ISO 国际标准知识图谱的构建方法研究

方思怡

（上海市质量和标准化研究院）

摘　要：标准数字化转型对标准知识组织形式和相关服务模式提出了全新的要求。知识图谱是标准数字化转型的关键核心技术之一，能有效解决ISO标准信息和知识服务在数据颗粒度和关联性等方面的局限性。本研究聚焦标准知识服务重点关注的ISO标准核心要素，通过深入分析其文本结构特性，在综合比较不同ISO标准数据存储格式后，提出适用于ISO的标准知识图谱构建方法，并在塑料制品、橡胶等市场监管关注的领域开展初步应用，以期能够为标准数字化转型提供一定的技术参考。

关键词：标准知识图谱，ISO，国际标准，标准数字化，实体抽取，知识组织

DOI编码：10.3969/j.issn.1674-5698.2024.12.012

Research on the Construction Method of ISO International Standard Knowledge Graph

FANG Si-yi

（Shanghai Institute of Quality and Standardization）

Abstract: The digital transformation of standards has put forward new requirements for the organizational form of standard knowledge and related service models. Knowledge graph is one of the key core technologies for standard digital transformation, which can effectively address the limitations of ISO standard information and knowledge services in terms of data granularity and correlation. This study focuses on the core elements of ISO standards which are the key focus of standard knowledge services. Through in-depth analysis of their text structure characteristics and comprehensive comparison of different ISO standard data storage formats, a standard knowledge graph construction method suitable for ISO is proposed, and preliminary applications are carried out in the fields of market supervision concern such as plastic products and rubber, in order to provide technical reference for the digital transformation of standards.

Keywords: standard knowledge graph, ISO, international standards, digitalization of standards, entity extraction, knowledge organization

0　引言

标准是经各利益相关方协商一致形成的技术性文件。在不同类型的标准中，国际标准是在全球范围内广泛使用的技术性制度工具[1]。作为世界范围内影响力最大的标准化组织，ISO历来重视标准的推广应用[2]，联合IEC共同提出了机器可读标准的概念和相应的解决方案。近年来，面向机器可读标准的标准数字化研究已逐渐成为标准领域的重大战略方向，并催生了标准化工作的极大变革。

基金项目：本文受上海市质量和标准化研究院院立项目“国际标准核心要素标注方法研究”（项目编号：YRY202406）资助。

作者简介：方思怡，硕士研究生，工程师，研究方向为标准数字化、标准知识服务、标准数据挖掘与分析。

随着数字经济时代的深入发展和人工智能、大数据等技术的不断普及，标准化工作已步入数字化发展的新阶段[3]，进而对标准情报服务提出了全新的发展要求[4]。在标准情报服务中，标准信息服务和标准知识服务的质量与标准数字化技术的应用深度密切相关。当前国内外机器可读标准的能力等级普遍处于较低水平，标准信息和知识服务大多存在服务手段单一、技术方法落后、数据颗粒度不够细等问题[5]。在数字化转型的背景下，标准信息和知识服务亟需实现多元化、细粒度、深层次、关联性的数据挖掘与组织形式。作为业内标准数字化转型公认的关键核心技术[6, 7]，知识图谱是一种以图形式存储和表征大规模数据及其关系的结构化知识库[8, 9]，因此在标准知识组织方面享有一定的优势，能体现不同标准核心要素的关联性，并提供面向特定标准应用的图谱计算以支撑标准化活动的相关决策。

当前国内外的标准数字化技术研究大多处于初级阶段，知识图谱在ISO国际标准中的应用还存在较大的提升空间。本研究从机器可读标准的视角出发，通过深入分析ISO标准核心要素的文本结构特性，在综合比较不同ISO标准数据存储格式后，提出适用于ISO的标准知识图谱构建方法，并选取特定领域开展初步应用，以期能够为标准数字化相关工作提供一定的技术参考。

