从数据质量到数据产品质量

蔡莉[1]，朱扬勇[2,3]
1. 云南大学软件学院，云南 昆明 650504；
2. 复旦大学计算机科学技术学院，上海 200438；
3. 上海市数据科学重点实验室，上海 200438

摘要
长期以来，数据质量研究主要是为了满足组织自身信息系统正常运行的需求。随着数据要素市场的建设与发展，数据的质量需求从“自用需求”转变为“他用需求”“监管需求”。数据市场中的数据产品质量问题是数据使用者（购买者）和市场监管机构重点关注的内容。分析了数据产品质量的使用者需求和监管者需求，创新性地提出了一个数据产品质量体系框架；在此基础上，以盒装数据产品为例，从时间、空间和内容完整性3个方面构建了对应的质量维度、质量指标和质量评测模型。该质量体系可以对资源类数据产品进行检测和评定，能够为数据产品购买者和市场监管机构提供行之有效的检测依据和标准。

关键词
数据产品；数据质量；质量管理；盒装数据

中图分类号：TP301 文献标志码：A doi:10.11959/j.issn.2096-0271.2022040

From data quality to data products quality

CAI Li[1], ZHU Yangyong[2,3]
1. School of Software, Yunnan University, Kunming 650504, China
2. School of Computer Science, Fudan University, Shanghai 200438, China
3. Shanghai Key Laboratory of Data Science, Shanghai 200438, China

Abstract
For a long time, the purpose of data quality research is to fulfill requirements of the normal operation of the organization's own information system. With the construction and development of data market, the requirements on data quality have changed from "self-use" to "other use" and "need for supervision". The data products quality in the data market is the focus of data users (buyers) and market regulators. The demands of users and regulators for data product quality were analyzed, and a framework of data product quality was proposed innovatively. On this basis, taking BoxedData as an example, the corresponding quality dimensions, quality indicators and quality assessment models were construct from three aspects of time, space and content integrity. The quality framework was suitable for detecting and assessing resource data products, and could provide effective detection methods and standards for data product buyers and market regulators.

Key words
data product, data quality, quality control, BoxedData

0 引言

数据作为信息化的副产品，长期以来处于自产自用的状态，数据质量研究也集中在数据自产自用过程中的质量管理和控制方面。数据质量是随着信息系统的发展而出现的，数据质量会直接影响信息系统的运行效果，因此需开展数据质量研究[1]。数据质量逐渐形成一个专业的研究领域，并涌现出许多重要的研究成果。在20世纪70年代至90年代，数据质量问题的研究更多来源于行业应用，如会计领域、管理领域、统计领域和计算机领域[2]，没有一个关于数据质量的统一知识体系[3]；在1990—1999年，美国麻省理工学院（MIT）的数据质量研究小组在Wang R Y教授[4]的带领下提出了全面数据质量管理（total data quality management，TDQM）的理论，美国国会要求联邦政府的行政管理和预算局（Office of Management and Budget，OMB）制定新的政策，确保所发布数据的可靠性，即数据要有质量[5]；2005年，国际标准化组织（International Organization for Standardization，ISO）下设的委员会开始组织撰写ISO 8000标准[6]，2001年美国国会正式批准“信息质量法”[7]。

市场上流通的产品被称为商品，任何一种在市场上流通的商品在上市前都需要满足一定的产品质量标准、规范或要求，数据产品亦不例外。因此，数据从自用到商品这个质的变化也必将表现在数据质量上，有关数据的质量研究和实践需要从关注原始数据质量到关注数据产品质量、从内部质量控制到外部质量检测，即数据用户和政府监管部门要对数据产品的质量提出要求并进行检测。本文针对数据产品的质量需求，构建了一个数据产品的质量体系，该体系包括数据产品质量的使用需求、数据产品质量的监管需求、数据产品质量评测等6个部分。该体系能为监管机构或消费者提供切实可行的检测依据和标准。本文以盒装数据为例，将数据产品质量体系具体化。

1 关于数据的质量新需求

1.1 数据产品质量现状

农业经济时代的关键生产要素是劳动力和土地，工业经济时代的关键生产要素是资本和技术，而自大数据出现以来，数据是数字经济的关键要素成为共识[8]，从数据满足企业自身信息系统运行到将数据拿到市场上流通，这是数据的质的变化。

美国农业部经济研究服务机构下设数据产品审查委员会，该委员负责监督和实施数据产品必须遵循的质量需求，确保每个数据产品都符合实用性、客观性、透明度、完整性和可访问性标准[9]。美国国家环境信息中心世界海洋数据库（world ocean database，WOD）对其发布的海洋剖面和海洋生物观测数据产品有着严格的质量控制流程，保障了数据产品的稳定性和权威性[10]。上述关于数据产品质量的做法仍然局限在某些部门或领域，不是严格意义上的数据产品质量，其数据产品并不具有通用性和市场流通性。在国内数据要素市场建设方面，有30多家数据交易机构基本没有对数据产品的质量进行监管，仅2021年11月25日成立的上海数据交易所对交易标的的数据质量进行了明确要求[11]。

在市场上流通的数据应该是数据产品，数据产品具有数据类别格式多种多样、数据规模大小不一、数据对象内容千

差万别等特点，因此要形成一个被广泛认可的数据产品标准形态，在此基础上才能构建出一个合理的、具备权威性的数据质量体系。一旦数据产品质量体系构建完成，市场监管部门就可以根据数据产品质量标准检测市场上流通的数据产品质量，而数据产品生产企业就可以根据数据产品质量标准管控数据生成过程各个环节的数据质量问题，提升数据产品质量，达到产品质量标准。

1.2 数据产品质量的使用者需求

数据产品在市场上流通，给他人使用，即数据的“他用需求”。那么使用者（购买者）对数据产品的质量有什么需求呢？目前，在数据交易市场上，数据产品的使用者对产品质量的需求有如下几个方面。

