·实验室建设与管理·

基于知识图谱的实验室数据治理

李宏伟

（南京邮电大学 管理学院，南京 210003）

摘要：实验室工作中的相关数据存在种类多、类间联系复杂、存储类型多样等特点，给数据集成及基于数据的科学决策造成困难。目前实验室工作人员对于数字化管理只有初步了解，但对于在日常工作中如何具体展开，困于没有清晰的思路和方法。该文从实验室数据治理切入，介绍了一种基于知识图谱技术的数据治理方法，即以自顶向下的知识图谱构建策略为指导，首先构建实验室管理业务及相关数据集中的概念模型，然后从现有业务数据中抽取实体及实体间关系，并结合实情对存储模式进行优化设计，最后在图数据库中进行数据存储。通过对该方法的试验，初步实现了实验室数据治理，为后续基于数据的管理奠定了基础。

关　键　词：实验室；数据治理；知识图谱；Neo4j

中图分类号：G482　　**文献标志码：**A　　DOI: 10.12179/1672-4550.20240062

Laboratory Data Governance Based on the Knowledge Graph

LI Hongwei

（School of Management, Nanjing University of Posts & Telecommunications, Nanjing 210003, China）

Abstract: There are many types of related data in laboratory work, complex inter-class connections, and various types of storage, which make it difficult for data integration and data-based scientific decision-making. At present, laboratory staff only have a preliminary understanding of digital management, but there is no clear idea and method on how to carry out in daily work. Starting from the laboratory data governance, a data governance method based on the knowledge graph technology is introduced, which is guided by the top-down knowledge graph construction strategy. Firstly, a conceptual model of laboratory management business and related data sets is constructed. Then, entities and relationships between entities are extracted from the existing business data, and the storage mode is optimized according to the actual situation. Finally, data storage is performed in the graph database. Through the experiment of this method, the laboratory data management is preliminarily realized, which lays a foundation for the subsequent data-based management.

Key words: laboratory; data governance; knowledge graph; Neo4j

《中国教育现代化 2035》[1] 提出要“加快信息化时代教育变革，推进教育治理方式变革，加快形成现代化的教育管理与监测体系，推进管理精准化和决策科学化。”高校实验室相关数据是实现实验室精准化管理，进行科学决策的重要依据，对于高校建设与发展具有重要的价值。为了使相关数据能够为实验室科学管理决策所用，需要对实验室管理工作过程中所涉及的数据进行有效的治理。数据治理是数据资源及其应用过程中相关管控活动、绩效和风险管理的集合[2]，其目标是支持组织机构对自身数据的有序管理、应用并提升数据价值[3]。知识图谱是一种人工智能的底层技术，是对现实世界语义化的表示形式[4]。通过构建实验室数据资源的知识图谱，可以实现对实验室数据的治理，为各类智能化应用奠定基础。

数据治理对于提升“双一流”高校建设治理效能，强化共建、共享、共治“双一流”治理格局起着辅助和支撑作用[5]。文献 [6] 从治理理念、基础设施、核心技术、应用载体和根本保障五个方面阐述了高校大数据治理体系的优化路径。文献 [7] 从战略目标、组织架构、数据标准、数据集成等 6 个方面提出了高校数据治理框架并进行了

收稿日期：2024−02−21

基金项目：南京邮电大学实验室工作研究课题(2022XSG13)。

作者简介：李宏伟，硕士，高级实验师，主要从事实验室管理、知识管理方面的研究。E-mail: lihw@njupt.edu.cn

应用实践。在高校实验室数据治理方面，文献 [8] 建议通过扩大数据采集渠道、构建统一数据平台、完善实验室系统交互功能、重视数据分析工作等措施进行数据治理。文献 [9] 以教育部高等学校实验室信息统计数据为核心，构建了实验室教学管理知识图谱本体模型。上述研究都是从工作体系、理论分析层面进行探讨，对实验室数据治理实践有一定的参考价值，但缺乏对数据治理思路、步骤、方法以及工具的深入介绍，不便于实验室一线管理人员开展数据治理实践。本文采用基于知识图谱的数据治理方法，通过实验室工作相关数据概念模型的设计，对数据类与类之间的关系进行了梳理，并使用 Neo4j 图数据库进行了实验室数据的存储试验，目的是对实验室数据治理的思路、方法及工具进行实践，为实验室数据治理工作的深入开展积累经验。

