ICS 35.240.30
CCS L70

T/CCUA

中 国 计 算 机 用 户 协 会 团 体 标 准

T/CCUA 043—2024

文献资源知识图谱构建 技术要求

Constructing knowledge graph of literature resources - Technical requirement

2024－12－16 发布　　2025－1－16 实施

中国计算机用户协会　　发 布

目　　次

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

本文件由中国计算机用户协会提出。

本文件由中国计算机用户协会归口。

本文件起草单位：中国国家版本馆、中国计算机用户协会创新技术应用分会、中南出版传媒集团股份有限公司、《全国新书目》杂志有限责任公司、天闻数媒科技（湖南）有限公司、湖南大学、中电长城科技有限公司、星环信息科技（上海）股份有限公司、湖南超绘智能科技有限公司。

本文件主要起草人：刘成勇、王志庚、杨俊杰、张 琦、唐 卓、胡昌华、林 峰、刘 剑、刘轶铭、耿 锐、马 驰、马腾飞、田 维、李谟毫、张嘉鹏、邹 璞、肖丽晶、刘杨兵、邓 筱、刘 斌、符利华、李 苏、郭 峰。

引　言

随着信息技术的快速发展和互联网的普及应用，知识图谱作为一种结构化知识表示和组织方法，在各个领域的知识管理和智能应用中发挥着越来越重要的作用。知识图谱是实现文献资源智能应用的重要基础，同时利用大语言模型在语义理解、内容生成等方面的技术优势，实现大语言模型对知识图谱构建至知识图谱应用各环节的增强，提升知识图谱构建效率和质量。在实际应用中，为了保证知识图谱的质量和可用性，需要制定一套文献资源知识图谱构建的标准流程。

文献资源知识图谱构建 技术要求

1 范围

本文件确立了文献资源知识图谱架构和构建流程，规定了相关技术要求。
本文件适用于相关组织文献资源知识图谱的开发和维护。

2 规范性引用文件

本文件没有规范性引用文件。

3 术语和定义

下列术语和定义适用于本文件。

3.1

大语言模型 large language model

经过预训练和微调的大规模人工智能模型，可以理解指令并基于大量数据生成人类语言。

[来源：WDTA AI-STR-02，3.2]

3.2

文献资源知识图谱 knowledge graph of literature resources

以一种结构化的形式描述文献资源领域中概念、实体及其关系的方式。

注：文献资源知识图谱将文献资源的海量信息表达成更接近人类认知世界的形式，提供了一种更好地组织、管理和理解文献资源海量信息的能力。

3.3

本体 ontology

表示实体类型以及实体类型之间关系、实体类型属性类型及其之间关联的一种模型。

注：又称本体模型

[来源：GB/T 42131-2022，3.8]

3.4

实体 entity

独立存在的对象。

[来源：GB/T 42131-2022，3.2]

3.5

关系 relation

实体、实体类型、实体组合或实体类型组合间的联系。

注：关系用于描述实体类型和实体类型、实体类型和实体、实体和实体之间的关联方式。

[来源：GB/T 42131-2022，3.11]

3.6

实体识别 entity identification

一种信息提取技术。从文本数据中获取人名、地名等实体数据。

[来源：《计算机科学技术名词（第三版）》，07.0419]

3.7

实体链接 entity linking

将文本中的实体链向其在给定知识库中目标实体的过程。

[来源：《知识图谱：方法、实践与应用》，4.5.1]

3.8

关系抽取 relation extraction

识别文本中提到的实体之间关系的任务。

[来源：GB/T 41867-2022，3.3.4]

4 缩略语

下列缩略语适用于本文件。

API：应用程序编程接口（Application Programming Interface）

CSV：逗号分隔值（Comma-Separated Values）

JSON：轻量级的数据交换格式（JavaScript Object Notation，JS对象简谱）

RDF：资源描述框架（Resource Description Framework）

RESTful：基于REST表述性状态转移架构风格的Web服务设计方法（Representational State Transfer）

SQL：结构化查询语言(Structured Query Language)

XML：可扩展标记语言 (Extensible Markup Language，XML)