（1）数据量充裕

不同行业或者应用场景下，数据购买者对数据量的需求有所不同。例如，一家做医药O2O（online to offline）的电商平台希望购买能提供药品-病症之间的关系的数据集。国内市场上销售的常规药品的数量达到6万种，如果所购买的数据产品中的数据对象能涵盖这6万种药品，那么数据量就符合购买者的需求。再如，购买者需要利用出租车的全球定位系统（global positioning system，GPS）轨迹数据分析居民出行的热点区域[12]，假定购买者所在城市大约有7 300辆出租车，如果数据集能涵盖全部出租车的运行数据，那么数据量也符合购买需求。此外，数据量还与时间有一定关联。一个月的出租车运行数据肯定比一周的运行数据更加充足，从中获取的数据分析或者数据挖掘的结果也更加准确。因此，数据量表示了在某一应用场景下，数据购买者对数据产品所涵盖数据集的广度和深度的要求。

（2）来源权威

数据产品是否由权威机构提供，或者由权威专家或专业人员参与数据产品的采集、处理、实现和发布，以及比对的标杆是否来源于权威资料，也是数据购买者关注的质量需求之一[13]。以前文的药品数据产品为例，通常能提供药品信息的权威机构是药品监督管理局，但其提供的数据并不包括疾病方面的信息，无法满足购买者的需求。于是，购买者退而求其次，只能从一家提供药学服务的公司购买所需要的数据产品。

（3）数据准确

数据产品的准确性是数据购买者关注的第三个质量需求，数据产品的准确性越高，其可信度越高，所能产生的数据价值也就越高；反之，则可信度越低，数据价值也越低[14]。准确性的衡量比较困难，当有标准数据集或者参考数据集时，可以将数据对象与之进行对比，确定其准确性。否则，只能在一定误差范围内确定准确性。在上述例子中，可以将药品数据产品中的部分信息与药品监督管理局提供的药物信息进行对比，以确定内容的准确性。但是，出租车的GPS轨迹数据没有对应的标准数据集或者参考数据集，只能在一个给定的限制条件下判断其准确性。例如，如果一辆出租车在工作日早高峰某个时间点的车速达到120 km/h，基本可以判断这一数值是错误的。

（4）数据之间的一致性

数据产品中的数据对象都有一些属性或者字段，有些属性之间会存在一定的关联关系或者映射关系，这些关系可以被统称为一致性。例如，邮政编码与地址信息存在一种映射关系，邮政编码涵盖了周边一定投递范围内的地址信息。如果两者不匹配，那就破坏了一致性的质量需求。另外，有些数据产品直接来源于数据库中不

同表之间的连接查询结果，一张表中某个属性的取值范围由另一张表中对应属性的取值确定，这也是一致性需求的体现。

（5）数据产品的时间

有一些应用场景对数据产品的发布时间或者更新时间有明确要求，甚至希望能提供近乎实时的数据。例如，某导航公司准备提供实时路况的查询功能，因而需要购买浮动车数据。所谓浮动车就是安装了GPS设备的车辆，通过网络将实时的经纬度位置、车头方向、速度等值传递到处理中心，进而计算出全市主要道路的路况信息[15]。通常，浮动车包括出租车、长途客车、物流车辆等，其中最重要的车辆就是穿梭于城市各种道路的出租车。还有一些应用场景则希望数据产品的更新时间能与自己的业务相匹配，以获得更优质的服务[16]。例如，一个外卖平台与提供高分天气预报的公司合作，想结合天气预报做更多的场景挖掘，比如分钟级降雨预报，以此判断接下来2 h订单量是否激增，外卖员的平均送单时间是否增加等。

（6）数据产品的获取方式

数据产品的获取方式多种多样，有一些数据产品可以直接到交易平台购买；另一些数据产品由于数量较大，交易平台上只会提供样本数据，全量数据需要经过一定授权后通过应用程序接口（application programming interface，API）下载，或者经过协商后采取远程查询数据库的方式获取。因此，数据产品获取方式的难易程度也是购买者关注的一个质量需求。

（7）质量反馈

某些数据产品的适用场景较少，购买者数量不多，导致该产品的评价或者反馈意见很少。还有一些数据产品由于适用场景较为广泛，出现了数量较多的购买者。如果数据产品也能像普通商品一样提供用户购买后的使用体验或者质量反馈，就能帮助新的购买者判断这一产品是否符合自己的需求、是否值得购买。

（8）元数据信息

元数据是用来解释数据的数据，它可以帮助购买者理解数据产品的各种信息和真实语义，是数据提供者和购买者之间沟通和理解的桥梁[17]。元数据记录了数据计算文档、语法和语义描述、质量指标、访问控制策略、数据“血缘关系”等信息。

1.3 数据产品质量的监管者需求

数据产品流通市场需要政府监管才能保证市场的公开、公平和公正，才能形成一个良性市场。数据市场监管者对数据产品质量的需求就是“监管需求”，包括如下4个方面。

（1）数据产品的合规性

数据产品是在充分挖掘数据价值的基础上帮助用户进行决策（甚至行动）的一种产品形式。数据产品来源于数据，因此，数据采集或爬取是否符合国家的法律法规成为监管者最关注的监管需求。当前，数据产品的提供者主要是企业，而企业数据合规风险来自由大量个人信息构成的运营数据，我国现行法规要求企业在采集公民个人信息时坚持同意、合理、最小化3项基本原则[18]。在交易数据产品之前，市场监管部门需要调查数据来源的合法性，调查因素包括被收集人是否知晓该数据被数据产品提供方收集、数据流通行为是否已经得到被收集人同意、数据利用形式是否已告知被收集人并得到同意以及接收数据的种类等。除了通过业务采集的数据，一些企业还会通过爬虫技术抓取外部数据。非法的数据爬取会带来不正当竞争、侵犯商业秘密等民事纠纷或非法获取计算机系统数据罪的风险，这些风险也需要监管部门予以考虑[19]。

(2)有效的数据产品质量标准

数据产品在市场上交易之前，最好能通过相应的质量检测，现阶段这一工作主要由数据产品提供方自行完成。由于我国并未出台针对数据产品的国家质量标准，数据产品提供方会依据自己制定的质量标准完成检测。质量标准不统一使得监管部门或者购买者难以判断数据产品的质量，进而影响后续的数据定价以及质量问题维权。此外，现有参与交易的产品质量检测报告大多由数据产品提供方自己提供，很少由第三方质量检测机构出具，缺乏一定的公信力[20]。如果国家层面或者行业层面能出台一个有效的数据产品质量标准，那么该标准既可作为数据产品生产、检验和评定质量的技术依据，又能为数据要素市场的发展提供强有力的服务保障。