1　ISO国际标准知识图谱的相关概念

本研究将标准知识图谱界定为专业知识图谱在标准领域的一大分支。标准知识图谱是以标准文本及相关数据为来源、经由一定技术所形成的结构化知识库，通过图的形式来组织和存储标准知识[10]。与常规的知识图谱类似，标准知识图谱在逻辑架构上可以分为模式层、数据层和应用层，其中模式层用来存储标准知识的本体概念，也即标准核心要素的类型；数据层用来存储模式层对应的实例数据；应用层则涵盖了标准知识图谱所涉及的智能计算。就数据类型而言，标准知识图谱通常由标准实体和标准关系构成，其中标准实体通常是指标准文本中的具体核心要素，例如：标准名称、标准号、标准指标、标准分类号等，而标准关系则是用来描述标准实体之间的具体联系。

ISO国际标准知识图谱是标准知识图谱面向ISO文本的特定类型，以ISO文本及相关数据为来源，旨在存储和表征ISO文本中的标准核心要素及其关联性特点。本研究参考了ISO的文本编写要求[11]，并紧密结合标准数字化发展的业务需求和ISO标准核心要素的文本结构特性，在已有研究的基础上[12]设计了ISO国际标准知识图谱的模式层。本研究重点聚焦标准信息服务和知识服务所关注的ISO标准核心要素，包括标准号、标准英文名称、标准发布时间、标准版本情况、标准化技术委员会、被代替标准号、标准ICS分类号、标准范围、标准规范性引用文件名称及标准号、标准术语、标准指标和基于标准全文的标准主题关键词。在上述标准核心要素中，除了基于标准全文的标准主题关键词外，均直接来自于标准文本。与其他核心要素相比，标准术语和标准指标在构成上更为复杂，可根据具体的元素功能进一步划分为更小的知识单元，其中标准术语可以划分为标准术语编号、标准首选术语、标准首选术语缩略语、标准弃用术语、标准术语同义词、标准术语领域、标准术语定义、标准术语示例、标准术语条目注释、标准术语来源；标准指标则可以划分为标准指标名称、标准关联指标名称、标准下一级指标名称、标准指标值及单位、标准指标定义、标准指标描述说明、标准指标表格名称、标准指标条目注释、标准指标示例、标准指标符号、标准指标偏好值、标准指标最小值、标准指标分类和标准指标下一级分类。本研究以上述ISO标准核心要素为ISO国际标准知识图谱模式层的标准实体类型，以标准号与上述核心要素的名称指向形式“ISO标准核心要素+是”（例如：Term definition is）的英文表述为标准关系类型。

