

面向人工智能的数据治理实践指南 (1.0)

王妙琼

中国信息通信研究院云计算与大数据研究所

2024年5月

数据治理需求与方法论不断演进



第一阶段-DBMS 01

数据治理主要依赖于数据库管理系统

- 1988 年，麻省理工启动全面数据质量管理计划（TDQM），DAMA（国际数据管理组织协会）同年成立。
- 2002 年，学术论文《数据仓库治理》中开始出现“数据治理”的实践内容。
- DGI、DAMABOK等数据治理框架开始发展。



第二阶段-EDW 02

企业级数仓及大数据平台建设加速数据治理发展

- 2017年，中国信通院《数据资产管理实践白皮书（1.0）》发布。
- 2018 年，国家标准《数据管理能力成熟度评估模型》发布。
- 2023年，《数据资产管理实践白皮书》更新至6.0，《DataOps实践白皮书（1.0）》发布



第三阶段-AGI 03

通用人工智能的快速发展为数据治理带来新的挑战

- 2020年，GPT3，生成式人工智能的成功席卷全球，对整个社会的生产模式产生影响。
- 2021年，吴恩达提出人工智能将从“以模型为中心”转向“以数据为中心”，对高质量数据集提出要求。

更高的数据集 质量要求

面向更复杂的数据类型、更多维的评价维度、更冗长的处理环节，需要进一步构建面向人工智能的数据质量评价体系和质量提升工艺。



更显性的偏见与 歧视

训练数据集中不易察觉的偏见和歧视信息，会直接影响到训练结果的输出，随着人工智能应用逐步走向产业化，潜在的道德伦理问题会引起严重的后果。

更复杂的安全与 隐私风险

模型训练和应用过程中存在诸多安全风险，除去传统关注的数据丢失、隐私泄露等风险，还需要额外关注数据投毒、逆向还原等更关注内容治理的新型风险。

面向人工智能的数据治理

(DG4AI, Data Governance for Artificial Intelligence)