(3)数据产品的可溯源性

一些数据产品是由原始数据集经过一定的处理形成的衍生产品，这些处理涉及流转、复制、迁移、集成、抽取、计算等操作。如果没有对原生数据的溯源信息进行记录，将在很大程度上降低数据产品的真实性和有效性[21]，从而为特定的数据应用场景带来风险。溯源信息可被看作数据的元数据，通常包括what、why、when和where 4个方面的元素[22]。其中，what描述影响数据发生的事件，包括创建、使用、存储和转换，甚至涉及数据的存档；why描述事件发生的原因；when记录事件发生的时间；who是这些事件涉及的人或组织。数据产品的可溯源是指利用标记、数字指纹等方式，实现对数据产品整个生命周期内所经历的全部操作及变换信息的描述，确保由原始数据衍生的数据产品真实可靠，也是建立信任和实现责任制的重要基础。

(4)应用场景明确

数据产品的产生和交易是为了满足用户的某些需求，其应用场景描述了关于产品、用户及其环境的背景信息、用户的目的或目标、一系列活动和事件等内容。由于用户的需求类型多样，明确应用场景一方面可以帮助监管部门判断数据产品是否合规，另一方面也可以提供切合实际管理和应用需求的数据产品和业务应用。

2 数据产品质量体系框架

根据上述数据产品的质量需求，本文创新地提出了一个质量体系框架，如**图1**所示。该质量体系框架主要由应用场景确认、数据产品管理、质量需求描述、质量维度选择、评估模型及方法建立和数据产品质量监控6个部分构成。

(1)应用场景确认

在数据交易市场中，数据本身具有可复制性，因此不同的使用者和不同的使用场景具有不同的价值，不同行业下的应用场景对同一数据产品的需求大相径庭。为了避免违法违规，甚至禁止交易的数据产品或目前不宜交易的数据产品流入交易市场，数据产品的提供者需要明确给出产品的使用场景，以供市场监管方评估及核查。

(2)数据产品管理

按照产品的呈现形式和使用方式，数据产品可分为数据资源类、数据服务类以及数据咨询/决策类3种类型，不同类型的数据产品在质量维度选择和评估模型及方法建立上有较大区别。数据产品管理是将相同或者类似的产品按照应用场景进行归类和存储，从而方便后续的质量评估和监测。

(3)质量需求描述

数据产品质量需求主要有两个来源，分别为使用者和监管者，前者对应数据产

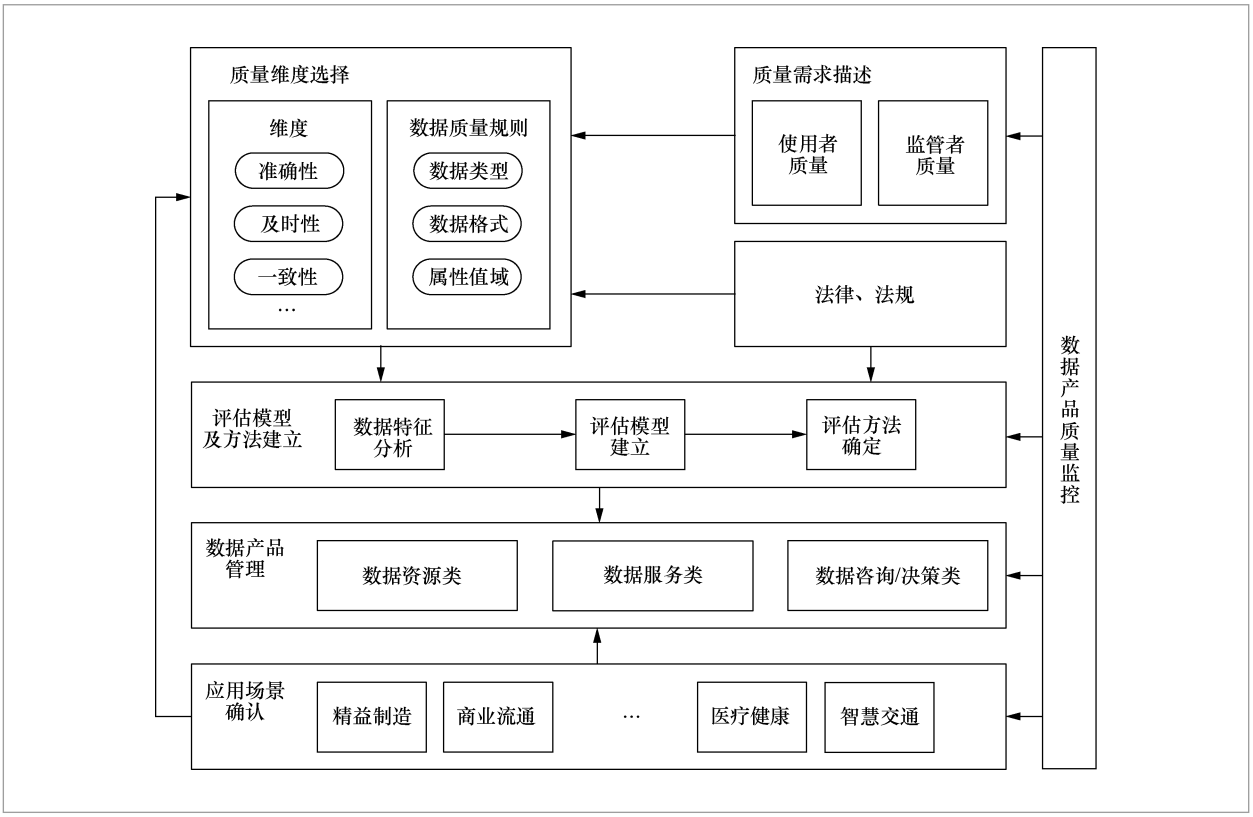

图1 数据产品的质量体系框架

品应用角度的要求，后者对应数据产品监管目标。数据产品质量需求涉及范围和影响程度不一，较小的需求以数据集中的数据对象修改为单位，处理方式简单直接；较大的需求以整个数据集为代表，剖析数据来源，甚至包括采集方式和业务规则的调整。

