基 础 科 学 数 据 共 享 网 项 目 标 准

TR-REC-011

数据资源加工指导规范

2011 年 3 月

国家科技基础条件平台建设基础科学数据共享网项目组

目　录

数据资源加工指导规范

1 范围

本规范提出国家科技基础条件平台建设项目基础科学数据共享网（以下简称基础科学数据共享网项目）中科学数据资源采集加工过程的规范化要求，包括对组织管理方面的要求、文档方面的要求、数据约定和数据采集加工流程方面的要求等。

本规范是对基础科学数据共享网项目中数据采集加工过程的指导性规范，适用于项目内各类数据资源的采集、加工或更新，各数据库主要承担建设单位应参照本规范建立本数据库的实施细则。

2 规范性引用文件

下列规范性引用文件通过本部分的引用而成为本规范的条款。凡是注日期的引用文件，其随后所有的修改（不包括勘误的内容）或修订版均不适用于本规范。但是，鼓励根据本规范达成协议的各方，研究是否可使用这些文件的最新版本。凡是不注日期的引用文件，其最新版本适用于本规范。

- TR-REC-014 数据集核心元数据标准
- TR-REC-017 唯一标识符规范
- TR-REC-018 基础科学数据分类规范
- TR-REC-062 技术文档参考规范

3 术语和定义

3.1 科学数据资源

科学数据资源是科技活动或通过其它方式所获取到的反映客观世界的本质、特征、变化规律等的原始基本数据，以及根据不同科技活动需要，进行系统加工整理的各类数据集，用于支撑科研活动的科学数据的集合。

3.2 元数据

关于数据的数据。本规范中，具体指描述数据及其环境的数据。

3.3 数据类型

对数据的有效值域及对该值域中的值所允许的操作的规定。例如，整型、实型、布尔型、日期类型、字符串类型等。

3.4 数据集

由相关数据组成的可标识集合。数据集的大小在理论上是不确定的，一个简单的数据表可以成为一个数据库集，几个相同类型的表也可以被成为一个数据集。

3.5 数据项

属性数据中不可再分的最小的单元。

3.6 数据产品

数据产品是遵从统一的标准规范，对基础数据进行集成、加工、处理后生成的新的数据集。该数据集的生产过程和数据质量控制措施可以被人工或计算机详细描述、记录，可被其他人或计算机重复操作。数据产品揭示数据间的内在联系，通过重新组合和再分析，表征某一规律性的现象或过程。

3.7 数据资源加工

生成数据产品的过程，包括数据加工模型、数据处理过程、数据产品质量评价等内容。

4 科学数据资源加工总体要求

数据资源采集加工过程中，数据库承建单位应采用基础科学数据共享网项目发布的有关标准规范，以及相关的国家标准、国际标准、学科领域标准规范或其应用方案，完成对采集加工工作的组织管理、制订数据约，规划数据资源加工流程，并严格贯彻实施，保质保量完

成数据采集加工任务。

对科学数据资源采集加工工作的要求包括多个方面，它规范人员操作，设备要求，数据采集、录入、筛选清理、预处理、处理加工、审核与更新等流程，是科学数据资源高质量建设的有效保障。

4.1 数据资源加工内涵

数据产品具有增值的普遍特征。作为数据产品，必须是经过实质性加工、具有智力投入的成果。有的数据虽然表达形式变化了，但由于没有进行实质性加工和智力投入，并未有效提高数据资源的信息量，也不能称之为数据资源加工。

4.2 数据资源采集加工指导思想和一般原则

4.2.1 统一领导，统筹规划

数据资源采集加工工作应在数据库牵头建设单位的领导下，统一决策，同一数据库范围内工作方法统一，技术指标统一，从而达成数据产品的一致性。

4.2.2 突出重点，注重基础

数据资源的内容选择应在突出重点和注重基础两者之前取得平衡。数据库承建单位应根据当前具备的工作基础以及国内外相关数据库建设情况，确定所承建数据资源的特点和重点内容，对重点内容加以重视，适当提高质量规格。

同时，数据库承建单位应注重基础性和共性数据的建设，确保所承建数据资源的广度，提升所承建数据资源的通用性、易用性，保证数据资源具有一定的用户范围。

4.2.3 需求导向、务求实效

确定资源采集的内容和范围时，既要考虑数据资源单位的数据资源特点以及工作的复杂、难易程度，不能选取太多，过于复杂不便实际使用；又要充分满足工程建设以及用户的查询、使用数据的需要，不能过于简单。数据资源建设工作应当切实以用户需求为导向，以应用为目标，做真正用户需要的数据，而不是盲目地扩大数据内容范围和提升技术指标。