ISO国际标准知识图谱模式层的框架图如图1所示。

2　ISO国际标准知识图谱的构建方法

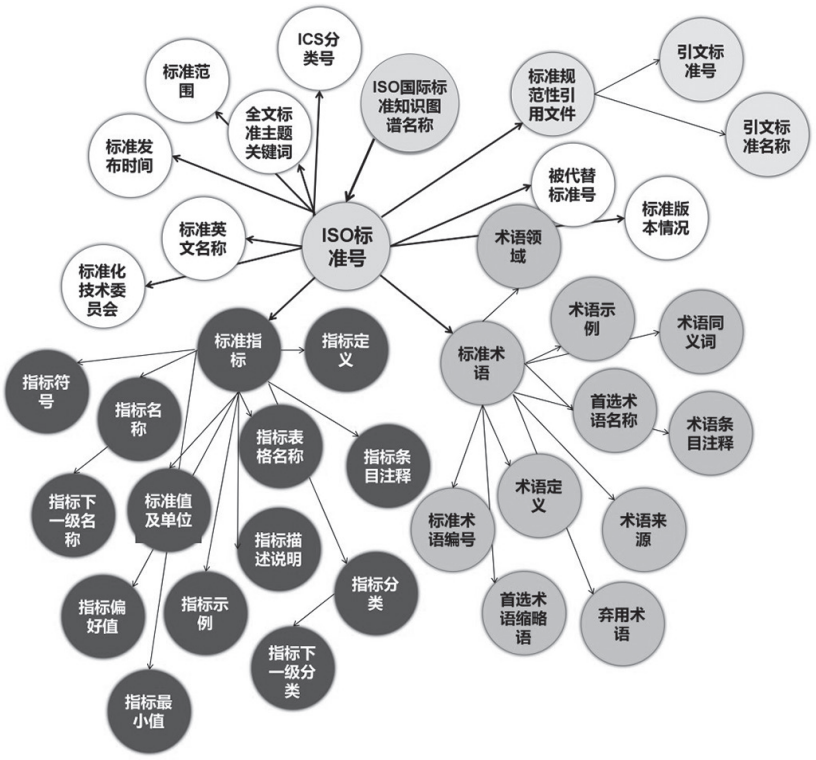

图1 ISO标准知识图谱模式层的框架

2.1 ISO国际标准文本数据资源的比较

ISO文本数据是ISO国际标准知识图谱的知识来源，因此在开展ISO国际标准知识图谱的构建之前，本研究选取上海市质量和标准化研究院标准文献馆中的部分ISO馆藏资源，将ISO标准的数字PDF文本转化为TXT、Word docx和XML文档，对PDF、TXT、Word和XML格式的ISO文本采用人机协作的方式进行数据读取和标注效果的比对分析，所得的比较结果见表1。

经过详细比对可知，在具有高质量PDF资源的前提下，ISO的XML文本在数据读取准确性、表格数据的语义完整性和关联性、数据标注的效果等方面表现优于其他类型的文本格式，因此本研究选取ISO的XML文本作为ISO国际标准知识图谱的数据来源格式，在ISO的XML文本数据基础上开展标准实体抽取和标准知识图谱的构建工作。

2.2 ISO国际标准知识图谱的构建流程

标准是具有明确编写规范的技术性文件，ISO标准文本在章节结构、要素构成、层次编排等方面遵从一定的编写要求。本研究从ISO标准的文本编写特点入手，结合知识图谱的通用性构建步骤，制定了适用于ISO国际标准知识图谱的构建流程，主要包括ISO标准知识获取、ISO标准知识表示、ISO

表1 不同格式的ISO文本数据的数据读取和标注效果比较

文本格式	数据读取准确性	是否包括图片信息	表格数据的语义完整性	表格数据的语义关联性	数据标注的效果
PDF文本	高度依赖读取PDF数据的工具水平，目前普遍不高，存在一定比例的数据读取错误	包括	表格的读取形态取决于PDF数据读取的工具水平，整体语义完整性较强	表格数据可直接衔接上下文的段落数据，与上下文的语义关联性较强	可通过PDF软件自带的标注功能进行数据标注，需将标注内容与PDF文本中的数据映射，映射的质量取决于PDF数据读取的准确性
TXT文本	视PDF文本的质量而定，当PDF文本质量较高时，经由PDF转化而成的TXT文本的数据读取准确性较高	不包括	无法保留表格的基本形态，语义完整性一般	同上	不支持TXT阅读软件中直接标注数据
Word docx文本	同上	包括	表格形态完整，整体的语义完整性不够强	表格数据在机器读取时无法直接衔接上下文的段落数据，语义关联性较弱	可直接在Word文本的软件中直接标注数据
XML文本	同上	包括	表格形态完整，整体语义完整性强	表格数据直接衔接上下文的段落数据，语义关联性较强	可通过在Word文本的软件中标注数据后转为XML文本

标准知识存储和可视化这4个步骤（如图2所示）。

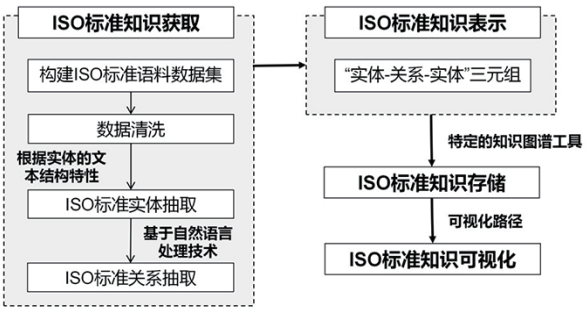

图2 ISO国际标准知识图谱的构建流程

在ISO标准知识获取阶段，旨在将ISO标准文本转化为机器可直接读取的ISO标准语料数据集，在完成数据清洗后采用特定技术抽取ISO标准实体，采用自然语言处理和文本挖掘技术批量生成ISO标准实体对应的ISO标准关系。在ISO标准知识表示阶段，主要是将上一个阶段获取到的ISO标准实体和ISO标准关系转化为ISO标准知识图谱的基本数据存储形式，也即实体关系三元组（实体—关系—实体），最终形成ISO标准实体关系三元组结构化数据集。在完成ISO标准知识表示后，采用特定的知识图谱工具读取ISO标准实体关系三元组结构化数据集，完成ISO标准知识存储和可视化呈现。