(4)质量维度选择

在数据质量研究中，研究者提出的质量维度多达20余个，这些维度从不同角度反映了测量和管理数据质量的需求。质量维度的选择主要由数据产品质量标准来确定，但现阶段并未出台针对数据产品的国家标准或者行业标准。因此，可以依据数据产品质量需求、国家的相关法律法规以及应用场景来确定。同时，将质量维度应用到实际的评估模型时，还应该分析数据类型、数据格式和属性值域的分布，以建立每一个维度下的具体评估指标。

(5)评估模型及方法建立

评估模型及方法建立指对各类数据的特征进行分析，根据分析结果和所选择的质量维度及其评估指标，建立评估模型。之后，确定评估方法及其详细过程。评估方法可以采用定性评估、定量评估或者综合评估方法[3]。

(6)数据产品质量监控

数据产品质量监控覆盖数据产品在交易平台上的全流程，并对其进行质量监管和检验，具体任务包括数据产品登记、数据产品合规审查、数据产品溯源、数据产品质量评估、质量报告生成、数据产品交易追踪和数据产品质量反馈等内容。

3 盒装数据产品的质量框架和质量指标

数据产品有多种类型，而盒装数据是叶雅珍等人[23]提出的一种资源型的数据产品标准形态，包括盒内数据和盒外包装两个部分。其中，盒内数据是指“时间+空间+内容”三维度的数据立方体组织，一般包括图像、图形、视频、音频、文本、结构化数据等多种类型的数据；盒外包装是包括产品登记证书、使用说明书、质量证书、合规证书等内容的数据盒外部形态[23]。

3.1 质量维度

盒内数据是用时间维度、空间维度、内容维度来表示的，因此数据质量也可以从这3个维度来评测。**图2**显示了本文提出的针对盒装数据的质量评测体系。整个质量评测体系是一个两层的多维度、多指标的结构。数据产品质量维度是一个可以测量和改进的数据产品的某个特性或者属性。事实上，质量维度提供了一种用于测量和管理数据产品质量以及信息的方式[24]。数据产品质量指标归属于质量维度，是质量维度更细化的评测形式。

3.2 质量指标

建立了盒装数据产品的3个质量维度后，每个维度还需要细分为2~5个质量指标，这些指标可以定量地评估盒装数据产品的质量。时间完整性维度划分为时间覆盖率、时效性和可溯源性3个指标，空间完整性维度划分为空间覆盖率和空间一致性两个指标，而内容完整性维度划分为属性覆盖率、准确性、一致性、可获取性和权威性5个

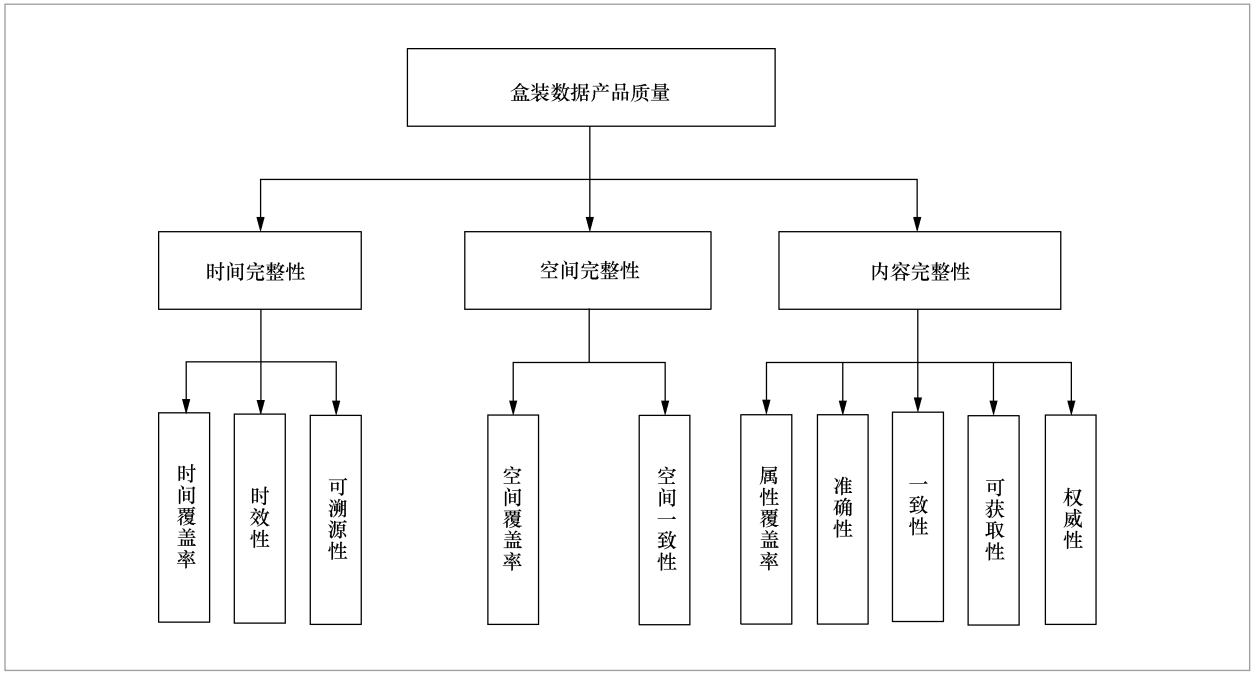

图 2 盒装数据产品的质量评测体系

指标。这10个指标的具体含义见**表1**。

(1)时间完整性维度

时间完整性从3个方面刻画数据产品是否满足时间的质量需求。一是时间覆盖率，指数据集中的数据对象在数据产品所描述的各个时间点上有具体的值，没有缺失。如果数据对象在某些时间点上没有值或者存在描述时间之外的值，则都属于不完整的情况。二是时效性，指数据产品能否在需要的时候得到保证。例如，数据购买者需要购买A医院2021年心脏病患者诊断检验数据集，但是A医院只能提供2020年的相关数据，则无法满足需要提供最新诊断检验数据集的需求。三是可溯源性，指数据产品可以进行溯源。