4.2.4 前瞻性、科学性

资源采集加工的内容不但要满足现阶段科学数据资源的使用需求，更应该考虑将来一定时间内由于科技快速发展等原因可能产生的数据资源应用需求，这样建立的数据资源才会更有生命力。确定数据资源采集范围时，可以积极采用国内和国外先进标准。

4.2.5 延续性

对于连续采集数据，数据采集加工的内容应在一定时间范围内具有较好的延续性，使数据资源建设的内容相对保持稳定，增加数据的时间可比性，数据资源采集加工的内容确定应相对慎重，不断地增删数据内容对数据资源积累形成信息造成很大的负面影响。

5 组织管理

数据库主要承建单位负责所承建数据库内数据资源采集加工过程的领导、组织、协调和管理。

数据库各参加建设单位共同承担所承建数据库的数据采集加工工作。

数据采集和加工承担人员应具备以下条件：具有一定的政治素质，爱岗敬业，工作认真负责，细致严谨，熟练掌握数据采集和加工过程所需的学科领域知识和计算机技术。

6 文件要求

为保证所承建数据库数据资源采集加工过程规范健壮，降低人为因素的影响，使标准的技术方法长期延续并加深项目主管单位和用户对数据资源的了解，数据库承建单位应将所所承建数据库在采集加工过程中所采取的政策措施，标准的流程、技术和方法等形成数据资源采集整理工作指南，并发布实施，同时，还应对数据资源采集加工过程的执行情况建立加以记录。

适用时，数据资源采集整理工作指南应包括以下内容：

- 数据来源说明，如资料列表，数据准入原则等
- 数据约定，对拟建数据库规格的约定，包括数据采集的文件格式，数据库模型，指标设置，各项指标的定义、公式、测量方法、精度要求，以及数据采集所使用的样表等。
- 数据采集加工的过程要求，为保证数据资源采集加工工作正常完成所必须执行的工

作过程，每个过程的目标，执行人，设备要求，必要步骤和过程产出结果的要求等。

适用时，采集加工过程的执行情况记录应包括以下内容：

- 工作时间
- 人员
- 相关的环境因素
- 设备运行情况
- 执行情况
- 异常和处理

数据资源采集加工过程的相关信息应填入所承建数据库的元数据对应元素当中。关于数据库核心元数据的更加详细规定参照《TR-REC-014 数据集核心元数据标准》要求执行。

必要时，建库单位应保留数据采集的原始记录一定时间，以备查证使用。

文档书写方面更加详细规定参照《TR-REC-062 技术文档参考规范》要求执行。

7 数据约定

在正式开展数据资源采集加工工作之前，数据库承建单位应以用户需求为出发点，立足于当前承建单位的数据建设能力，对数据资源采集加工直至形成产品的过程和产品的规格进行商讨，并形成约定。

数据约定是数据采集加工工作策划的重要输入项，数据约定的内容中至少应包括以下方面：

- 范围约定

根据学科领域和应用特点确定数据选取范围，保证数据完整性、准确性和连贯。

  - 时间范围约定：数据集描述的起止时间
  - 空间范围约定（如适用）：数据集描述的地理空间范围
  - 学科范围约定：

- 数据量
- 数据类型约定
- 数据质量期望，如填充率水平、差错率水平、主要数据来源等
- 数据库模型，如 ER 图等
- 数据字典

对于每个数据元素，应在以下方面进行描述：

  - 数据来源
  - 采集方法，如采集的部分，拍照要求，计算公式等
  - 设备要求

- 编码方法
- 精确度
- 参照系

对数据采集加工内容的确定应特别注重其规范性，相关的规范包括项目规范、任何可能存在的国家标准、国际标准或行业标准等。其中应特别注重涉及唯一标识符的内容设计应参照《TR-REC-017 唯一标识符规范》要求；涉及分类编码的内容设计应参照《TR-REC-018 基础科学数据分类规范》的要求。

8 数据资源加工流程

8.1 过程策划

规范的采集加工业务流程是保障科学数据资源质量最重要和关键的环节。数据库承建单位应对数据资源采集加工过程进行策划，以需求为导向，对数据采集加工工作的过程方法进行设计，确定有效和高效实现数据加工目标所必须的过程，以及每个过程应该遵循的技术与规范，以及为达成数据采集加工目标所必须的过程输入输出规格要求。