2.3 ISO国际标准知识图谱的标准实体抽取方法

在深入分析ISO标准核心要素的文本结构特性的基础上，本研究提出了适用于不同ISO标准核心要素的标准实体抽取方法，主要分为基于规则的文本挖掘方法、基于有监督的深度学习方法和基于无监督的机器学习方法，其中标准号、标准名称、标准发布时间、标准化技术委员会、标准版本、被代替标准号、标准规范性引用文件名称及标准号、标准范围、标准术语可采用基于规则的文本挖掘方法来自动抽取；基于标准全文的标准主题关键词则需要采用无监督的机器学习方法来获取，鉴于潜在狄利克雷分布主题模型（Latent Dirichlet Allocation Topic Model, LDA Topic Model）的理论相对成熟，本研究采用LDA主题模型来自动获取ISO标准全文范围内的主题关键词；而对于不存在明显规则的标准指标，考虑到当前尚未推出高度适配于ISO标准文本的大语言模型（Large language model, LLM），本研究采用了有监督的深度学习方法，基于循环神经网络模型（Recurrent Neural Network, RNN）及其亚型组合，通过人工标注指标数据和训练神经网络模型的方式实现自动抽取术语的相关实体。

2.4 ISO国际标准知识图谱的可视化路径实现方法

在综合比较不同的图谱可视化工具后，本研究选取Neo4j平台作为ISO标准知识图谱的存储和可视化呈现工具。作为当前应用最为广泛的图数据库，Neo4j自带包括构建Web应用程序、机器学习图算法以及图计算与分析相关的Graph Data Science Library（GDS库）等工具的大型生态系统，可充分满足ISO国际标准知识图谱的快速存储和功能模块研发需要。本研究采用Python编写了调用Neo4j平台的程序，实现了ISO国际标准知识图谱中各个实体和关系的可视化路径。

3 ISO国际标准知识图谱的应用

在形成ISO国际标准知识图谱的构建方法后，本研究在综合考量上海市市场监管的业务需求基础上，选取与民生密切相关的塑料制品、油漆和清漆、橡胶及橡胶制品等领域的7篇ISO文本开展ISO国际标准知识图谱的构建方法验证与初步应用。经过统计可知，上述小样本ISO国际标准语料数据集共计231,655个字符，生成的ISO国际标准知识图谱涵盖了1251个标准实体和1474个标准关系。ISO国际标准知识图谱的Neo4j平台界面截图如图3所示。

4 总结与展望

4.1 总结

在标准数字化转型的背景下，本研究紧密结合标准信息和知识服务的业务发展需求，通过深入分析ISO国际标准的文本结构特性，聚焦标准信息和知识服务所重点关注的ISO标准核心要素，经详细比对ISO不同格式的文本特点后，以XML格式的ISO文本为数据来源，采用基于规则和深度学习相结合的技术打造了适用于ISO的国际标准知识图

