医疗全量数据模型管理工具建设与实践*

严晓明 吴 龙 陈秀娟 李 琴 刘立宇 张 军 韦志强

（广东省人民医院 广州 510080） （生命奇点（北京）科技有限公司 北京 100089）

张 帆 高云鹤 梁会营 杨小红

（广州市妇女儿童医疗中心 广州 510623） （广东省人民医院 广州 510080）

〔摘要〕 介绍建立数据模型工具的意义、数据模型构成与功能，阐述医疗全量数据的数据模型管理工具系统设计及功能、应用效果，指出其应用可实现对数据模型全生命周期管理，为后续数据治理奠定良好基础。

〔关键词〕 国际数据管理协会数据管理体系；医疗数据建模；医疗数据模型

〔中图分类号〕R－058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2022.06.004

Construction and Practice of Model Management Tools for Medical Full Data

YAN Xiaoming, WU Long, CHEN Xiujuan, Guangdong Provincial People's Hospital, Guangzhou 510080, China; LI Qin, LIU Liyu, ZHANG Jun, WEI Zhiqiang, Gennlife (Beijing) Technology Co. Ltd., Beijing 100089, China; ZHANG Fan, GAO Yunhe, Guangzhou Women and Children's Medical Center, Guangzhou 510623, China; LIANG Huiying, YANG Xiaohong, Guangdong Provincial People's Hospital, Guangzhou 510080, China

〔**Abstract**〕 The paper introduces the significance of building data modeling tools, and the composition and functions of data models, expounds the system design, functions and application effect of data model management tools for medical full data, and points out that its application can realize the whole life cycle management of data models, and lays a good foundation for the subsequent data governance.

〔**Keywords**〕 data management system of Data Management Association (DAMA); medical data modeling; medical data model

1 引言

随着医院信息化的快速发展，医院信息系统积累了大量临床医学数据与临床影像数据。这些分散在不同业务系统的医疗数据是医疗活动的记录，通过数据采集、数据建模、数据清洗等一系列数据治理工作建立基于科研主题的医疗全量数据平台，可为医院疾病防控、临床辅助诊断、药品监督以及精准医疗等多方面赋能[1]，推进医院科研高质量发展，提升医院核心竞争力。但目前

〔修回日期〕 2022－06－04

〔作者简介〕 严晓明，软件工程师，发表论文 8 篇；通信作者：杨小红，主任医师，硕士生导师。

〔基金项目〕 国家重点研发计划“面向不确定需求的测试数据集配置平台研发”（项目编号：2019YFB1404803）；国家重点研发计划“医学人工智能产品检测共性关键技术及标准研究”（项目编号：2019YFB1404804）。

医疗数据存在重创造轻管理、重数量轻质量、重业务轻增值的现象，在服务创新、数据质量、开放共享、安全合规等方面面临严峻挑战[2]。通过整合不同业务系统医疗数据建立医疗全量数据平台，需要构建完整的数据治理体系以及符合新时代需求的医疗数据质量监管保障机制[3]。保障机制需涵盖元数据、主数据、数据质量管理等内容[4]，其中数据治理核心之一是数据模型即数据元的管理。目前数据建模过程仅停留在建模的过程文档中，并未形成统一的建模方法以及质量管理体系，建模过程中缺少管理工具辅助。在数据治理过程中容易出现模型定义混乱、模型变更过程缺少监管与模型标准管理等问题。本文以国际数据管理协会（Data Management Association，DAMA）数据管理体系相关理论为基础，建设基于医疗全量数据的模型管理工具，提升数据建模效率，实现数据建模质量管理以及模型全生命周期管理，为后续数据治理奠定基础。

2 建立数据模型工具的必要性

随着医疗数据量和数据应用需求的迅速增长以及数据仓库、大数据等技术的成熟应用，信息互联互通和大数据应用成为医疗信息化过程中的刚性需求。数据治理的重要性日益突显，高效的数据治理被视为医疗信息互联互通与医疗数据价值有效挖掘的重要基础[5]。医疗信息系统多呈现碎片化特征。大数据应用往往需要将医院中几十个甚至上百个异构医疗信息系统产生的各类数据进行集成汇聚并进行统一建模。只有依据国家相关标准建立的数据模型才能在适应性、共用性和稳定性方面满足物联网、大数据、人工智能等新一代信息技术应用需求，达到国家对医疗健康信息互联互通标准化的要求[6]。一般数据建模后的数据表结构大概有几百个甚至几千个，如果缺乏有效管理则可能导致以下问题：数据表、字段定义或者注释缺失，导致字段意思含糊不清、同名不同义或同义不同名、冗余字段和表、枚举值不一致等问题；数据模型变更合理性未得到有效控制，对变更过程缺乏记录，无法进行追溯；业务流程发生变化时未同步修改数据模型，导致数据模型与应用系统中数据不一致；无法及时、准确地为管理者提供数据模型的全生命周期过程相关信息。因此在建设医疗全量数据平台的数据治理过程中，为提升数据建模效率、规范数据模型标准、实现数据模型的全生命周期管理，建设数据模型管理工具十分必要。

