对流程进行了良好的管理,低劣的数据质量也将始终存在。基于上述原因,根据实验的独特目标进行数据质量评估是质量管理系统至关重要的组成部分。

科学数据产品首先是一个信息产品。因此,信息质量的原理可以用于生成产品质量。TIQM 整合了 W. Edwards Deming 关于质量的 14 个要点、持续改善的要素以及 Philip Crosby 和 J. M. Juran 提出的原理[9]。此方法将这些要点和原理与管理数据和信息质量的各方面相联系,确定了 6 个流程(表 4)。

在表 4 中,P1 和 P3 适用于质量管理的规划阶段,而 P2 则应用于质量的评估和保障。P4 与控制有关,P5 和 P6 映射了持续改进。科学数据质量管理所要求的 4 个高级要素中的任何一个都是现成的,但数据的再设计和修改(P4 流程)需要在按维度定制生产产品的环境中进行大量的检验。

有一些差别未在 TIQM 中进行专门论述,但是对于有效管理不同级别的科学数据产品的质量却是必需的。在 TIQM 中并未提出相对于非随机原因变异,正常原因变异对确定数据质量的重要性(表 4)。其实,这对于科学数据产品生产流程的各个阶段都至关重要。

由于数据产品最终是要存档的,所以流程质量问题必须解决,否则生成大量低质量的数据,归档后影响数据产品的质量。此外,P4 中要求对错误数据进行重新设计和修正通常并不现实,在科学领域中是不可能实现的,特别对于观测系统生成的科学数据产品更为困难,更何况能够用于数据整理的算法需要花费很长时间才能出现。修复数据的优先级应该比寻找能自动修正流程的方法的优先级低得多,并由此产生一个自我修复系统。质量管理模型除提供客观的方法外,还必须提供评估和控制数据质量的主观方法。

这些考虑因素被整合纳入 TIQM 后,为 GBT 的数据特例形成了定制的、详尽的 TIQM 实用描述[10]。

表4 全面信息质量管理(TIQM)

序号	活动	描述
P1	评估数据定义和信息构架的质量	评估数据模型、数据库和数据仓库设计的质量，以及在企业中为数据质量提供支持构架的稳定性和一致性。
P2	评估信息质量	标识适当生成的数据如何符合数据质量的维度，数据质量的维度在组织中非常重要。
P3	度量非优质信息和风险	制定一个质量成本计划，对于与预防、评估、内部故障和外部故障相关的成本进行度量和跟踪。
P4	重新设计和修正数据	生成了有误的数据后，要修复那些数据。结合P5计划确保确定故障的原因，并且改进了流程，进而消除导致错误的主要根源。
P5	改进信息流程质量	改进易发生错误的流程，在向下游蔓延并导致返工之前解决上游的问题，从而为组织带来价值。
P6	营造信息质量环境	应用TIQM关于信息质量的确14个要点来建立质量文化。

在文献〔11〕的附录1中描述了如何在科学数据管理应用中解释TIQM流程的一个例证，文献〔11〕的附录2中概括了信息质量的14个要点，它们可以完全用于建立信息质量文化。尽管附录1描述的研究目前尚未经过经验性的验证而形成科学数据生产的通用质量管理模型，但是它的发展是以前面章节中描述的定义和结构为基础的。

7 总结及未来的工作方向

概念化或分类法的目的不是提供说明性的指导方针，并且也不能进行经验性的验证，其价值在于其在获取更佳的洞察力方面可作为一个有用且有效的工具。使用本文中开发的分类，应该能够为科学中的质量管理建立理论基础，也能够在已经应用到其他学科的大量知识的基础上确定和开发质量管理的技术。理解质量在科学数据生产中的作用就可以在未来研发中准确实施质量管理。由于这样的原因，质量经理要想将这些发现普及到其他环境中的应用上可能是不切实际的。不过，流程工业和科学数据产生流程的相似性表明，工业可以通过研究科学数据生产环境下的质量管理(尤其是标准研究领域)而受益。

通过探讨弗吉尼亚Green Bank的美国国家射电天文台科学数据产品的性质，将它们的产生流程与其他科学数据生产流程进行比较，我们可以得出如下基本论断：

(1)科学数据及其生产流程的独特性体现在3个方面：不同于TDQM，其生产流程包含了产生原始数据以及根据原始数据派生出产品的过程；存储在档案中的数据产品的生命周期要长于生成这些数据的仪器的生命周期，并且必须主观和客观地对它们的质量进行评估；持续改进的过程必须包括档案中数据产品的精炼，并源源不断地将知识整合进生产流程中，以帮助实现前述的精炼。