指在人工智能应用中管理和控制数据的过程与实践。用以确保数据的质量、可靠性、安全性与合规性，数据能够被准确地用于训练和部署AI模型，同时保护数据的隐私和安全。

数据质量管理

在人工智能应用中确保高质量数据，包括数据的准确性、完整性、一致性和时效性等。

数据安全和隐私保护

保护数据免受未经授权的访问、泄露和滥用，并确保数据隐私得到尊重。

数据合规性管理

数据处理符合政策、法律、法规和行业标准，确保数据合法合规使用。

数据治理流程优化

数据治理的组织结构、决策权责和流程，确保数据在整个周期中得到管理和控制。

数据伦理管理

在人工智能数据治理中应充分考虑伦理和社会责任的重要性。

技术工具和平台建设

包括支持人工智能数据治理的技术工具和平台，例如隐私保护工具、安全分析平台等。

成本治理与优化

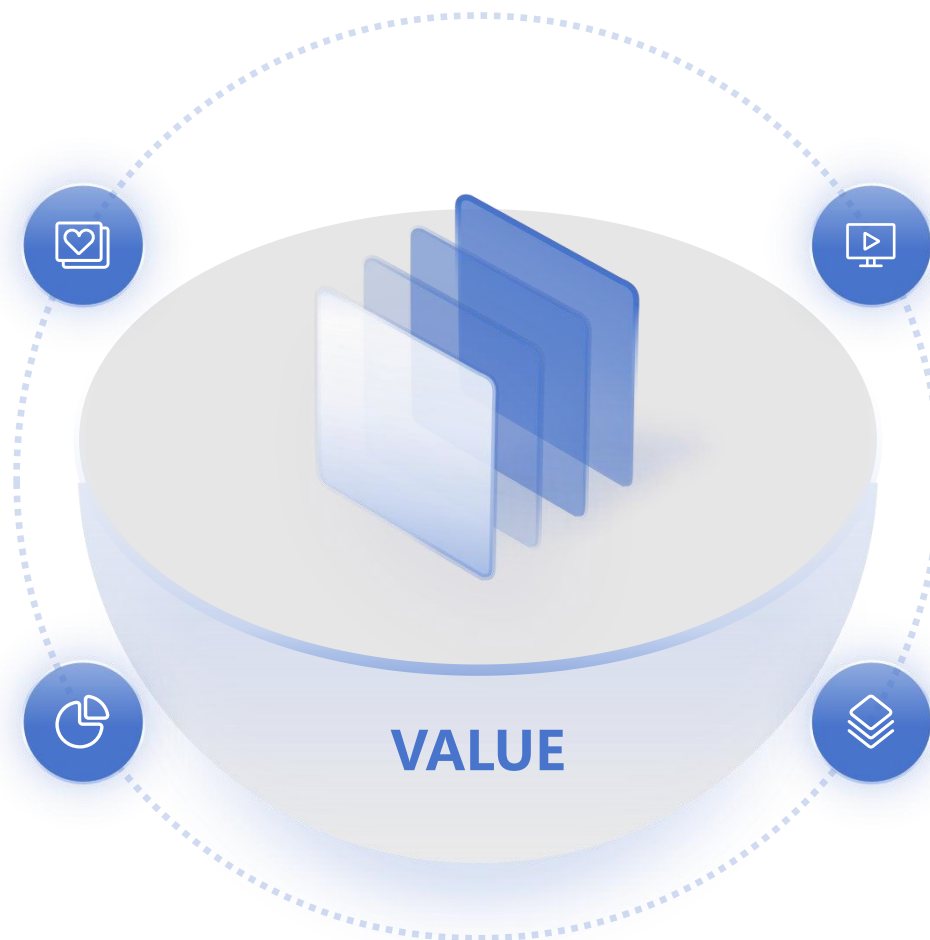
指对人工智能场景下数据的存储、计算、加工等成本的优化与治理。

提升人工智能应用效果与性能

通过提升数据的质量和和安全水平，从而显著提升AI模型的训练效果和预测准确性。

加强人工智能数据安全与隐私保护水平

确保数据的安全和合规性，防止数据泄露和滥用，保护个人隐私和企业机密。



提升人工智能应用的规模化落地效率

缩短人工智能模型的开发构建全流程周期，从而提升人工智能应用的规模化落地效率。