3 数据模型简介

3.1 概述

DAMA 数据管理知识体系对数据管理职能形成共识。其提供常用的数据管理职能、交付成果、角色和其他术语标准的定义，确定数据管理指导原则与范围、界限，引导相关数据管理人员接触更多资源并加强对数据管理的理解。

3.2 定义及构成

DAMA 数据管理知识体系中，数据建模被定义为发现、分析和确定数据需求的过程，用数据模型的精确形式表示和传递这些数据需求[7]。数据模型按照描述详细程度的不同，每种模式可分为 3 层模型：概念模型、逻辑模型和物理模型。每种模型都包含一些组件，如实体、关系、事实、键和属性；数据模型数据可以采用多种不同模式表示，其中最为常见的 6 种模式分别是：关系模式、多维模式、面向对象模式、事实模式、时间序列模式和 NoSQL 模式。在建模活动中 DAMA 强调建模规划、正向逆向建模、模型审核、模型维护等过程，需要通过持续改进来控制模型质量以及促使模型保持最新状态。

3.3 功能

根据 DAMA 体系理念，通过构建数据模型管理工具实现命名规范、数据标准管控、值域管理、缩写管理、数据映射管理、版本管理以及模型质量管理等功能。通过管理工具持续改进模型质量，保证模型在长期工作中保持最新状态，为数据治理奠定良好基础。

4 系统设计

4.1 设计原则

4.1.1 概述 医疗全量数据模型管理工具基于 LINUX 操作系统，运用 Kubernetes 和 Docker 虚拟化技术构建大规模 Hadoop 集群，提供大规模高性能分布式数据存储和在海量数据中的映射能力。系统设计遵循标准化、规范化等原则。

4.1.2 标准化和规范化 系统遵循医疗行业标准并与医院实际数据情况相结合，建立医院标准化数据管理规范，提供相应标准化组件功能实现模型化管理服务，降低复杂的医疗数据管理和整合成本、改善数据整体利用效果。

4.1.3 完整性和实效性 系统建设坚持完整性原则，统筹规划、统一设计，采取有力的组织措施和严格的制度保障，保证数据建模、数据采集及数据使用等过程模型管理功能的完整性和实效性。

4.1.4 先进性和实用性 在设计理念、技术体系等方面要求具有先进性和成熟性，以期满足系统在较长生命周期内具有可维护性和可扩展性。系统设计必须考虑易维护和管理性，保证系统在运行过程中能够快速、准确地定位和排除故障。系统界面应简单、美观、容易理解且易于操作，方案选择和功能设置应追求实用性，必须切合全量医疗数据模型管理的实际需求。

4.1.5 兼容原有业务信息系统 充分发挥已有系统功能，利用现有医院数据库的数据架构，通过数据集成和转换快速形成对应数据模型。不仅需要支持业务系统现有数据和历史数据的数据模型，还需支持临床文档结构（Clinical Document Architecture，CDA）进行数据建模与数据映射，满足异构数据源的数据建模。

4.2 系统架构

4.2.1 概述 数据建模是集成医院各临床、运营、管理等全量数据的基础，在以科研、患者、管理等数据维度进行整合过程中使用统一数据模型，并对数据进行质量控制、标准化管控和数据治理。基于规范化、归一化后的全量标准化数据可以构建面向数据分析的各类应用，如临床研究、临床辅助决策支持、智能患者服务、智能药品研究、绩效管理、运营管理等方面。按照系统功能将数据模型管理工具分为 3 个层次，见图 1。

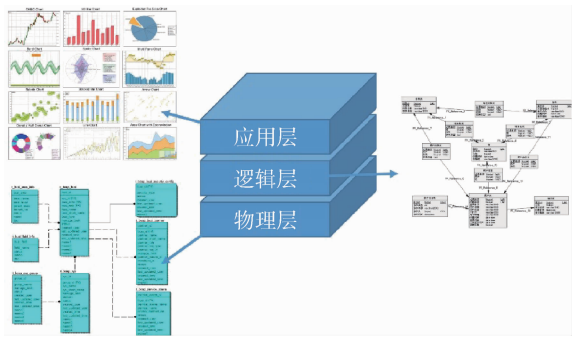

图 1 系统功能架构

4.2.2 物理层 实现依据模板对数据模型增加、变更、模型转换以及版本管理等功能。可根据不同数据库系统表信息生成数据模型，快速进行模型采集与管理，是数据建模的基础。