过程策划的输入可以包括但不限于以下方面：

- 用户和其他相关方的需求和期望；
- 对数据资源特性的评估；
- 对服务过程特性的评估等。

特别地，数据资源建设的相关建设应该格外关注是否存在任何可能存在的相关国际标准、国家标准、行业标准或其它相关标准规范可以作为输入项。

对数据资源采集加工流程的约定由数据库主要承建单位负责协商形成，并敦促各承建单位遵照实施。所拟定的各项技术与规范都应写入数据资源采集整理工作指南。

下列流程为不同类型科学数据库资源采集加工常见的业务流程，以及每个业务流程相对通用的原则和质量要求，数据库承建单位可参照选择适宜之条款建立所承建数据库的采集加工过程方法。本规范对下列流程的执行顺序没有要求，但数据库建设单位在数据资源采集整理工作指南中应指出其采集加工过程方法的执行顺序。

在正式展开工作之前，数据库承建单位应对数据资源采集加工过程进行策划，以需求为导向，对数据采集加工工作的过程方法进行设计，确定为达成数据采集加工目标所必须的过程输入输出规格要求。策划结果应该能支持数据采集加工工作有效和高效的实现。

过程策划的结果应该包括：

- 实现数据加工目标所必须的过程，以及过程之间的关联
- 每个标准化过程所应达成的目标和应遵循的规范：

- 目标
- 人员要求
- 资源要求
- 过程的输入
- 一般执行方法
- 过程的输出
- 相关文档

8.2 数据采集

数据采集录入是指对科学数据资源进行收集并形成原始记录的过程。

数据的采集是数据库业务流程的源头，数据采集的质量如何直接关系到信息的质量问题，必须予以高度重视。

8.2.1 数据采集的原则

- 保证采集数据的全面真实。采集的数据必须根据规定的要求，采集到所需要的全部数据，并且保证数据准确真实。
- 因不同的数据调查对象而异，采用不同的采集方法和不同的质量控制要求。

8.2.2 数据采集的技术要求

- 数据采集的内容和各项指标的采集方法根据事先拟定的规则进行，力争做到不缺不漏，其中核心指标项必须填写著录。文字表达应当规范、简明、正确、严谨，含义清楚。
- 如涉及图像拍摄，一般拍摄对象的正面及侧面图像，必要时还应拍摄细部、标题等部位的图像。
- 数据收集中，对有明显错误或不符合规律的数据亦予以剔除。
- 如果存在相关的国家标准或行业标准，数据采集和指标测量应严格遵照相关的标准规范进行。

8.2.3 数据采集工作流程

- 从数据来源查询获取数据，并按照一定的规则整理收集；

- 在数据记录中采取注明实验条件和实验误差的方法给用户提供参考；
- 相关专家考察、审核相关数据；
- 数据由工作人员填写原始记录表格或原始记录入库；
- 如果存在计量单位不一致的情况，则先进行换算单位，应注明单位换算的情况。

8.3 数据录入

涉及数据录入时，数据库承建单位应对录入设备，录入人以及必要的质量控制措施等等相关信息加以记录。

8.3.1 数据录入的技术要求

- 所使用的录入系统必须是指定的录入系统；
- 输录要完全忠实于采集得到的资料；
- 必填内容不得为空。

8.3.2 数据录入的方法

- 文本数据手工填报；
- 文本数据计算机手工录入；
- 二维图像信息拍摄或计算机自动扫描；
- 三维音像信息多媒体摄像制作；
- 原有数据的格式转换。

8.4 原始数据的保存

必要时，数据库承建单位应设定原始数据保存时间要求，并对数据采集得到的原始数据加以妥善保存，以备需要时复查使用。

如有必要，数据上交时应附带原始记录及相关数据。

8.5 来源筛选

为确保数据产品的质量，数据库承建单位应对原始数据获取来源进行选择，建立数据来源的准入门槛制度，从开始阶段就对数据资源质量进行控制。

数据来源可以是其它数据库资源，也可以是文献，书籍等其它媒体形式的资源。

考虑到所收集数据的可靠性，数据来源均应为公开发表的国内外一级或核心科技刊物的发表论文，原始文献以书籍、手册、综述等为来源的数据。

数据来源筛选的原则可以包括但不限于以下方面：

- 数据生产者和提供者的口碑；
- 数据来源的时间、空间、学科范围符合本数据库的使用预期；
- 数据来源的数据规模满足需求；
- 数据来源使用的数据格式符合需求；
- 数据来源遵循某一国际或国内知名的数据标准建立；
- 数据来源的技术指标，如准确度，精确度水平等；
- 数据来源的主要内容；
- 数据来源是否具有完整的元数据或相关资料描述。