1　实验室数据及特点

本文以南京邮电大学管理学院文科实验室为例。文科实验室注重学科重组、文理交叉，把新技术融入文科实验课程，为学生提供综合性的跨学科实践的实验室[10]。文科实验主要是对社会、企业等不同场景的模拟、仿真，以计算机为载体，以软件为实验平台，主要仪器设备是计算机、软件、VR\AR 等[11]。文献 [12] 建议新文科实验室建设过程中，应打造智能化综合管理平台，优化管理流程、提升管理效率及管理服务质量。若要实现对实验室的智能化管理，需要对实验室管理相关的数据进行治理。

学院管理类文科实验室主要是教学型实验室，相关管理工作包括：实验室资源管理，如实验场地、设备管理，实验室预约安排，软硬件维护；实验教学管理，如课内外实验项目管理，教学资料管理等；实验室教学辅助工作，包括协助实验教学，开发及改进实验项目等；实验室安全工作，如安全隐患巡检及整改，安全制度规范及知识的管理等；实验室建设工作，如建设项目申请撰写，设备的采购等。在这些具体工作中，至少涵盖 5 类数据，这些数据存在以下 3 个特点。

1）数据种类多，数据间关系复杂

实验室工作直接对接的上级部门分别是实验室处和教务处，几乎同时与学院所有人群都有交集。在所有管理业务中，至少包含基础、人员、课程、设备、文档等 5 类数据，各项业务之间也存在着一定的重叠交互，使得数据类与类之间存在着复杂的联系。

2）数据存储类型多样

实验室管理工作信息化水平不高，各类数据的保存类型主要以 Word、Excel 文档为主，也有部分业务基于信息系统，相关数据以关系型数据库存储。此外，从数据结构化视角来看，存储类型还涉及结构化数据和非结构化数据，常规的数据统计表都以结构化数据为主，但规章制度、软硬件操作规范等则是以文档、视频等非结构化形式存储。

3）数据存储位置分散

实验室管理人员面向的是工作第一线，后端对接多个上级部门的不同业务，工作有一定的分工。因信息化不充分，使得数据以文本类型为主，相应的数据也分散地保存在各业务负责人手上。

2　实验室数据治理现状分析

实验室管理相关数据呈现种类繁多、关系复杂、类型多样、存储分散的特点，但实际中又缺乏数据治理方面的深入工作，使得数据存在孤岛化、共享度低、一致性差等问题，难以发挥数据的价值。实验室数据治理工作的欠缺，主要体现在数据治理意识和工作方法及工具两个层面。

1）数据治理意识层面。数据管理是利用计算机软硬件对数据进行收集、存储、处理和应用的过程，目的是发挥数据的效用。数据治理是发挥数据管理功能的保障。实验室工作人员的工作重心主要在于处理日常业务，没有形成基于数据进行精细化管理的意识，在数据管理和数据治理方面都缺乏有效的工作。针对意识层面的不足，最重要的是通过案例学习，让实验室工作人员看到精细化管理的效用，明确用数据驱动实验室管理的目标，从而在日常工作中树立数据治理的意识。

2）工作方法及工具层面。目前，没有有效的工作方法用于梳理现有数据及数据间的关系并进行治理。大部分数据都是以二维表格的文档形式进行存储，难以发现数据间一致性问题，也无法有效发掘数据蕴含的规律，如在做设备采购需求调研时，无法有效查询已经采购的设备是否可用于新实验相关知识点的训练等。针对该层面的不足，需要选择工作人员易于学习及操作的方法和

工具。

知识图谱是以图的形式呈现客观世界中的概念和实体及其之间关系，其构建过程目标明确，步骤清晰，易于理解和操作，可以作为实验室数据治理工作方法。目前，知识图谱有 3 种构建策略：自底向上、自顶向下和二者混合的策略[13–14]。在实验室数据治理工作中，可以借鉴自顶向下的知识图谱构建策略，其中包括本体构建和实体识别两个主要环节。本体构建主要任务是明确数据所属概念及概念间联系，完成这项工作，需要实验室工作人员清楚相关工作的业务流程，以及在流程中涉及的对象及类别。实体学习主要任务是建立实体与概念之间的映射，这需要实验室人员对工作过程中所处理的实体相关数据有正确的认识和理解。最后，可以借助 Neo4j 进行知识图谱的存储及可视化。