5 架构与流程

5.1 构建文献资源知识图谱技术架构

构建文献资源知识图谱技术构架见图 1。

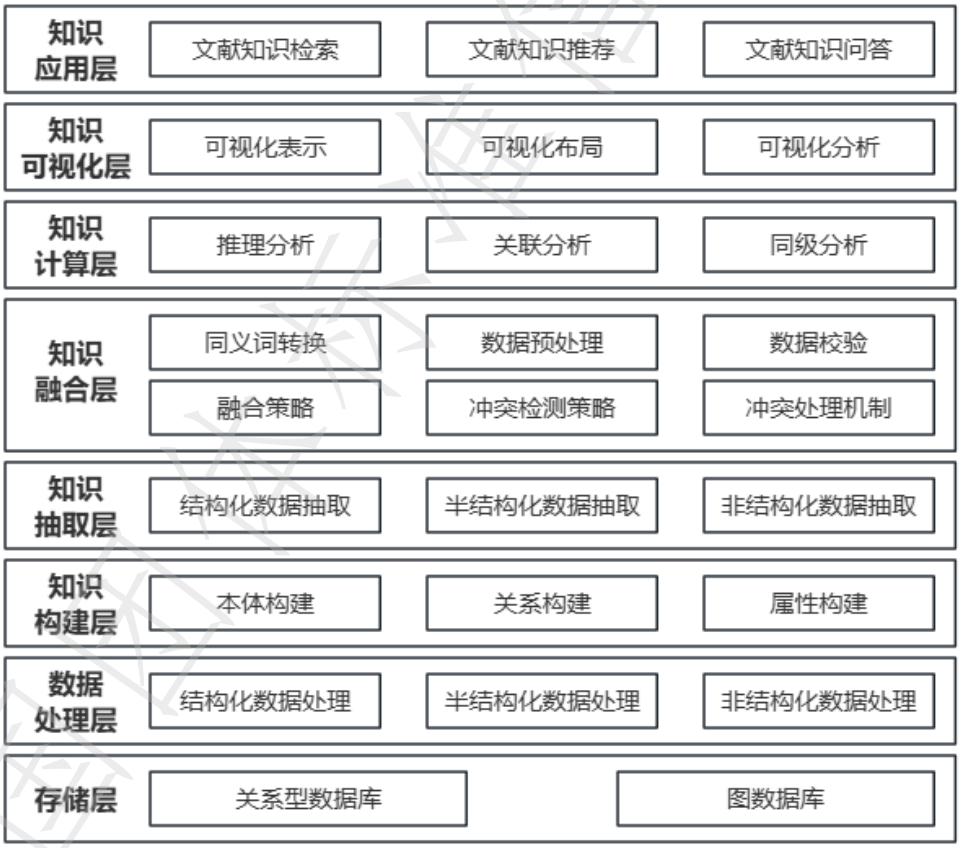

图1 构建文献资源知识图谱技术架构

图1中：

a) 存储层

提供分布式存储管理，为文献资源数据存储提供高可用的存储支撑，主要包括关系型数据库、图数据库等；

b) 数据处理层

支持对结构化数据、半结构化数据和非结构化数据的接入与清洗；

c) 知识构建层

基于统一的数据接入，通过可视化知识模型构建的方式，提供配置化的模式实现知识本体构建、关系构建、属性构建。依据文献资源的内容交叉性，构建具有领域特色的本体模型，确保知识图谱的专业性和准确性；

d) 知识抽取层
实现从海量异构文献资源数据中抽取知识，基于统一的可视化抽取任务管理页面，提供结构化数据抽取、半结构化数据抽取、非结构化数据抽取，支持实体、关系、属性等知识抽取；

e) 知识融合层
提供知识本体融合、知识更新、实体链接等功能，为文献知识融合提供工具支撑；

f) 知识计算层
集成通用的图挖掘分析算法库，为各类图分析应用提供基础算法支撑。同时提供知识推理分析、关联分析、同级分析等知识计算功能；

g) 知识可视化层
基于统一的 2D/3D 知识可视化展示框架，提供知识图谱可视化表示、知识可视化布局以及知识图谱可视化分析等功能。同时提供交互式界面，能够在知识图谱中进行探索式检索，以发现新知识；