完善数据治理体系

在传统数据治理体系上增强对非结构化数据、数据伦理的关注，是对基础数据治理的升级。

数据治理理论已有多年的发展，针对人工智能领域的的数据治理，在治理范围、成效评价、治理价值上存在差异。

范围

传统的数据治理

结构化数据为主

面向AI的数据治理

+ 非/半结构化数据、多模态数据

更多样

成效评价

由人类评价为主

由算法评价为主

更客观

治理价值

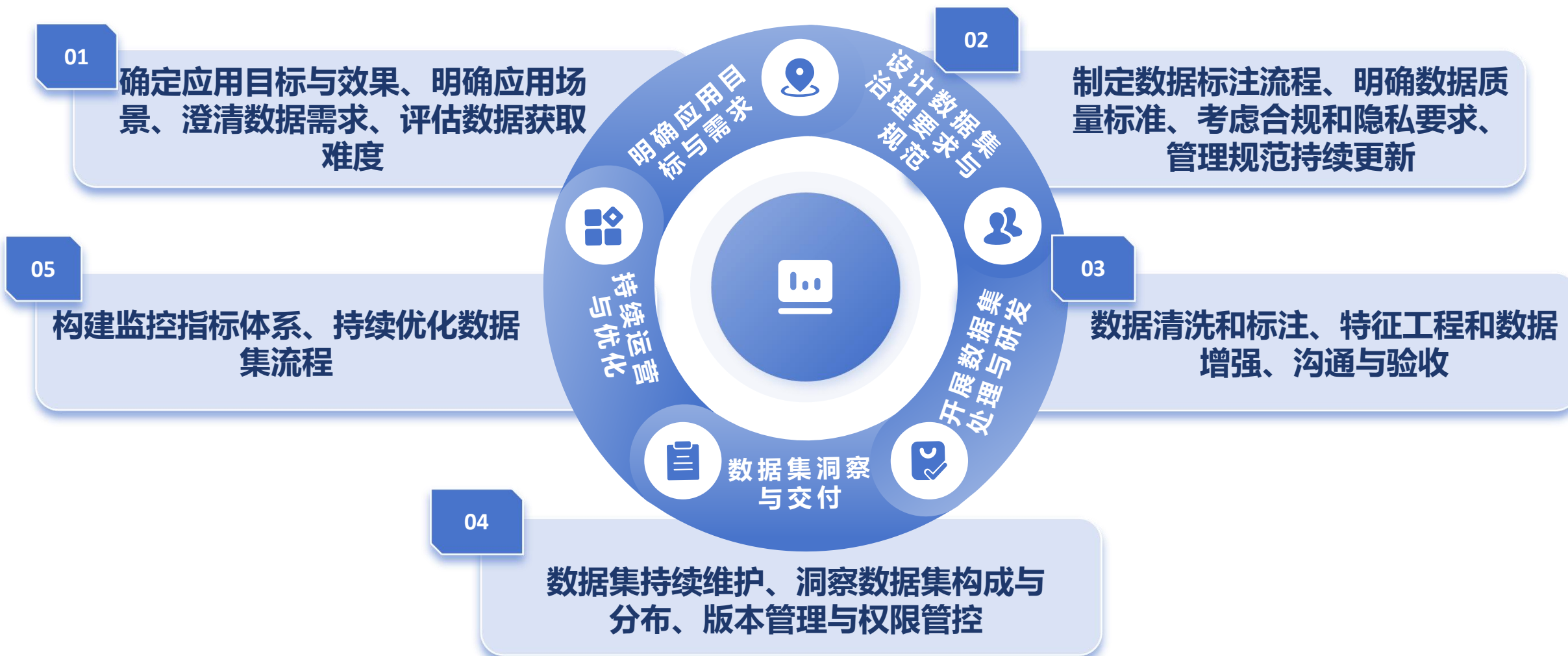
较难度量

直接影响算法应用效果

更显性



面向人工智能的数据治理实践方法





人工智能数据产业分工更加明确

供给人工智能所需的数据集产业，在采集、加工、交易、消费等环节将更加清晰和成熟，通过市场化的调节机制合理分配产业的人才、资金与资源，更高效的推进人工智能应用发展。

数据治理成为大模型的胜负手

随着研发成本的不断攀升和市场集中度的提高（马太效应），只有少数企业可能在这场竞争中胜出。尽管算力和算法方面的技术优势并不明显，但通过有效的数据治理，企业仍有可能获得竞争优势。

DG4AI服务化

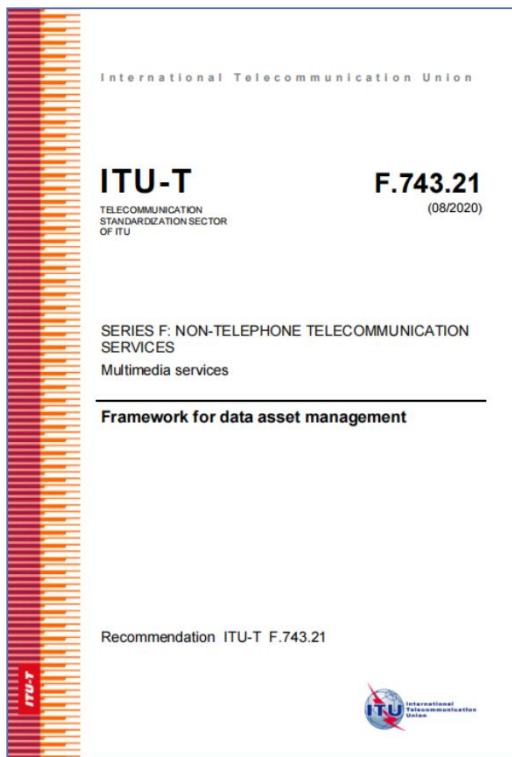
随着DG4AI技术、实践和理论的成熟，DG4AI将更加标准化与流程化，进而服务化，高效、高质量、安全可控的提供标准的数据产品（数据集、语料库）。