8.6 原始数据标准化预处理

为避免原始数据过于庞大，信息过于复杂，数据受噪声数据、空缺数据和不一致性数据的侵扰，必要时，数据库承建单位应对采集得到的原始数据进行标准化预处理。

数据处理的主要目的在于

- 减少误差。消除数据中的一些明显错误、粗差或系统误差。
- 提高数据的系列性，尤其是在时间和空间序列上的连续性。
- 提高数据的完整性，对单一要素数据进行综合。

一般的原始数据预处理方法包括数据清理、数据集成和变换、数据归约等。

8.7 数据清理

数据清理用于填充空缺值、识别孤立点、消除噪声、纠正数据不一致。常用的数据清理方法包括：

8.7.1 空缺值的清理

- 忽略元组
- 人工填写空缺值
- 使用一个全局常量填充空缺值
- 使用属性的平均值填充空缺值
- 使用与给定元组属同一类的所有样本的平均值
- 使用最有可能的值填充空缺值

8.7.2 噪声数据

- 分箱
- 聚类
- 计算机和人工检查结合
- 回归

8.7.3 不一致数据

对于有些事务，所记录的数据可能存在不一致。有些数据不一致可以使用其他材料人工地更正。知识工程工具也可以用来检测违反限制的数据。例如，知道属性的函数依赖，可以查找违反函数依赖的值。

8.8 数据集成

数据集成用于将来自不同数据源的数据整合成一致的数据存储。元数据、相关分析、数据冲突检测和语义异种性的解析都有助于数据集成。

主要方法包括：

8.8.1 模式匹配

利用数据库的元数据对异构数据进行映射转换，形成模式匹配。

8.8.2 消除冗余

利用相关行分析的方法检测冗余，消除重复数据。

8.9 数据变换

将数据转换成适合使用的形式。

主要方法包括：

8.9.1 平滑

去掉数据中的噪声。这种技术包括分箱、聚类和回归。

8.9.2 集

对数据进行汇总和聚集。

8.9.3 数据概化

使用概念分层，用高层次的概念替换低层次的“原始”数据。

8.9.4 规范化

将属性数据按比例缩放，使之落入一个小的特定区间，如-1.0 到 1.0 或 0.0 到 1.0。

8.9.5 属性构造

由给定的属性构造和添加新的属性，以帮助提高精度和对高维数据结构的理解。

8.10 数据归约

对数据处理的技术，如数据立方体聚集、维归约、数据压缩、数值归约和离散化都可以用来得到数据的归约表示，而使得信息内容的损失最小。

8.10.1 数据立方体聚集

聚集操作用于数据立方体中的数据。

8.10.2 维归约

通过删除不相关的属性（或）维减少数据量。通常使用属性子集选择方法。

8.11 数据加工模型和算法

数据库承建单位应根据基础数据的类型，建立相应的数据加工模型和算法。例如，针对属性数据加工的要求，建立属性数据加工模型和算法；针对栅格数据加工的要求，建立栅格数据加工模型和算法；针对矢量数据加工的要求，建立矢量数据加工模型和算法。

数据加工应基于统一的模型，如概念模型，地理坐标系，高程参照系，时间模型，统一的文件格式等。

属性数据加工模型的核心是对属性数据进行规范化处理，包括赋予属性数据以空间特征，以及基于数学模型对属性数据进行均一化处理等。

间格网化模型可以使属性数据生成标准的数据产品。

专题数据产品突出反映一种或几种主要要素或现象。

8.12 数据处理加工与产品生产

数据加工处理是指对已经采集的数据按照拟定的数据加工模型和算法进行汇总、计算、分析及数字化处理的过程。数据按要求，开发处理系统，进行加工处理，产生需要的数据、报表等。图形、多媒体数据按照业务要求进行加工，可以和相应的制作、转换工作相结合。