相比传统的表格治理方式，基于知识图谱的数据治理方式主要有以下两点优势。

1）能够让实验室工作人员对业务和相关数据有更深入的理解。传统的数据处理方式，工作人员只忙于对基本表格的填充，并不关心数据之间的联系。基于知识图谱的治理，需要实验室工作人员从业务流程中实体与实体间的联系入手，在填充实体相关数据的同时，还要明确实体之间的联系，有助于深入理解业务。同时，这也是一种管理知识沉淀的方式，阶段性成果可以作为新员工培训的资料，为实验室工作的快速交接提供保障。

2）借助知识图谱的可视化操作能从多方面提升人员工作效率，包括快速理解业务相关数据，快速发现数据内在规律和趋势，保证较高的分析准确性，这些都是使用传统表格治理数据所不具备的。

基于上述原因及从人工智能发展的趋势来看，将知识图谱技术应用于实验室数据治理的实践是一种有现实意义的探索。

3 基于知识图谱的数据治理工作方法

基于自顶向下的知识图谱构建策略，实验室数据治理工作的步骤可以分为以下 3 个步骤：

1）实验室管理人员从业务层面梳理数据所属相关概念及概念间联系，构建知识图谱的概念模型；

2）工作人员基于所梳理的概念模型和工作相关数据报表，建立概念与实体之间的映射，明确实体间的联系；

3）在信息技术人员的协助下完成图数据存储。

3.1 实验室数据概念模型构建

实验室管理人员通过概念模型构建工作，可以实现对实验室相关业务的统一理解，也会对业务中所使用的数据及存储要求有一定的认识，这将有助于管理人员后续工作的开展，为工作中高质量收集实验室数据打好基础。

概念模型构建过程中，实验室管理人员可以从各项管理流程以及相关的数据集中抽取出概念及概念间的关系（或称为类及类间的关系）。管理业务如本文第 1 节所述，数据集可包括：实验中心（室）相关信息统计工作中涉及的实验室基本情况表、专任实验人员表、实验设备情况表等；实验室绩效考评工作中涉及的资产利用率计算明细表、实验室绩效考评表等。概念模型的描述，可采用框图描述，如图 1 所示，形式上易于实现及理解，便于一般管理人员掌握及使用。

本文结合学院管理类文科实验室工作中的相关概念和数据集，总结了 8 个概念类，部分类别中包含子类。如人员类中包含了实验室管理员、教师和学生类；将课内实验、STITP 项目、学科竞赛、实验室开放项目统一归类到课程大类中；将设备采购、维护及异常解决，实验室安全巡查及整改等业务统一归类到事件大类中；将各项工作中涉及的文档统一归结到文档类。类之间的关系用于描述不同类之间的交互模式和逻辑联系，如学生课程之间的“选修”关系，实验项目与设备之间的“使用”关系等。最终，根据主要类及类之间的联系构建出实验室数据知识图谱的概念模型，如图 1 所示。

3.2 实体及实体关系获取

本阶段实验室管理人员需要基于前文所梳理的概念模型，并结合工作中的数据表，建立概念与实体之间的映射规则，以及明确实体间关系的构建规则，最后借助工具进行自动处理。实体及实体关系通常是指清楚的、事实性的信息，这些信息来源和结构可能不同，相应的获取方法也不同。如从结构化关系型数据库中可以采用 D2R 工具，从非结构化文本中获取则需要通过自然语言处理技术实现信息抽取。