h) 知识应用层
提供统一的 RESTFul 接口，提供基于文献知识的知识检索、知识问答、知识推荐等知识服务。

5.2 文献资源知识图谱构建流程

文献资源知识图谱构建流程见图2。

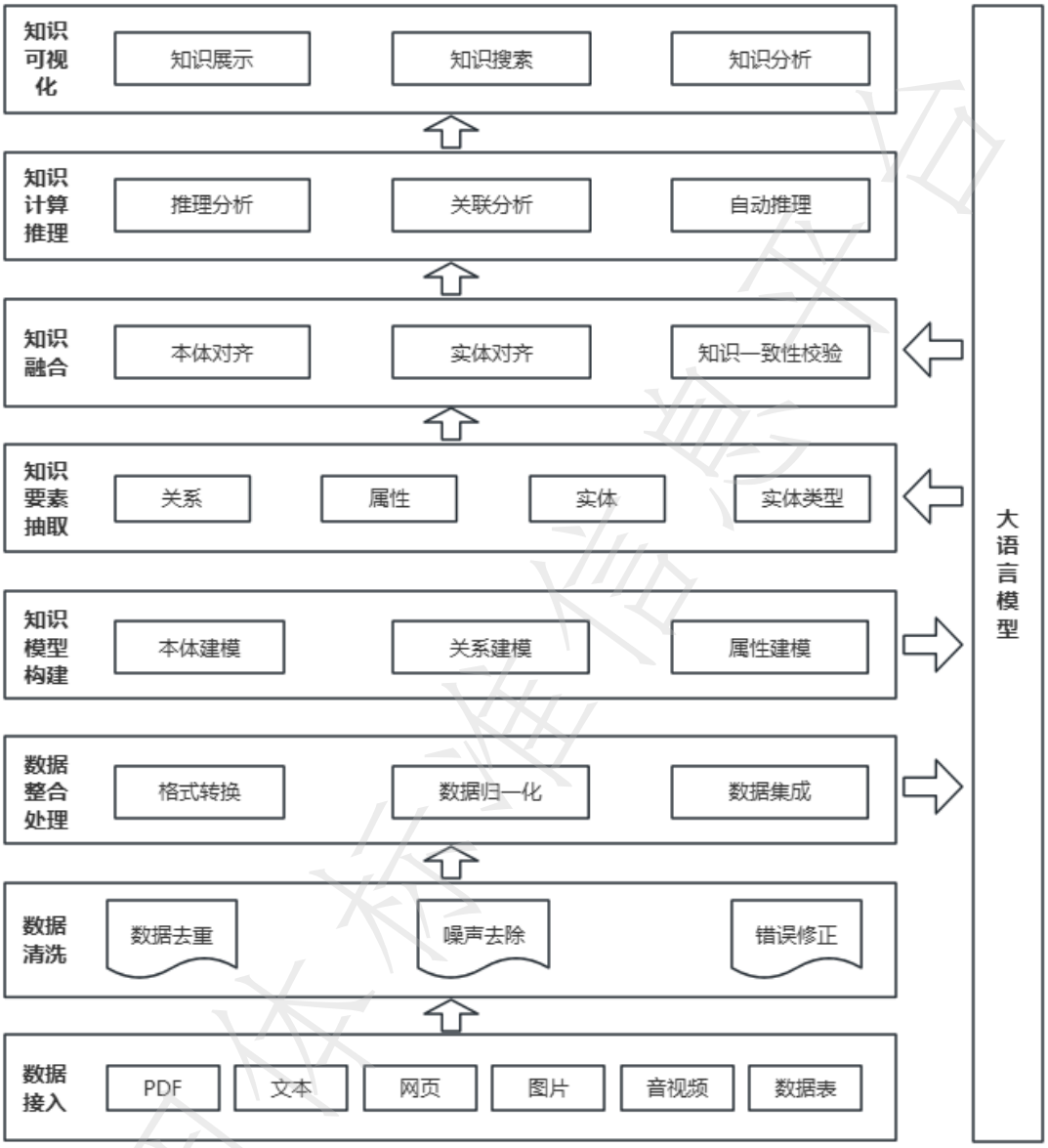

图2　文献资源知识图谱构建流程

5.2.1　数据接入

数据接入将明确数据源的选择，确定合适的数据接入方式，以确保数据的完整性和准确性。

5.2.2　数据清洗

数据清洗通过对接入的数据进行预处理，进一步提高数据的质量和一致性。

5.2.3　数据整合处理

数据整合处理通过数据格式转换，将不同来源的数据统一格式，而后通过数据归一化确保数据中的实体和属性具有统一的表示，消除歧义。最后，数据集成将处理后的数据融合为一个整体。