4.2.3 逻辑层 实现对数据模型与物理模型映射配置以及查询，由数据结构、数据操作、数据映射和数据完整性约束条件组成。

4.2.4 应用层 提供数据模型的可视化显示、统计与查询等功能，辅助数据模型管理人员或者数据分析人员清晰地了解数据模型整个生命周期。

5 系统功能（图 2）

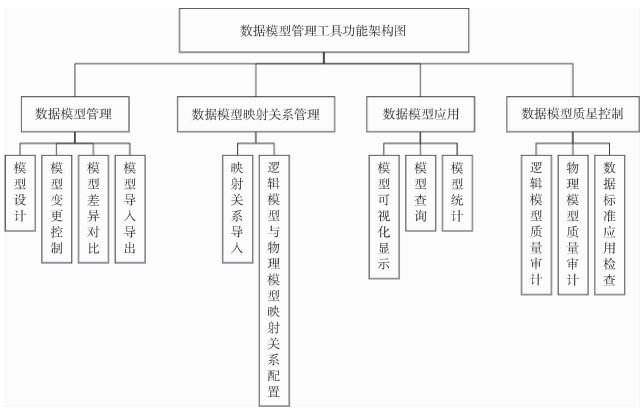

图 2 数据模型管理工具功能设计

5.1 数据模型管理

5.1.1 概述 数据模型管理包含正向与反向的建模设计、版本管理、变更控制、差异对比、导入导出等功能。方便管理人员能高效、持续管理数据模型，实现数据模型的全生命周期管理，见图 3。

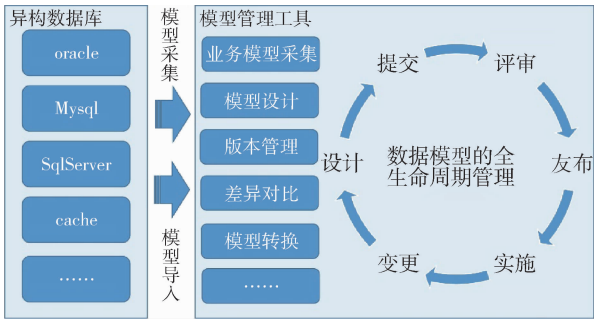

图 3 数据模型的全生命周期管理

5.1.2 业务模型采集 支持自动采集业务系统的数据结构，支持源端 Oracle、Mysql、SqlServer、Cache 等市场中的主流数据库，实现在数据模型管理工具中创建与源端一致的数据结构。

5.1.3 数据模型设计 支持管理模型、管理工具从物理模型到数据库的正向建模，通过生成数据库定义语言（Database Definition Language，DDL）在数据库中创建数据表；支持对原有系统的逆向工程能力，可根据数据库系统表信息生成数据模型；支持在字段中引用数据标准，在模型工具中可以全面查阅和寻找已发布的数据标准，并将其和数据字段进行绑定，实现数据模型的引标落标。

5.1.4 数据模型版本管理与变更控制 DAMA 管理体系认为数据模型需要保持最新状态，需求或者业务流程发生变更时都要对数据模型进行变更。数据模型管理工具支持对数据模型设计、提交、评审、发布、实施到消亡的全过程实现流程化的变更管理。支持数据模型可视化设计和修改，在模型变更时可自动生成差异化的 DDL 语句提交到测试环境中并对模型进行评审，评审通过后模型才能发布上线；提供数据模型版本化管理，可自动生成版本号以及版本变更明细信息。支持回溯任意时间点的数据模型设计状态，实现对各系统数据模型的有效管控和治理，强化医院对其数据模型的掌控能力。

5.1.5 数据模型差异对比 支持模型管理工具的数据模型与业务数据库之间的模型进行自动对比，可发现设计的数据模型和实际业务数据库中模型不一致问题。通过提供数据库表结构差异、数据关系差异的可视化报告，辅助用户监控数据模型的质量问题，提升数据模型设计和建模的质量。

5.1.6 数据模型导入导出 可以将模型文件（如 PD、ERWin 等数据 DDL 文件）直接导入到数据模型工具中生成数据模型；支持将模型工具中的模型导出为数据库 DDL 脚本，进而在数据库中创建已经规划好的模型。

5.1.7 物理数据模型转换 实现将已定义好的数据模型转换成新数据模型的功能，支持不同数据库模型之间的直接转换。通过数据模型转换功能可实现数据仓库/数据中心等各层级数据模型的快速转化，提高数据模型的结构一致性和建模效率。

5.2 数据模型映射关系管理

实现逻辑模型与物理模型的映射关系配置，同时支持管理人员通过不同格式的模板维护逻辑模型与物理模型映射关系。可通过批量导入模型格式的方式将数据模型维护到管理工具中，在映射关系通过审核后进行快速配置。