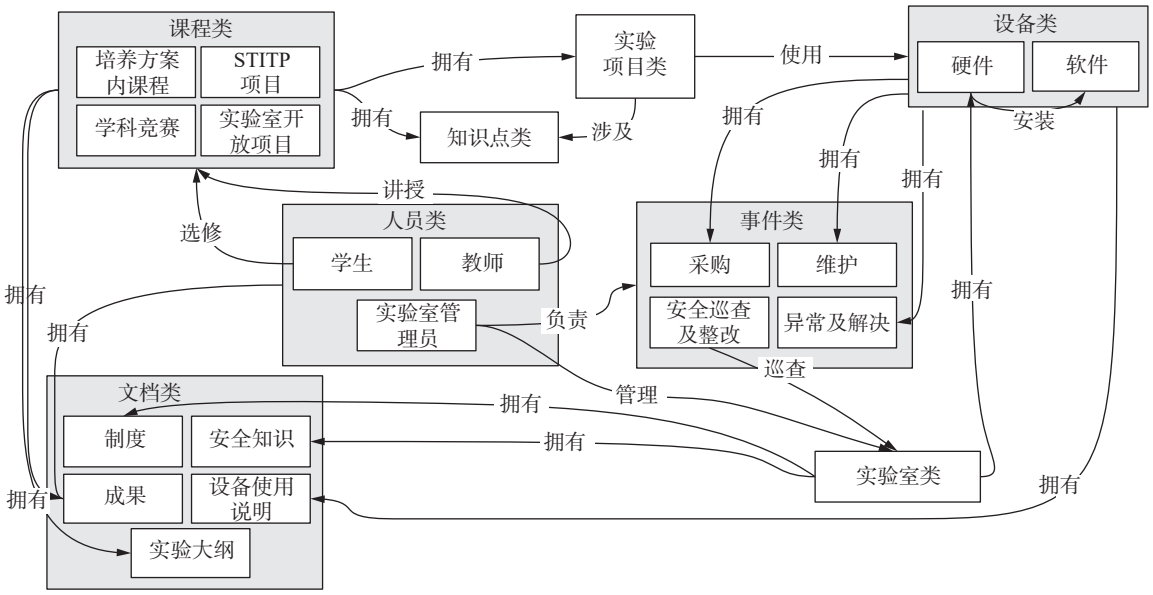

图 1　实验室数据所属类与类间关系示意图

本文中对实体及实体关系的获取主要包括以下 3 个步骤：

1）建立概念和实体之间的映射规则，并确定实体的相关属性，以实验室管理员类及相关实体映射为例，先汇总实验室管理员类相关的 Excel 数据表，包括“专任实验室人员表”“实验室基本情况表”等，然后将各表格中工号相同的数据整合成一条记录，一行记录表示一个实体，实体属性综合了数据表中的相关列名；

2）依照 Excel 数据源表内部及表之间的数据联系，明确各实体之间关系的构建规则；

3）依照前两步确定的相关规则，用 Python 语言编写数据处理程序，从原始工作表中抽取相关信息，转换成实体及实体间关系数据集，以 csv 格式文件进行保存。

3.3　实验室数据存储试验

3.3.1　数据存储前的存储模式设计及完善

本阶段主要的工作是在实验室数据的概念模型的基础上，结合业务场景实情，对实体相关数据的存储模式进行优化设计，数据存储模式的设计工作对于实验室一般管理人员有一定的难度，需要有具备一定计算机技术能力的信息技术人员协助完成。

目前知识图谱主流的存储方法是采用图数据库进行存储，图数据库存储能有效利用其以关联数据为中心的数据表达、存储和查询。在现有的图数据库中，Neo4j 是一款开源、稳健、可伸缩的高性能图形数据库[15]。在 Neo4j 中，基本元素包括：标签（label）、节点（node）、关系（relationship）以及属性（property）[16]。节点表示的是实体，通过多个键值对形式描述其属性，通过标签标记其类型。标签本质上是类，用以标记某个节点归属于某个类，一个节点可以贴多个标签，表示某个实体归属多个类。关系用来描述两个实体之间的联系，关系也有属性，通过关系可以展现数据之间深层次的关联。本文根据实验室管理工作的相关数据实情及多次实验，对数据的图存储模式做了以下两个方面的设计。

1）父子类关系处理

父子类之间是从属关系，可以通过关系形式来体现，也可以通过对同一个节点贴两个标签的方式实现。在保证数据检索需求的前提下，为了减少数据可视化时关系繁多的情况，本次试验采用了后一种方法。

2）两个相同实体之间发生的多次联系情况的处理

如同一个教师在不同学期承担相同的课程，需要在教师和课程之间建立具有不同学期属性的关系。如果这样，在数据可视化时，两个实体之间将有多个同名关系，不利于对数据的检索、观察及理解。本文将学期属性作为关系名的组成部分进行存储，如将关系定义为 [:承担 2022_2023_1]。