5.2.4　知识模型构建

知识模型构建是指建立知识图谱的概念模型，即采用什么样的方式来表达知识，构建一个概念模型对知识进行描述。在模型中需要构建本体、属性以及本体之间的关系。

5.2.5　知识要素抽取

知识要素抽取依据文献资源中的信息，借助自然语言处理技术进行实体的识别以及关系和属性的抽取。

5.2.6 知识融合

知识融合是将新抽取的知识与已有知识进行融合，消除新抽取知识与已有知识的歧义。

5.2.7 知识计算推理

知识计算推理在已有的知识库基础上进一步挖掘隐含的知识，从而丰富、扩展知识库。

5.2.8 知识可视化

知识可视化是将知识转化为一种人类的视觉形式，知识可视化包括知识可视化表示、知识可视化布局、知识可视化分析等功能。通过知识可视化，用户可直观的对数据进行全局感知。

6 技术要求

6.1 数据接入与清洗

数据接入与清洗是在选择明确的文献资源文本数据源基础上，通过数据清洗和其他预处理方法，消除数据中的噪声和异常值，进一步提高数据的质量和一致性。

a) 应优先选择权威性强、质量高的数据源，如知名出版社、学术机构等，以提高知识图谱的可信度；
b) 为考虑数据的时效性，应优先选择提供实时或定期更新的数据源；
c) 应确保数据源提供的 API 或数据导出格式与知识图谱构建系统兼容；
d) 应确保数据传输过程中遵循相关的安全标准和协议；
e) 应确保选择的数据源符合法律法规和隐私政策要求，避免侵犯知识产权和个人隐私；
f) 应支持结构化数据、半结构化数据、非结构化数据接入；
g) 应对文献资源进行文本清洗，去除无关信息；
h) 可实现数据源的自动接入和整合；
i) 应支持数据表、SQL 语句、文件等接入方式；
j) 对于大型数据集，应提供批量导入功能，支持多种数据格式（如 CSV、JSON、XML 等）的导入；
k) 可对文本进行标准化处理，如统一大小写、去除标点符号等；
l) 可利用高效的去重算法和技术，去除数据中的重复项，确保知识的唯一性和准确性。

6.2 数据整合处理

数据整合处理包括数据格式转换、数据归一化和数据集成等。首先将不同来源的数据统一为构建知识图谱所需的格式。其次，数据归一化确保数据中的实体和属性具有统一的表示，消除歧义。最后，数据集成将处理后的数据融合为一个整体，为后续的知识图谱构建和分析提供坚实的数据基础。

a) 应处理涉及信息安全保护的数据；
b) 应将所有数据源的数据转换为统一的、标准化的格式，如 JSON、XML 或 RDF，以便于后续的整合和分析；
c) 应将来自不同数据源的数据进行融合，消除数据之间的矛盾和冲突，形成一致的知识表示；
d) 应检查数据的完整性和全面性，确保关键信息没有遗漏；
e) 应利用领域知识和规则，验证数据的准确性，确保知识的可靠性；
f) 可将各种非结构化或半结构化数据（如 PDF、Word 文档、OFD、网页等），解析为结构化数据；
g) 可利用实体链接和实体消歧技术，将不同数据源中的同名实体对齐到知识图谱中的同一实体上，确保实体的唯一性和一致性；
h) 可对实体的属性进行归一化处理，如将日期、数字、单位等转换为统一的格式和标准，以便于后续的查询和分析。

6.3 知识模型构建

知识模型构建可建立文献资源知识图谱的概念模型，即采用什么样的方式来表达知识，构建一个概念模型对知识进行描述。知识模型构建的过程是知识图谱构建的基础，高质量的知识模型能避免许多不必要、重复性的知识获取工作，有效提高知识图谱构建的效率。

a) 应支持以可视化、拖拽等方式构建知识模型；
b) 应按照领域建立本体，以更准确地描述该领域的概念、实体及其相互关系；
c) 应定义一组跨领域通用的核心本体，作为不同领域本体的基础，以支持跨领域应用；
d) 应对常用关系进行约定，确保不同知识图谱在描述相似概念时使用一致的词汇，减少歧义；
e) 应清晰、明确地定义知识图谱中的实体和关系，确保每个概念和实体都有唯一的定义和解释；
f) 应保持实体和关系定义的连贯性和一致性，避免歧义和重复；
g) 应在设计实体和关系时考虑未来可能的扩展，确保知识图谱可以随着知识的发展而不断扩展；
h) 应为每个实体和关系定义清晰的属性，包括属性名称、数据类型、取值范围等，确保知识的精确表示；
i) 应明确实体之间的关系类型和层次，包括父子关系、兄弟关系、属性关系等，形成丰富的关系网络；
j) 应支持利用联邦学习技术，以实现跨机构的知识图谱构建与更新；
k) 应支持增量更新、全量更新两种方式的图谱构建；
l) 可导出已构建完成的知识模型；
m) 可从外部导入知识模型。