(2)空间完整性维度

空间完整性从两个方面刻画数据产品是否满足空间的质量需求。一是空间覆盖率，指数据产品中的全体数据对象都应该包含在数据集中。如果没有包含某些数据对象，则是不完整的；如果包含了数据产品描述之外的数据对象，则也是不完整的。二是空间一致性，指描述数据对象之间的空间逻辑关系是否与现实世界相匹配，例如，某一空间数据产品提供昆明市2021年的地图数据集，但是，其中有些用来表征面的数据对象存在重叠和空隙，这就不满足空间逻辑关系一致的需求[25]。

(3)内容完整性维度

内容完整性从5个方面刻画数据产品是否满足内容的质量需求。一是属性覆盖率，指数据集中的数据对象内容完整，没有遗漏，也没有多余。例如，在GPS数据产品中，每辆出租车当天的经度、纬度、运行状态、方向和车速共同构成完整的行驶数据。如果这5种属性少了某一种或某几种属性，那么内容就是不完整的；当然，如果多了某些属性，则也是不完整的。二是准确性，指数据对象的取值是否真实、准确地描述应用场景或者误差能在一定的允许范围内。例如，2020年9月7日出租车云A*****的经度、纬度、运行状态、方向和车速与实际情况相符，那它的数值准确无误；或者某一兴趣点(point of interest，POI)的经纬度误差控制在0.000001° 以内，则数值准确。三是一致性，指数据产

表1 盒装数据产品的质量指标

质量维度	质量指标	定义
时间完整性	时间覆盖率	数据对象在数据产品所描述的各个时间点上的完整程度
	时效性	数据产品从产生到获取再到利用的一个很显著的时间差值
	可溯源性	是否提供数据产品在生命周期内所经历的操作及变换信息的描述
空间完整性	空间覆盖率	数据对象包含在数据产品所描述的空间中的完整程度
	空间一致性	不同的数据对象在空间上的逻辑关系是否正确和完整
内容完整性	属性覆盖率	数据产品在属性集上的完整程度
	准确性	数据产品的正确性、可靠性和可鉴别的程度
	一致性	数据对象在不同属性上的逻辑关系是否正确和完整，或者不同对象之间的逻辑关系是否正确和完整
	可获取性	数据产品可以被获取或者允许授权用户进行下载和使用的方便程度
	权威性	数据产品来源真实且可靠的程度

品间属性或数据内容的一致程度。例如，“2021年高德地图中上海市POI数据集”数据产品（以下简称POI数据产品）中，邮政编码与地址信息要一致。四是可获取性，表示数据产品可以方便地获取或者允许授权用户进行下载和使用。例如，POI数据产品可以直接在交易平台购买或者通过API授权下载。五是权威性，表示数据产品由权威机构或者专业人员提供，可靠性和可用性都很高。例如，POI数据产品由高德提供，高德是国内数字地图、导航和位置服务解决方案提供商，具备国家甲级导航电子地图测绘和甲级航空摄影资质，因此它是一家地图类数据产品的权威提供商。

4 盒装数据产品的质量评测模型

为了更形式化地描述盒装数据产品的质量评测模型，本文给出如下变量定义，见**表2**。下面将描述各评价指标对应的评测模型。

4.1 时间完整性评测模型

时间完整性的评测模型如下：

表 2 变量定义表

变量名	定义
P	数据产品
o	数据对象，$P=\{o_1, o_2, o_3,\cdots, o_N\}$
N	数据对象的数量
A	数据产品的属性集，$A=\{A_1, A_2, A_3,\cdots, A_M\}$
M	属性集的数量
K	数据产品的时间范围，将其切分为K个时间点
LN	空间数据产品中数据对象所在层的数量

$$\text{时间完整性}=w_1\times \mathrm{PT}_{\mathrm{COV}}+w_2\times \mathrm{PT}_{\mathrm{TL}}+w_3\times \mathrm{PT}_{\mathrm{PRO}} \quad (1)$$

其中，w_1~w_3表示权重，$w_1+w_2+w_3=1$，可以根据实际需求或者评测指标的重要性确定权重的取值。$\mathrm{PT}_{\mathrm{COV}}$、$\mathrm{PT}_{\mathrm{TL}}$和$\mathrm{PT}_{\mathrm{PRO}}$分别表示时间覆盖率、时效性和可溯源性的评测结果。

（1）时间覆盖率评测模型

数据对象o_i如果在某个时间点上有缺失，就会影响数据产品的时间完整性。假设映射函数$F(x)$表示数据对象在某个时间点上是否存在，则有：

$$F(\mathrm{oT}_{ij})=\begin{cases}1, & o_i\text{在时间点}t_j\text{上有数据}\\ 0, & o_i\text{在时间点}t_j\text{上没有数据}\end{cases} \quad (2)$$

故时间覆盖率评测模型$\mathrm{PT}_{\mathrm{COV}}$如下：

$$\mathrm{PT}_{\mathrm{COV}}=\frac{\sum_{i=1,\cdots,N,j=1,\cdots,K}F(\mathrm{oT}_{ij})}{N\times K} \quad (3)$$

$\mathrm{PT}_{\mathrm{COV}}$的取值范围是(0, 1)，越接近1，表示数据产品的时间覆盖率越好；反之，则越差。

（2）时效性评测模型

时效性评估反映数据产品的产生或提供是否及时，可以通过计算数据产品产生或提供的时间与当前时间的差值来表示。假设以当前时间作为基准时间并设为t，则时效性评测模型$\mathrm{PT}_{\mathrm{TL}}$如下：

$$\mathrm{PT}_{\mathrm{TL}}=1-\frac{t-t_P}{t} \quad (4)$$

其中，t_P表示数据产品P的创建或提供时间，为了便于计算，可以将t_P和t转换为整数进行处理，在转换时，有相应的函数可以计算当前时间距离1970年1月1日0点0分0秒的总毫秒数。$\mathrm{PT}_{\mathrm{TL}}$的取值范围是(0, 1)，越接近1，表示数据产品的时效性越好；反之则越差。

（3）可溯源性评测模型

可溯源性评测模型主要以定性评

估为主，可将需要溯源的信息设计为打分项，然后检查数据产品中各溯源要素是否由提供者提供。如果是由提供者提供，则获得相应的分值；否则，该项分值为0。最后，将所得分值相加即最终的评测结果。

4.2 空间完整性评测模型

空间完整性的评测模型如下：

$$空间完整性 = w_1 \times \mathrm{PS}_{\mathrm{COV}} + w_2 \times \mathrm{PS}_{\mathrm{CON}} \tag{5}$$