3.3.2　数据存储试验

本阶段的工作主要由信息技术人员完成，信

息技术人基于图存储模式，基于已获取的用于试验的部实体相关数据，在 Neo4j 中通过 Cypher 语言进行实体及实体间的联系的创建，进行图数据库中的存储试验。本试验创建工作包括以下方面。

1）节点创建

以创建人员实体为例，“管理员王一一”节点的创建代码如下，其中“管理员”是其标签，表示其所属类别，大括号中的键值对描述其属性。

CREATE (s:管理员 {name:'王一一',id:'20200011', gender:'女', phone:'13812345678', birthYear:199705, title: '实验师', position: '一般', eduDegree: '硕士研究生', type: '实验室建设与管理'})

2）标签创建

以父子类处理为例，如对已创建的节点贴第二个标签时，先查找实体，进一步通过 SET 指令向其添加第二个类标签“人员”。

MATCH (s:管理员 {name: '王一一'}) SET s:人员 RETURN s

3）关系创建

以同一教师在不同学期任课为例，将不同学期的关系分别定义为 [:GUIDE2022_2023_1] 和 [:GUIDE2021_2022_1]。具体创建代码如下：

MATCH(m:`教师` {name:'黄蓉'}) MATCH (n:`课程` {name:'管理信息系统'})

MERGE(m)-[:GUIDE2022_2023_1]->(n);

MATCH(m:`教师` {name:'黄蓉'}) MATCH (n:`课程` {name: '管理信息系统'})

MERGE(m)-[:GUIDE2021_2022_1]->(n);

在试验数据录入结束后，可以进行图形化展示，能够直观地查看到所有实体及实体间的关系，部分节点及关系如图 2 所示。同时，实验室管理人员可以使用 CALL db.schema.visualization 指令输出图存储模式，结合已有的概念模型，检查类与类之间关系定义的正确与否，若有问题，可以进一步调整概念模型及存储模式。在确定概念模型及存储模之后，参照 3.2 节中的方法，进行实验室管理整体数据抽取与转换，最终采用 Neo4j 提供的 LOAD CSV 方法实现批量数据录入。

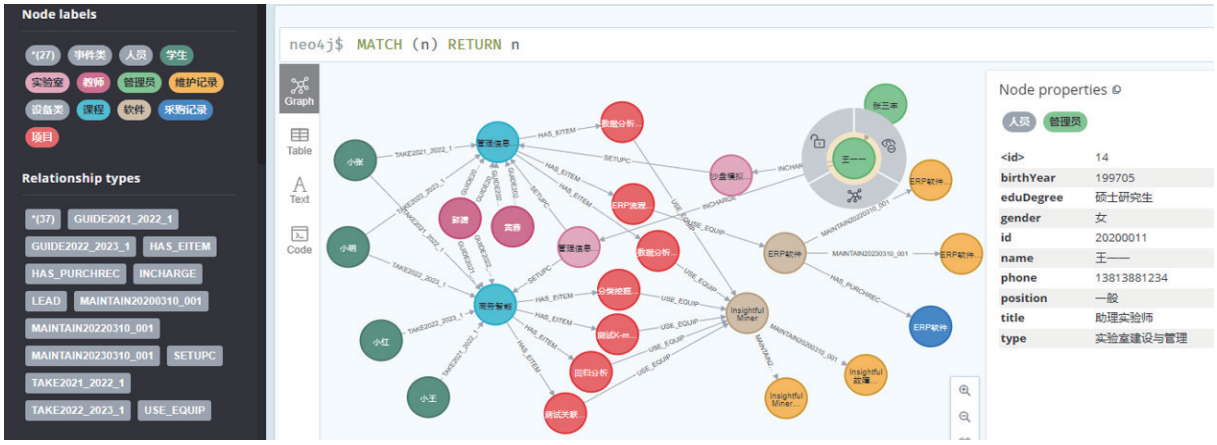

图 2 数据局部视图

至此，从实验室业务流程及相关数据的梳理出发，通过构建数据概念模型、获取实体及实体关系，设计图存储模式，到最后的数据录入以及可视化验证，基本实现了实验室数据治理工作，所创建的结果可为软件采购等实验室工作的科学决策提供依据。这一流程可以作为实验室一线工作人员参与并开展实验室数据治理工作的思路及方法。过程中涉及实验室一般管理人员与信息技术人员，对于占人员组成主体的一般管理人员来说，他们可以参与概念模型的构建、概念模型与实体及实体间关系的映射、图存储模式的设计完善及基础数据的梳理工作。