6.4 知识抽取

知识抽取依赖于文献资源中的信息，借助自然语言处理等技术进行实体识别和关系抽取。实体抽取是从文本中识别并提取出具有实际意义的实体，如人名、地名等。关系抽取则关注于揭示实体之间的关系，如亲属关系、职业关系等，以此丰富和完善图谱内容；

a) 应抽取文献资源中的所有相关关系，确保知识图谱的完整性；
b) 应支持数据字典、抽取规则、抽取模板等多种抽取策略；
c) 应支持通过界面自定义数据字典、抽取规则、抽取模板等配置；
d) 应支持单属性多模型的抽取模型组合策略能力；
e) 抽取策略应易于适应新的文献资源或领域，支持知识的持续更新和扩展；
f) 可支持自动化抽取，减少人工干预，提高知识图谱构建效率；
g) 可针对文献资源领域的特殊性，对模型进行领域适应性训练，以提高实体识别的准确性；
h) 应准确识别文献资源中的实体，包括人名、地名、组织机构名、专业术语等，确保识别的实体与知识图谱中的实体相匹配；
i) 应支持查看知识抽取结果（包括当前及历史）；
j) 应支持对抽取结果进行修改审核；
k) 应支持查看知识抽取结果报表（包括当前及历史）及数据详情。

6.5 知识融合

知识的产生是一个不断更新、不断完善、动态产生的过程，知识抽取后需要将抽取的知识与已有知识进行融合。知识融合是通过对相关知识对齐、关联、合并使其成为一个有机的整体，是一种提供更全面知识共享的重要方法。

a) 应支持从知识实体、本体、属性、关系 4 个层次进行融合；
b) 应支持采用分布式算法，以实现不同来源知识图谱的实体对齐和融合；
c) 可支持同义词转换、数据预处理（转换、格式化，比如：大小写转换、日期格式化）、数据校验（过滤、正则等规则，比如：身份证，邮箱，手机校验等）等多种融合预处理策略；
d) 可支持关键词、多属性相似度等多种实体冲突检测策略；
e) 可支持实体链接替换、保留、合并等多种冲突处理机制；
f) 可实现文献资源中识别出的实体与知识图谱中实体的唯一性映射，确保每个实体在知识图谱中都有唯一的标识；
g) 可利用上下文信息和实体链接技术，正确解析实体所指，避免歧义。

6.6 知识计算推理

知识计算推理是在已有的知识库基础上进一步挖掘隐含的知识，从而丰富、扩展知识库；

a) 应确保推理的结果准确无误，确保生成的知识符合事实和定义；
b) 应确保推理过程保持逻辑一致性，避免产生矛盾的知识；
c) 应支持常用算法推演查询，包括中心性算法、社区检测算法、路径寻找算法、相似度算法、图嵌入算法；
d) 应快速处理大量的数据，保证推理过程的高效性；
e) 宜支持添加、修改和删除推理规则；
f) 可利用定义明确的推理规则，从文献资源中抽取的信息中推导出新的关系；
g) 可利用图模型进行推理，通过图结构中的模式发现新的关系；
h) 对于推理过程中可能存在的不确定性，应使用概率模型或置信度评估来处理；
i) 应防止因过度推理导致知识图谱中出现错误或不准确的信息；
j) 可支持自动化的推理过程，减少人工干预。

6.7 知识可视化

知识可视化是将知识转化为一种人类的视觉形式，直观、形象地表现、解释、分析、模拟、发现或揭示隐藏在知识内部的特征和规律。知识可视化包括知识可视化表示、知识可视化布局、知识可视化分析等功能。通过知识可视化，根据业务需求设计合适的数据展示布局和交互形式，用户可直观的对数据进行全局感知，也能够了解数据结构背后的数据关系，对结果进行追根溯源的分析。

a) 应使用标准化图元（如圆形、方形、箭头等）来表示不同的元素，以减少认知负担；
b) 应合理使用颜色来区分不同实体、关系和属性，同时确保颜色的对比度和可访问性；
c) 宜采用合理的布局减少认知复杂度，例如使用力引导布局来优化节点之间的关系；
d) 应提供放大、缩小、移动、搜索、过滤等交互功能；
e) 应在有限的视觉空间内展示尽可能多的相关信息，同时避免过载；
f) 应确保可视化中呈现的数据与知识图谱中的数据精确对应，不丢失信息；
g) 可提供有效的导航机制，使用户能够轻松地在图中定位和跳转；
h) 可允许高级用户根据需要定制可视化的某些方面，如颜色方案、图元样式等；
i) 应支持处理大规模的知识图谱数据，保持良好的性能和可扩展性。