5.3 数据模型质量控制

5.3.1 概述 在完成数据模型设计后，需要对设计好的模型进行评审。评审过程按步骤分为逻辑模型质量审计、物理模型质量审计以及数据标准应用情况，见图 4。

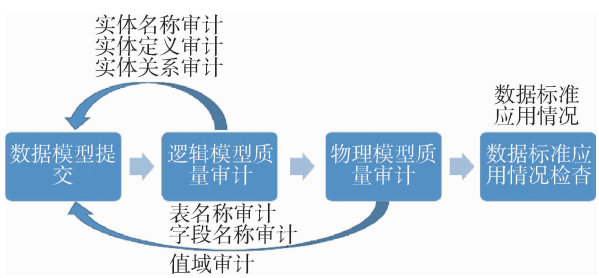

图 4 数据模型质控流程

5.3.2 逻辑模型质量审计 对模型中的实体名称与中文属性名称进行解析，审查名称是否符合数据

标准，自动检查实体定义和属性定义，对实体孤立进行审核，审查是否存在与其他没有关系的实体。

5.3.3 物理模型质量审计 系统自动对表名称、字段名称进行解析，审核其英文名称是否符合数据标准，对模型中的域进行审核，审核其是否符合标准化要求。

5.3.4 数据标准引用情况检查 实现生成数据模型对数据标准引用情况进行详细报告的功能，报告内容包括模型引用数据标准情况，为后续数据治理提供详细依据。

5.4 数据模型应用

为辅助数据模型管理人员或者决策成员清晰地了解医院数据资产，模型管理工具提供数据模型可视化显示功能，提供数据模型图形化视图，通过图表形式展示数据表以及表间关系；实现数据模型统计功能，支持对所有数据模型进行统计分析，可展示逻辑数据模型、物理数据模型的数量及标准落地情况；实现数据模型查询功能，支持逻辑数据模型查询与物理数据模型查询，可以通过表、视图、字段等属性进行对维度的查询，以便管理人员更好地组织和利用数据资产。

5.5 应用效果

医疗全量数据的数据模型管理工具建设，完成数据建模与数据生产工具、数据管理工具的一体化设计。在医院建设全量数据平台的数据建模过程中实现了模型、流程、质控、监控可视化等功能，并在建模的同时同步完成数据生产流程设计，提供一站式数据处理服务，弥补传统建模方式对数据模型管控的缺失与不足。与传统建模方式相比数据模型管理工具建模具有一定优势，见表 1。

表 1 传统建模方式与数据模型管理工具建模的效果差异对比

建模效果	传统方式建模	数据模型管理工具建模
建模速度	人工方式、速度慢	标准流程、快速建模
数据标准符合度	人工验证、易出错	自动核验、完全符合
用户体验	手工方式、步骤繁琐	图形化界面管理
模型复用	标准模型复用	标准、自定义模型均可
版本管理	人工方式、易混乱	完全支持
灰度升级	不支持	完全支持
差异化版本发布	不支持	完全支持

6 结语

在医院实际应用中，通过数据建模工具高效地完成 20 个业务系统的数据建模，转换了 49 个标准数据表，涉及数据 6.3 亿条，共管理 213 个实体、1 051种属性、109 个血缘关系。形成数据模型的命名管理、数据标准管控、值域管理、映射管理、版本管理以及数据质量管理等统一管理体系，达到了对数据模型全生命周期管理要求，为后续数据治理奠定良好基础。同时数据建模工具基本上支持市场中的主流数据库，在数据湖、科研数据库建设等涉及数据建模的工作中都具有广泛应用前景。

参考文献

1 侯丽，洪娜，李露琪，等．OHDSI 通用数据模型及医学术语标准国内应用现状分析［J］．医学信息学杂志，2020，41（2）：9.

2 周光华，徐向东，张学高，等．国家全民健康信息平台数据治理体系设计［J］．中国卫生信息管理杂志，2019，16（2）：131－134.

3 王志勇，吴骋，王立鹏，等．医疗大数据背景下的数据治理与质量监管［J］．中国数字医学，2021，16（4）：5.

4 王利亚，邱航，陈若雅．基于元数据可追溯性的健康医疗大数据治理方法及可视化呈现［J］．中国卫生信息管理杂志，2019，16（6）：6.

5 姬卫东，李琳，张振，等．互联互通背景下医疗数据治理面临的问题与对策［J］．中国数字医学，2021，16（11）：6.

6 赵霞，周毅，李琳，等．卫生信息标准开发建模技术研究［J］．医学信息学杂志，2020，41（12）：5.

7 DAMA中国分会翻译组译．DAMA 数据管理知识体系指南（第 2 版）［M］．北京：机械工业出版社，2020.