(2)科学数据可归类为：装置监控数据、原始数据、校正数据、派生数据、同化数据以及模型输出数据。如果数据集是为了供研究人员使用，则可将其视为科学数据产品。

(3)直到原始数据生产出来为止，适用性都是一个客观的过程，其可能包括了校正，这取决于校正操作所能提供的质量水平。在此阶段之后，适用性的评估开始与研究人员的特定科学意图相关。

(4)要描述旨在供用户使用的科学数据产品的质量特征，需要先理解使用科学数据产品的科学意图以及生产科学数据产品的流程。科学数据产品的质量是产品所要达到目的的一个基本条件。

(5)在理解科学决策流程环境下的质量目标方面，使用符号框架可能比单独使用质量目标更为恰当。因为，符号框架反映了在科学数据产品生产流程的各个阶段最重要的质量目标。

(6)应当走出误区：即将焦点放在对已生成的科学数据产品进行大量检验，只为得出其特征描述；科学数据中心管理模型对科学数据产品生产的商品化，这种趋势更具挑战性。更为有效的策略是将质量管理设计到科学数据产品生产流程中，在生产流程中及时反映组织的认识。

(7)制定管理措施，确定每一级科学数据产品的信息质量管理者。比较合理的方式是，生产设施只负责原始数据和校正数据，而研究人员负责派生数据。生产设施可能无法确保信息质量达到特定的级别，以供所有未来的实验使用。

已规划的未来的工作应建立在本研究的基础上。比如，准确地描述构成科学数据质量的各个方面是科学数据分类和质量成本调查的基础。接下来的工作将解决质量成本系统的建立，以帮助NRAO Green Bank确定如何更好地实现其质量目

标。在理解了科学数据产品和科学数据质量构成的基础上，可以研究数据质量与软件质量之间的关系。这是一个重要的联系,因为软件广泛地用于收集、同化和分析数据。比如,如果软件生成的结果不符合研究人员的意图,甚至,即使它在表面上满足了所有的要求，而实质上是在出色的团队有效且高效的操作下得出的，此软件仍不能被视为有很高的质量。目前尚未有文献对类似的重要联系进行论述。

参考文献

[1] Price R J, G Shanks. Empirical Refinement of a Semiotic Information Quality Framework[C] . In Proceedings of the 38th IEEE Hawaii International Conference on System Sciences, Hawaii, 2005.

[2] Wang R Y, D M Strong. Beyond Accuracy: What Data Quality Means to Data Consumer[J] . Journal of Management Information Systems, 1996, 12: 5 - 34.

[3] Redman T C. Data Quality for the Information Age [M]. Norwood, Mass.: Artech House, Inc. 1996.

[4] Loshin D. Enterprise Knowledge Management: The Data Quality Approach[M] . San Diego: Morgan Kaufmann, 2001.

[5] Spivak S M, F C Brenner, eds. Standardization Essentials: Principles and Practice [M] . New York: Marcel Dekker, Inc. 2001.

[6] Tsien P Y. Data Management: The Quest for Quality [M]. Accenture White Paper, 2004, August 24.

[7] English L. Improving Data Warehouse and Business Information Quality: Methods for Reducing Costs and Increasing Profits[M]. New York: John Wiley & Sons, Inc. 1999.

[8] Wang R Y. A Product Perspective on Total Data Quality Management[J] . Communications of the ACM, 1998, 41(2): 58 - 65.

[9] Evans J R, W M Lindsay. The Management and Control of Quality[M] . Mason, Ohio: South - Western College Publishing, 2005.

[10] Kujala J, P Lillrank. Total Quality Management as a Cultural Phenomenon[J] . Quality Management Journal, 2004, 11(4): 43 - 55.

[11] N M Radziwill. Foundations for Quality Management of Scientific Data Products[J] . The Quality Management Journal, 2006, 13: 7 - 21.

Foundations for Quality Management of Scientific Data Products

Nicole M. Radziwill [1], Su Ying [2]

(1. The U. S. National Radio Astronomical Observatory, VA 22901, USA;

2. Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract: Better scientific data quality means more accurate conclusions being made more quickly, and benefits can be realized by society more readily. To improve scientific data quality, and provide continuous quality assessment and management, the nature of scientific data and the processes that produce it must be articulated. The purpose of this research is to provide a conceptual foundation for the management of data quality as it applies to scientific data products. Definitions for data product and data quality tailored to the context of scientific decision making are proposed, given two typical scenarios: 1) collecting observational data, and 2) performing archive - based research. Two relevant extensions to the total quality management (TQM) philosophy, total information quality management (TIQM), and total data quality management (TDQM) are then examined to determine. Recommendations for planning, assessment/assurance, control, and continuous improvement are proposed, focusing on designing quality into the production process rather than relying on mass inspection.

Keywords: data center, data management, data quality, scientific data, semiotics, TDQM, TIQM, TQM