这一过程，可以是计算机自动处理、手工操作，或者是计算机与人工相结合方式进行。根据数据资源加工程度的不同，数据产品可分为多级。

8.12.1 数据加工的级别

- 0 级数据：未作任何处理的原始记录，其记录格式、符号、代码等大多由作业者本人或其服务的单位自行设置，外单位人员，即使是同行，也是无法理解这些数字的含义的。
- 人们对数据规范标准认识不断提高的今天，0 级数据正在逐渐消失。各部门、系统纷纷制定了数据标准和统一格式，科学数据从产生那刻起，就是标准的、他人可读的了。
- 1 级数据：经初步加工，包括数据项的必要注释、数据格式的简单转换等，成为能让他人理解的数据。这是原始数据记录生产地向上级主管部门报送的数据，这对于原始数值生产地而言是“数据成品”；而对于接受单位，特别是承担数据归档、服务的数据中心而言则是“原始数据”。
- 2 级数据：在数据中心对数据作进一步加工处理，主要是两个方面的工作：其一是标准规范化处理，其二是数据质量检查与订正，使数据真正成为可以被利用的数据。

- 3 级数据：在 1、2 级数据的基础上，进一步深加工而形成的科学数据产品。科学数据产品应当有统一的分类和编码系统，有统一的数据格式或能提供转换接口；应当置备标准、完善的元数据；应当有数据质量标准，并经规范的质量检验与修正；还要有标注明确的外包装。
- 4 级数据：为了特殊的用途，并非数据中心日常业务范围之内，而专门为之整理、加工和生产的科学数据产品。

8.12.2 数据加工处理的原则

- 数据在加工处理过程中必须始终保持与原始数据的一致性和完整性，不能出现丢失或改变原始数据的情况；
- 经过加工处理后的数据，必须是正确的数据，不能由于软件或操作的原因出现新的错误数据；

8.12.3 数据加工处理的技术要求

- 加工处理的数据必须是经审核通过的采集数据；
- 数据加工处理的软件必须是经测试和试用被证明是具有良好的稳定性、可靠性和容错性，并经过正式批准使用的软件；
- 数据加工处理人员必须是具有资格、并经过授权的专业人员。
- 采集的数据进行加工制作，包括查重、著录、标引、录入、校对、审核、入库等，并最终形成各种专题数据库。
  - 查重：对收集到的数据在已建数据库中查重。
  - 标引：分类标引和主题标引。
  - 录入：按数据库要求的格式录入标引后的数据。
  - 校对：对数据准确性、数据内容全面性、数据著录规范性等进行校对。
  - 入库：数据存入数据库。
  - 汇总（迭加汇总、超级汇总）：由原始数据汇总生成综合数据。
  - 计算：按各种数学模型和算法对数据进行计算；
  - 分析：对数据进行合理性、准确性、相关性、趋势性等各种统计分析，如对比分析、构成分析、相关分析、时间序列分析等，并生成相应的图形图表。
  - 修复：根据已有残缺或局部数据进行修复，或生成全貌完整数据。

8.13 数据审核

数据审核是一种评价过程。这种评价是以审核准则为依据，以审核证据为前提，做出客观的评价。数据审核就是对数据的有效性进行核实。

数据审核的目标是确保数据内容与被描述对象相一致，并且质量符合数据产品标准要求。

数据审核可以贯穿于整个数据资源加工过程之中，可以量化评价的内容包括数据来源质量评价、数据加工模型与算法质量评价、数据产品质量评价等。

数据审核可以由数据采集加工人员自检，也可由数据库主要承建单位专门进行。

适宜时，数据审核宜采取计算机辅助方法进行。

数据库主要承建单位应明确审核所参照的评估模型和方法以及技术要求等。如果学科领域内已存在相关的数据质量管理国际、国家规范或行业标准，数据审核宜采用这些相关标准。审核指标的设置应在符合实际的前提下尽可能不应与当前国际领先水平有太大差距。审核指标可以包括但不限于准确性，真实性误差等技术参数，特色数据和重点数据宜适当提高指标。

数据资源审核通过后方可正式对用户提供服务，未能通过审核的数据一般应返回到必要的流程进行修正或重新加工。

关于数据审核的方法与审核结果应包含在数据库用户手册和数据库对应的元数据当中。

8.14 数据更新

数据更新是对存储在数据库中的数据资源进行补充、修改和删除的工作。

数据更新的目标通常是为了维持所承建数据资源的现势性或使其具有连续性。

适宜时候，数据库承建单位宜采用数据更新流程，一般数据更新应订立数据更新计划，计划内容包括更新的频率和周期，数据更新的内容、范围和总量等。

执行数据更新时一般应重新执行本数据库采集加工的完整流程。