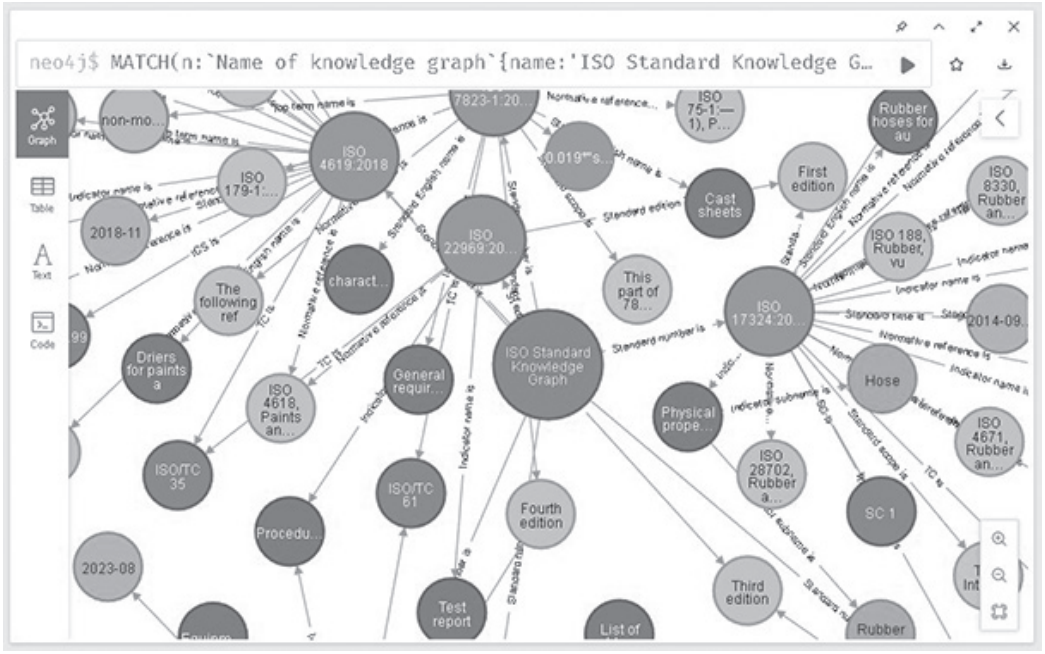

图3 ISO国际标准知识图谱的界面截图

谱构建方法，并在塑料制品、橡胶等领域的小样本ISO数据集上开展方法验证和初步应用，为后续的标准知识库构建和相关标准知识服务提供必要的技术支撑。

4.2 展望

标准数字化转型是标准未来发展的必然趋势。本研究将在后续工作中围绕以下几个方向开展深入研究：（1）拓展标准知识图谱的数据规模和应用功能，尝试提取ISO标准文本中图片数据的技术信息；（2）继续追踪以大语言模型为例的前沿技术，优化当前ISO标准知识图谱的实体抽取方法；（3）在ISO标准知识图谱的基础上进一步形成融合标准、专利、论文、法规等文件的标准综合知识库，为构建适用于标准领域的标准大语言模型和打造更为丰富的标准知识服务奠定数据基础。

参考文献

[1] 张晓刚. 国际标准化发展的新趋势[J]. 质量与标准化, 2022(10):1–4.

[2] 张宝林,侯常靓,邬雨笋,等.国际标准化组织机器可读标准工作动态[J]. 信息技术与标准化, 2022(10):18–22.

[3] 崔静,王立玺. 标准数字化工作路线图探究[J]. 信息技术与标准化, 2023(06):43–46.

[4] 彭国超,刘婕,张冰倩. 我国标准情报服务的分类及发展现状研究[J]. 情报科学, 2022,40(10):179–186.DOI:10.13833/j.issn.1007–7634.2022.10.023.

[5] 郝文建,魏梅,张浩,等. 标准知识图谱的构建与应用[J]. 信息技术与标准化, 2021(08):44–47.

[6] 范昊,王一帆. 知识关联视角下标准文档的多粒度知识组织方法研究[J]. 信息资源管理学报, 2024,14(04):133–145. DOI:10.13365/j.jirm.2024.04.133.

[7] 王一禾,吕千千,祝贺. 标准数字化转型关键技术及其应用分析[J]. 信息技术与标准化, 2022(10):51–55+59.

[8] 穆天杨,陈华达,杨玉婷,等. 知识图谱技术在机器可读标准中的应用[J]. 信息技术与标准化, 2022(10):56–59.

[9] 王萌,王昊奋,李博涵,等. 新一代知识图谱关键技术综述[J]. 计算机研究与发展, 2022,59(09):1947–1965.

[10] 方思怡. 标准知识图谱的技术路径与应用场景探讨[J].中国标准化, 2023(11):49–55.

[11] ISO/IEC Directives, Part 2: Principles and rules for the structure and drafting of ISO and IEC documents [S].

[12] 方思怡.基于文本挖掘的ISO标准术语自动识别与标准术语知识图谱构建研究[J]. 标准科学, 2024(08):84–89.