其中，w_1和w_2表示权重，$w_1+w_2=1$，权重的取值由评估者确定。$\mathrm{PS}_{\mathrm{COV}}$和$\mathrm{PS}_{\mathrm{CON}}$分别表示空间覆盖率和空间一致性的评测结果。

（1）空间覆盖率评测模型

空间覆盖率反映数据产品中的数据对象是否缺失或者多余，空间覆盖率评测模型$\mathrm{PS}_{\mathrm{COV}}$如下：

$$\mathrm{PS}_{\mathrm{COV}} = \begin{cases} \dfrac{\mathrm{count}(P)}{N}, 0 < \mathrm{count}(P) \leqslant N \\ 1 - \dfrac{\mathrm{count}(P) - N}{N}, N < \mathrm{count}(P) < 2N \\ 0, \mathrm{count}(P) \geqslant 2N \end{cases} \tag{6}$$

其中，函数count(P)表示对数据产品P计数。若$\mathrm{PS}_{\mathrm{COV}}$的取值为1，则说明数据对象没有缺失或者多余；$\mathrm{PS}_{\mathrm{COV}}$越接近1，则说明数据对象缺失或者多余的情况越少；$\mathrm{PS}_{\mathrm{COV}}$越接近0，则说明数据对象缺失或者多余的情况越明显。

（2）空间一致性评估模型

对于空间数据产品，除了检查空间覆盖率，还需要检查空间一致性。空间一致性是指在空间数据对象之间不存在明显的矛盾或冲突，主要通过拓扑关系来反映两个对象间的空间关系。本文使用空间拓扑关系的描述模型V9I来描述两个对象间的空间关系，这些关系包括相等（equal）、相接（touch）、相交（intersect）、包含（contain）、在空洞内部（cwithin）、内接（interior-contact）、包含于（contained-by）、直接邻近（immediate-adjacency）、被第三个空间实体隔开（2-order-adjacency）、在空洞内部且边界相接（cinterior-contact）10种[26]。在现实世界中，如果两个数据对象的距离超过2 km，则分析它们的拓扑关系一般没有太大意义。因此，需要在对象的邻域范围内考虑拓扑关系。下面给出空间一致性评估中用到的相关定义。

定义1：邻域对象。假设o_{ik}、o_{jl}分别代表第i层的第k个数据对象和第j层的第l个数据对象，若对象o_{ik}、o_{jl}之间的距离小于给定的阈值d_{ij}，则称o_{jl}为o_{ik}的邻域对象，记为$N(o_{ik})=\{o_{jl}|D(o_{ik}, o_{jl}) \leqslant d_{ij}\}$，其中$D(o_{ik}, o_{jl})$为两个对象的欧氏距离。

定义2：拓扑关系。假设对象o_{ik}、o_{ij}之间应该满足的拓扑关系为观测拓扑关系，记为Tp_{ikjl}，它属于10种拓扑关系中的一种，则：

$\mathrm{Tp}_{ikjl} \in \{w | w$ = equal, touch, intersect, contain, cwithin, contained-by, interior-contact, cinterior-contact, immediate-adjacency, 2-order-adjacency$\}$

为了评测空间数据一致性，假设$\mathrm{Tc}(o_{ik}, o_{jl})$为数据对象之间的一致性检查函数，$C(o_{ik}, o_{jl})$为映射函数，定义如下：

$$C(o_{ik}, o_{jl}) = \begin{cases} 1, \mathrm{Tc}(o_{ik}, o_{jl}) = \mathrm{Tp}_{ikjl} \\ 0, \mathrm{Tc}(o_{ik}, o_{jl}) \neq \mathrm{Tp}_{ikjl} \end{cases} \tag{7}$$

则空间一致性评测模型$\mathrm{PS}_{\mathrm{CON}}$如下：

$$\mathrm{PS}_{\mathrm{CON}}=\frac{\sum_{i,j=1,\cdots,\mathrm{LN},k,l=1,\cdots,N}C\left(o_{ik},o_{jl}\right)}{\sum_{i=1,2,\cdots,\mathrm{LN},j=1,2,\cdots,\mathrm{LN}}\mathrm{tf}_{ij}^{w}} \quad (8)$$

其中，tf_{ij}^{w}表示在层L_i和层L_j上拥有拓扑观测关系中某种类型w的邻域对象对o_{ik}和o_{jl}的数目。

4.3 内容完整性评测模型

与前面两个评测模型类似，内容完整性的评估模型如下：

$$\begin{aligned}&\text{内容完整性}=w_1\times\mathrm{PV}_{\mathrm{COV}}+\\&w_2\times\mathrm{PV}_{\mathrm{ACC}}+w_3\times\mathrm{PV}_{\mathrm{CON}}+\\&w_4\times\mathrm{PV}_{AC}+w_5\times\mathrm{PV}_{\mathrm{AU}}\end{aligned} \quad (9)$$

其中，w_1~w_5表示权重，$w_1+\cdots+w_5=1$，权重的取值也由评估者确定。$\mathrm{PV}_{\mathrm{COV}}$、$\mathrm{PV}_{\mathrm{ACC}}$、$\mathrm{PV}_{\mathrm{CON}}$、$\mathrm{PV}_{\mathrm{AC}}$和$\mathrm{PV}_{\mathrm{AU}}$分别代表属性覆盖率、准确性、一致性、可获取性和权威性的评测结果。

（1）属性覆盖率评估模型

若数据产品中数据对象的属性缺失，则会降低数据产品的可用性。变量oVal_{ij}表示第i个数据对象在第j个属性上的取值，则有映射函数：

$$Y(\mathrm{oA}_{ij})=\left\{\begin{matrix}1,\mathrm{oVal}_{ij}\neq\mathrm{null}\\0,\mathrm{oVal}_{ij}=\mathrm{null}\end{matrix}\right\} \quad (10)$$

故属性覆盖率评测模型$\mathrm{PV}_{\mathrm{COV}}$如下：

$$\mathrm{PV}_{\mathrm{COV}}=\frac{\sum_{i=1,\cdots,N,j=1,\cdots,M}Y\left(\mathrm{oA}_{ij}\right)}{N\times M} \quad (11)$$