4 结束语

随着人工智能、知识库等数字技术的发展及应用，数字化转型已经在各个行业逐步推进。高校实验室的数字化转型也迫在眉睫，应在完善信息化的基础上，借助数字技术，以数据驱动管理创新，提高实验室管理水平和成效。数据治理作为数字化转型的基础性工作，重要性不言而喻。数据治理是从目标、组织、管理、技术、应用的角度持续提升数据质量的过程。本文对实验室数

据治理的方法和工具进行了尝试，将知识图谱相关方法与技术，作为实验室管理人员开展数据治理工作的思路、方法和工具，对相关工作内容和方法进行了介绍说明，并结合南京邮电大学管理学院实验室的具体情况，进行了应用及试验，初步实现了实验室数据治理。在后续的工作中，还需要从数据治理的组织目标、组织架构、数据标准等方面不断推进实验室的数据治理。对于数据的存储，在后续工作中还需要结合实验室管理工作中的实际应用情况进行重构提升。

参 考 文 献

[1] 中共中央, 国务院. 中国教育现代化 2035[EB/OL]. (2019-02-23)[2024-01-03]. http://www.moe.gov.cn/jyb_xwfb/gzdt_gzdt/201902/t20190223_370857.html.

[2] 国家市场监督管理总局, 国家标准化管理委员会. 信息技术服务治理第 5 部分: 数据治理规范: GB/T 34960.5—2018[S]. 北京: 中国标准出版社, 2018.

[3] 栾瑞鹏, 张静, 刘立坤. 面向装备试验鉴定领域数据治理的知识图谱本体构建[J/OL]. 系统工程与电子技术, 2022. (2022-12-29)[2024-01-03]. https://kns.cnki.net/kcms/detail/11.2422.TN.20221228.2006.024.html.

[4] 夏毅, 兰明敬, 陈晓慧, 等. 可解释的知识图谱推理方法综述[J]. 网络与信息安全学报, 2022, 8(5): 1−25.

[5] 周江林. 数据治理: 大数据时代“双一流”建设的路径优化取向[J]. 教育发展研究, 2022, 42(5): 15−21.

[6] 胡水星, 荆洲, 王会军. 我国高校大数据治理体系的关键要素与优化路径研究: 基于 DEMATEL-ISM 的研究视角[J]. 电化教育研究, 2022, 43(11): 38−44.

[7] 郝志杰, 李莉, 荣娟. 数据治理在解决“一张表”问题中的实践[J]. 实验室研究与探索, 2019, 38(12): 261−265.

[8] 邱坤, 郭盛, 顾亦然, 等. 高校实验室信息化数据治理的探索[J]. 实验室研究与探索, 2022, 41(10): 265−269.

[9] 冯健文, 黄贤群, 林璇. 高校实验室教学管理知识图谱模型构建研究[J]. 电脑知识与技术, 2021, 17(23): 199−201.

[10] 陈兰. 加强实验室建设 培养应用型人才[J]. 实验室科学, 2017, 20(1): 173−175.

[11] 胡菲菲, 张思思. “新文科”背景下高校文科实验室建设特点与趋向[J]. 实验技术与管理, 2023, 40(1): 221−226.

[12] 王娜, 张应辉. 高水平本科教育背景下新文科实验室建设路径探索[J]. 实验技术与管理, 2020, 37(1): 32−35.

[13] 黄恒琪, 于娟, 廖晓, 等. 知识图谱研究综述[J]. 计算机系统应用, 2019, 28(6): 1−12.

[14] 王传庆, 李阳阳, 费超群, 等. 知识图谱平台综述[J]. 计算机应用研究, 2022, 39(11): 3201−3210.

[15] 赵雪芹, 李天娥, 曾刚. 基于 Neo4j 的万里茶道数字资源知识图谱构建研究[J]. 情报资料工作, 2022, 43(5): 89−97.

[16] 王红, 张青青, 蔡伟伟, 等. 基于 Neo4j 的领域本体存储方法研究[J]. 计算机应用研究, 2017, 34(8): 2404−2407.

编辑　王燕