本次实践指南的编写得到了通信运营商、金融企业、科研院所、媒体行业、科技公司等专家的大力支持。

(排名不分先后)

华为云	央视频	电信智科	浪潮科技
中兴通讯	亚信科技	奇点云	蓝象科技
中国移动	星环科技	普元	科大讯飞
中国联通	新华社	浦发银行	交通银行
中国电信研究院	新华三	平安银行	华傲科技
中国电信	小鹏汽车	南京大学研究中心	海致科技
中电科大数据院	小米集团	明瞳科技	海南数造
中电金信	腾讯	绿盟科技	广州信安
长城汽车	软通智慧	联通数科	枫清科技
OneSight	人民大学	联通软研院	

面向人工智能的数据治理在理论与实践还处在探索期，目前中国信通院已牵头在中国通信标准化协会、国际电信联盟（ITU）等立项相关行业标准、国际标准，后续还将在技术工具、企业能力建设以及实践方法论等发面持续开展研究、合作与生态建设工作。



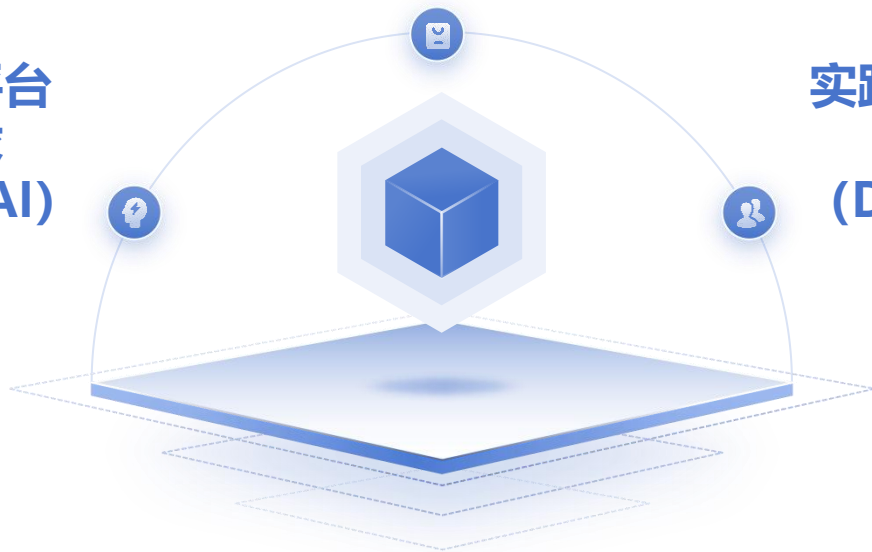
国际标准立项

- **ITU-T F.DADOCM-**
Requirements for data asset development and operations capability in multimedia
(多媒体数据资产开发运营能力需求)
- **ITU-T F.DPLM-REM-**
Requirements and evaluation methods of data provision capability for large models in multimedia applications
(多媒体应用中大模型数据供给能力要求与评估方法)

技术平台
要求
(DG4AI)

企业建设
成熟度
(DG4AI)

实践方法论
更新
(DG4AI)



《面向人工智能的数据治理实践指南（1.0）》将在6月19日于北京举办的“数据智能”大会上正式发布全文。中国信通院将持续就人工智能数据治理（AI4DG & DG4AI）开展研究、标准、生态工作。

—— 研究内容 ——

- 《面向人工智能的数据治理实践指南（1.0）》
- 《智能化数据治理白皮书（1.0）》
-

—— 标准计划 ——

- 《智能化数据治理工具技术要求》
- 《数据治理智能化成熟度模型》
- 《面向通用人工智能的数据研发、治理、运营一体化工具技术要求》（DataOps in AGI）
-

—— 生态培育 ——

沙龙研讨

创新实验室

案例汇编

谢 谢