6.8 质量评估和维护

知识图谱质量评估与维护是确保知识准确性和时效性的关键环节。可全面检查数据的准确性、一致性、完整性和时效性，确保图谱信息真实可靠。定期更新数据，实施严格的版本控制，持续优化图谱质量。

a) 应检查关键实体和关系的完备性，确保没有遗漏重要内容；
b) 应确保图谱中的实体、关系、属性等定义统一，无歧义；
c) 宜定期检查图谱中的信息是否过时，及时更新和维护；
d) 可设立定期的数据更新机制，确保图谱内容的实时性和准确性；
e) 可提供历史版本查询功能，便于追踪和恢复；
f) 应制定严格的数据访问和修改权限控制，确保图谱的安全性；
g) 可整合新的文献资源，对图谱进行增量更新；
h) 可对图谱执行版本控制管理，记录图谱的变更和更新。

6.9 知识抽取模型训练

首先选用合适的通用大语言模型，通过增量预训练注入领域知识，再训练其抽取实体和关系的能力，形成文献资源大语言模型。使用验证集和测试集评估模型性能，并根据评估结果调整模型参数或训练策略，以提高模型质量。

a) 应针对具有代表性、多样性和高质量的数据进行抽取，涵盖丰富的话题和语言风格。
b) 宜具备较强的并行计算能力，以支持大规模数据的训练；
c) 宜具备较好的过拟合控制能力，以保证模型在未知数据上的泛化能力；

d） 可支持模型剪枝和量化，以降低模型复杂度和计算资源消耗；
e） 可支持多语言训练，以满足不同国家和地区用户的需求；
f） 宜具备较好的可解释性，以便用户了解模型的工作原理和决策依据。
g） 应确保模型无偏见、符合道德标准，并遵守相关法律法规。

6.10 大语言模型赋能知识图谱

利用大语言模型在语义理解、内容生成等方面的技术优势，实现大语言模型对知识图谱构建至应用全生命周期各环节的增强，提升效率和质量。

a） 可支持知识图谱大语言模型针对知识图谱提供检索增强生成等能力；
b） 可具备知识图谱的自动构建能力，从非结构化数据中抽取实体、关系和属性等信息；
c） 可支持知识图谱的动态更新，以实时反映现实世界的变化；
d） 可支持知识模型的构建，以得到更准确和全面的知识模型；
e） 可具备知识图谱的查询和分析能力，为用户提供高效的知识检索服务；
f） 可支持知识图谱的分布式存储和计算，以满足大规模知识图谱的需求；
g） 可支持知识图谱的语义理解能力，为用户提供智能问答、推荐等服务；
h） 可支持知识图谱多模态知识对齐，以实现不同模态知识的对齐和整合。
i） 应遵守相关法律、法规、规章、标准和伦理，确保知识图谱的合法合规性。

参 考 文 献

[1] GB/T 5271.14-2008 信息技术 词汇 第14部分:可靠性、可维护性与可用性
[2] GB/T 35273-2020 信息安全技术 个人信息安全规范
[3] GB/T 41867-2022 信息技术 人工智能 术语
[4] GB/T 42131-2022 人工智能 知识图谱技术框架
[5] ISO/IEC 19510:2013 Information technology -- Open Distributed Processing -- Unified Modeling Language (UML) profile for RDF and OWL
[6] ISO/IEC 19763-10:2023 Information technology -- Metamodel framework for interoperability (MFI)
[7] ISO/IEC 20000-1:2018 Information technology -- Service management -- Part 1: Service management system requirements
[8] YD/T 4044-2022 基于人工智能的知识图谱构建技术要求
[9] T/HNIT 2-2021 领域知识图谱构建技术规程
[10]《计算机科学技术名词（第三版）》
[11] WDTA AI-STR-02《大语言模型安全测试方法》
[12]《知识图谱：方法、实践与应用》