其中，$Y(\mathrm{oA}_{ij})$为判断第i个数据对象的第j个属性取值是否非空的映射函数，$Y(\mathrm{oA}_{ij})$的取值为0或1。当属性取值非空时，$Y(\mathrm{oA}_{ij})$的值为1，否则为0。

（2）准确性评测模型

准确性反映数据对象是否真实、准确地描述应用场景，设属性集合$A=\{A_1,A_2,\cdots,A_M\}$在该场景下的参考值标准为$R=\{R_1,R_2,\cdots,R_M\}$，设$\varphi(\cdot)$为准确性判断函数，若对象o_i在属性A_k上的取值满足参考值标准R_k，则$\varphi(\cdot)$值为1，反之为0。准确性评测模型$\mathrm{PV}_{\mathrm{ACC}}$为：

$$\mathrm{PV}_{\mathrm{ACC}}(D_i)=\frac{\sum_{i=1,\ldots,N,j=1,\ldots,M}\varphi(o_{ij})}{N\times M} \quad (12)$$

其中，$\mathrm{PV}_{\mathrm{ACC}}$的取值范围为[0，1]，当$\mathrm{PV}_{\mathrm{ACC}}$取值为0时，数据对象的准确性很低；当$\mathrm{PV}_{\mathrm{ACC}}$取值为1时，数据对象的准确性很高。

（3）一致性评测模型

一致性评测用来判断同一数据对象中的不同属性之间的取值是否正确和完整。设A_k和A_l为存在一致性关系的两个属性，$\mu(\cdot)$为一致性判断函数，若对象o_i在属性A_k和A_l上的取值满足一致性关系，则$\mu(\cdot)$值为1，反之为0。则一致性评测模型$\mathrm{PV}_{\mathrm{CON}}$有：

$$\mathrm{PV}_{\mathrm{CON}}=\frac{\sum_{i=1,\ldots,N,k,l=1,\ldots,\mathrm{Cc}(M)}\mu(\mathrm{oVal}_{ik},\mathrm{oVal}_{il})}{N\times\mathrm{Cc}\left(M\right)} \quad (13)$$

其中，函数Cc(M)用来统计属性集A中存在一致性的属性数量。

（4）可获取性评测模型

可获取性是指用户可以获得数据产品的物理条件或者接口，可获取性评测模型$\mathrm{PV}_{\mathrm{AC}}$如下：

$$\mathrm{PV}_{\mathrm{AC}}=\frac{N-\mathrm{UN}}{N} \quad (14)$$

其中，UN表示不能访问的数据对象数量。

（5）权威性评测模型

数据产品的来源各不相同，依据各来源的实际情况，采用定性方法确定数据产品权威性的评测模型$\mathrm{PV}_{\mathrm{AU}}$为：

$$PV_{AU}=\begin{cases}(0.9,1.0], \text{国家行政机构}\\(0.8,0.9], \text{知名企业及公司}\\(0.7,0.8], \text{领域专家及学者}\\(0.6,0.7], \text{行业网站及机构}\\[0,0.6], \text{其他(自媒体等)}\end{cases} \quad (15)$$

如式(15)所示，本文针对不同数据来源，确定其打分范围。来自国家行政机构的数据权威性最高；其次，知名企业及公司、领域专家及学者、行业网站及机构等权威性依次降低；因目前互联网环境中自媒体、营销号大量存在，并且极易传播不实信息，故该来源的数据权威性最低。

5 结束语

数据流通是数据成为资源、成为资产、成为要素的必然，数据要素市场建设是“十四五”期间发展数字经济的重要任务，各地纷纷成立数据交易机构。然而，绝大部分的数据交易机构没有对数据产品的质量进行有效监管，这对于数据购买方来说是一个潜在风险，并影响了数据交易市场的健康发展。为此，本文构建了一个数据产品的质量体系，并以盒装数据为例，将数据产品质量体系具体化。由于数据产品有多种不同的分类形式，本文提出的数据产品质量体系主要适用于资源类数据产品的检测和评定，数据服务类以及数据咨询/决策类的数据产品还需要进一步的改进和完善。

参考文献:

[1] 蔡莉, 朱扬勇. 大数据质量[M]. 上海: 上海科学技术出版社, 2017.
CAI L, ZHU Y Y. The quality of big data[M]. Shanghai: Shanghai Scientific & Technical Publishers, 2017.

[2] SCANNAPIECO M, CATARCI T. Data quality under the computer science perspective[J]. Archivi & Computer, 2002, 2: 1–13.

[3] 蔡莉, 梁宇, 朱扬勇, 等. 数据质量的历史沿革和发展趋势[J]. 计算机科学, 2018, 45(4): 1–10.
CAI L, LIANG Y, ZHU Y Y, et al. History and development tendency of data quality[J]. Computer Science, 2018, 45(4): 1–10.

[4] WANG R Y. A product perspective on total data quality management[J]. Communications of the ACM, 1998, 41(2): 58–65.

[5] 涂子沛. 大数据[M]. 桂林: 广西师范大学出版社, 2013.
TU Z P. The big data revolution[M]. Guilin: Guangxi Normal University Press, 2013.

[6] 王军玲, 李华, 王强. ISO 8000数据质量系列标准探析[J]. 标准科学, 2010(12): 44–46.
WANG J L, LI H, WANG Q. Research on ISO 8000 series standards for data quality[J]. Standard Science, 2010(12): 44–46.

[7] 宋立荣, 彭洁. 美国政府“信息质量法”的介绍及其启示[J]. 情报杂志, 2012, 31(2): 12–18.
SONG L R, PENG J. Introduction and inspirations of the “information quality act” in the American federal government[J]. Journal of Intelligence, 2012, 31(2): 12–18.

[8] 朱扬勇, 熊赟. 数据的经济活动及其所需要的权利[J]. 大数据, 2020, 6(6): 140–150.
ZHU Y Y, XIONG Y. The required authorization to the data-centric economic activities[J]. Big Data Research, 2020, 6(6): 140–150.

[9] DENBALY M, GLASER L, VASAVADA U. ERS data product quality standards[Z]. 2021.

[10] 孙苗, 王子珂, 童心, 等. 典型海洋环境观测数据产品应用现状及对我国的启示[J]. 大数据,

2022, 8(1): 73-83.
SUN M, WANG Z K, TONG X, et al. Status of classical marine environmental observing data products application and enlightenment to China[J]. Big Data Research, 2022, 8(1): 73-83.

[11] 吴琼, 白廷俊, 庞清珊. 上海数据交易所今日揭牌成立[N]. 北京日报, 2021-11-25.
WU Q, BAI T J, PANG Q S. Shanghai Data Exchange was inaugurated today[N]. Beijing Daily, 2021-11-25.

[12] CAI L, WANG H Y, SHA C, et al. The mining of urban hotspots based on multi-source location data fusion[J]. IEEE Transactions on Knowledge and Data Engineering, 2021(99): 1.

[13] 闫鑫, 黄国彬. 科学数据分类研究述评[J]. 图书馆论坛, 2020, 40(5): 45-54.
YAN X, HUANG G B. A review of the study of research data classification[J]. Library Tribune, 2020, 40(5): 45-54.

[14] CAI L, ZHU Y Y. The challenges of data quality and data quality assessment in the big data era[J]. Data Science Journal, 2015, 14: 2.

[15] VANDENBERGHE W, VANHAUWAERT E, DEMEESTER P, et al. Feasibility of expanding traffic monitoring systems with floating car data technology[J]. IET Intelligent Transport Systems, 2012, 6(4): 347-354.

[16] 程静. 微天气与时间、位置和品类的交互对移动O2O平台商家销量的影响研究[D]. 哈尔滨: 哈尔滨工业大学, 2021.
CHENG J. Research on the impact of the interaction of micro-weather with time, location and category on the sales of merchants on mobile O2O platform[D]. Harbin: Harbin Institute of Technology, 2021.

[17] 黄明峰, 刘军, 靖剑波. 贵阳市政府数据开放平台设计与实现[J]. 电信科学, 2017, 33(9): 136-147.
HUANG M F, LIU J, JING J B. Design and practice of Guiyang open government data platform[J]. Telecommunications Science, 2017, 33(9): 136-147.

[18] 相丽玲, 贾昆. 中外个人数据保护标准研究进展与未来趋势分析[J]. 情报杂志, 2020, 39(2): 85-94.
XIANG L L, JIA K. Research progress and future trend of personal data protection standards at home and abroad[J]. Journal of Intelligence, 2020, 39(2): 85-94.

[19] 李延舜. 隐私政策在企业数据合规实践中的功能定位[J]. 江汉论坛, 2020(10): 136-144.
LI Y S. Privacy policy's functional orientation in the compliance practice of corporate data[J]. Jianghan Tribune, 2020(10): 136-144.

[20] 刘长玉. 政府、第三方检测机构和企业质量监管中博弈关系研究[J]. 东岳论丛, 2015, 36(10): 128-132.
LIU C Y. Study on the game relationship between government, third-party testing institutions and enterprise quality supervision[J]. Dongyue Tribune, 2015, 36(10): 128-132.

[21] 陈红玉, 翟军, 袁长峰, 等. 开放政府数据的溯源元数据研究及应用[J]. 情报杂志, 2017, 36(6): 148-155.
CHEN H Y, ZHAI J, YUAN C F, et al. Provenance metadata research of open government data and application discussion[J]. Journal of Intelligence, 2017, 36(6): 148-155.

[22] 胡韵, 胡爱群, 胡奥婷, 等. 大数据背景下数据可追踪性应用分析与方法研究[J]. 密码学报, 2020, 7(5): 565-582.
HU Y, HU A Q, HU A T, et al. Applications analysis and methods research of data traceability in big data content[J]. Journal of Cryptologic Research, 2020, 7(5): 565-582.

[23] 叶雅珍, 朱扬勇. 盒装数据: 一种基于数据盒的数据产品形态[J]. 大数据, 2022, 8(3): 15-25.
YE Y Z, ZHU Y Y. BoxedData: a data product form based on databox[J]. Big Data Research, 2022, 8(3): 15-25.

[24] MCGILVRAY D. Executing data quality projects: ten steps to quality data and trusted information TM[M]. San Francisco: Morgan Kaufmann Publishers, 2008.

[25] 蔡莉, 李永轩, 王淑婷, 等. 基于层次分析法的众源地理数据质量评估研究[J]. 测绘地理信息, 2021, 46(3): 98-102.
CAI L, LI Y X, WANG S T, et al. Quality assessment for crowdsourcing geographic data using AHP[J]. Journal of Geomatics, 2021, 46(3): 98-102.

[26] 任艳, 易宝林, 陈佳丽. 基于V9I模型的空间拓扑一致性发现与维护[C]//第二十三届中国数据库学术会议论文集. 出版地不详: 出版机构不详, 2006: 400-402.
REN Y, YI B L, CHEN J L. Spatial topological consistency discovering and maintenance based on V9I model[C]// Proceedings of the 23rd National Database Conference of China. [S.l.:s.n.], 2006: 400-402.

作者简介

蔡莉(1975-), 女, 博士, 云南大学软件学院副教授, 主要研究方向为数据质量、数据挖掘和智能交通。

朱扬勇(1963-), 男, 博士, 复旦大学计算机科学技术学院教授、上海市数据科学重点实验室主任, 复旦大学数据产业研究中心副主任。《大数据》期刊编委会副主任, 农业大数据产业技术战略联盟副理事长兼首席科学家, 大数据协同安全国家工程实验室副理事长, 中国自动化学会国防大数据分会副主任。国际数据科学倡导者, 提出数据界、数据学、数据身、数据自治、数据财政等概念和体系。发表学术论文200多篇, 出版《数据学》《漪漩数据》《特异群组挖掘》《数据自治》等专著, 并任《大数据技术与应用丛书》(22册)主编、《大数据资源》主编。主要研究方向为数据科学和数字经济, 近期研究重点为数字化转型、数据财政、数据资产、数据自治与数据跨境等。

收稿日期: 2022-02-07
通信作者: 蔡莉, caili@ynu.edu.cn
基金项目: 国家自然科学基金资助项目(No.61663047); 云南省软件工程重点实验室项目(No.2020SE314)
Foundation Items: The National Natural Science Foundation of China (No.61663047), Key Laboratory in Software Engineering of Yunnan Province (No.2020SE314)