ICS 35.240.80
L 70

团 体 标 准

T/BSIA 00x-2025

多模态医疗科研数据与共享平台标准

第 3 部分：平台架构

Multimodal medical Research Data and Sharing Platform Standards
- Part 3: Platform Architecture

（征求意见稿）

2025-12-XX 发布　　2025-12-XX 实施

北京软件和信息服务业协会　发布

目　　次

前　言

《多模态医疗科研数据与共享平台标准》预计分为6个部分：

——第1部分：总体要求；

——第2部分：心脑血管疾病科研数据集；

——第3部分：平台架构；

——第4部分：样本数据收集、存储及传递；

——第5部分：知识工程；

——第6部分：安全管理。

本文件是《多模态医疗科研数据与共享平台标准》的第3部分。

本文件依据GB/T1.1-2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》给出的规则起草。

本文件由北京软件和信息服务业协会提出和归口。

本文件由中国人民解放军总医院提出。

本文件起草单位：中国人民解放军总医院、北京大学、浙江大学、中南大学、国家人口健康科学数据中心、北京嘉和美康信息技术有限公司、北京嘉和海森健康科技有限公司、万达信息、北京软件和信息服务业协会。

本文件主要起草人：何昆仑、石金龙、乌日力格、周伟、张敬晨、闫雯、窦飞、陈鹏、龙飞、张磊、仓剑。

本文件为首次发布。

多模态医疗科研数据与共享平台标准

第 3 部分：平台架构

1 范围

本文件规定了多模态医疗科研数据与共享平台建设的系统架构、系统组成、接入规范、安全要求，支撑第 2 部分主题数据集运行，并为第 4-6 部分（治理/图谱/安全）提供技术载体。

本文件适用于医疗科研用户（医学科研教学单位、医院）、医疗行业软件企业和医疗数据处理服务企业或机构在相关业务规范中作为参照。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 34960.5-2018《信息技术服务 治理 第 5 部分：数据治理规范》

GB/T 42131-2022《人工智能 知识图谱技术框架》

3 术语和定义

GB/T 5271中包含的术语适用于本文件，并沿用第1部分术语，本文件中仅列出未包含的术语。

3.1 文档型数据库

一种非结构化数据管理系统，其核心特征是以文档为基本单位进行数据的存储、管理和查询。文档通常采用 JSON、BSON 或 XML 等格式，允许每个文档拥有独立、可变的结构（模式）。

3.2 分布式全文检索引擎

一种能够在分布式集群环境中，对文本数据建立倒排索引并提供近实时的、高性能的全文检服务的技术平台。其核心功能包括海量数据的分布式存储、索引、检索和分析。

3.3 数据治理

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。

[来源：GB/T 34960.5-2018]

3.4 知识图谱

以结构化形式描述的知识元素及其联系的集合。

[来源：GB/T 42131-2022]

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）

JSON JavaScript 对象表示法（JavaScript Object Notation）

RESTful API 表述性状态转移应用程序编程接口（Representational State Transfer Application Programming Interface）

PDF 便携式文档格式（Portable Document Format）

OCR 光学字符识别（Optical Character Recognition）

OAuth 2.0 开放授权 2.0（Open Authorization 2.0）

XML 可扩展标记语言（eXtensible Markup Language）

SFTP SSH 文件传输协议（SSH File Transfer Protocol）

CSV 逗号分隔值（Comma-Separated Values）

5 系统架构

5.1 逻辑架构

本平台采用分层模块化的逻辑架构设计，以实现高内聚、低耦合的系统组织方式，确保系统的可扩展性、可维护性与安全性。整体逻辑架构分为数据层、服务层与应用层三层结构。

数据层作为系统基础，负责多模态科研数据的存储与管理，包括结构化、半结构化和非结构化数据。该层采用混合存储策略，集成文档型数据库与分布式全文检索引擎，分别实现非结构化数据的灵活存储与高效检索功能。

服务层是系统的核心处理单元，提供数据处理、知识图谱构建与管理、自然语言处理、多模态语义对齐等核心服务。该层通过微服务架构实现功能模块化，各服务可独立部署、扩展和升级，包括数据解析引擎、自然语言分词与结构化引擎、知识获取与融合引擎、数据分析引擎等。

应用层面向最终用户，提供直观易用的操作界面，包括系统登录、数据检索、项目管理、数据分析、个人中心等功能模块。该层通过统一的API网关与服务层交互，确保用户操作的响应速度与系统安全性。

各层之间通过标准化接口进行通信，采用RESTful API与消息队列等多种方式，保证数据传输的可靠性、实时性与安全性。系统还设有统一的安全认证与权限管理机制，贯穿所有层次，确保数据与操作的安全可控。

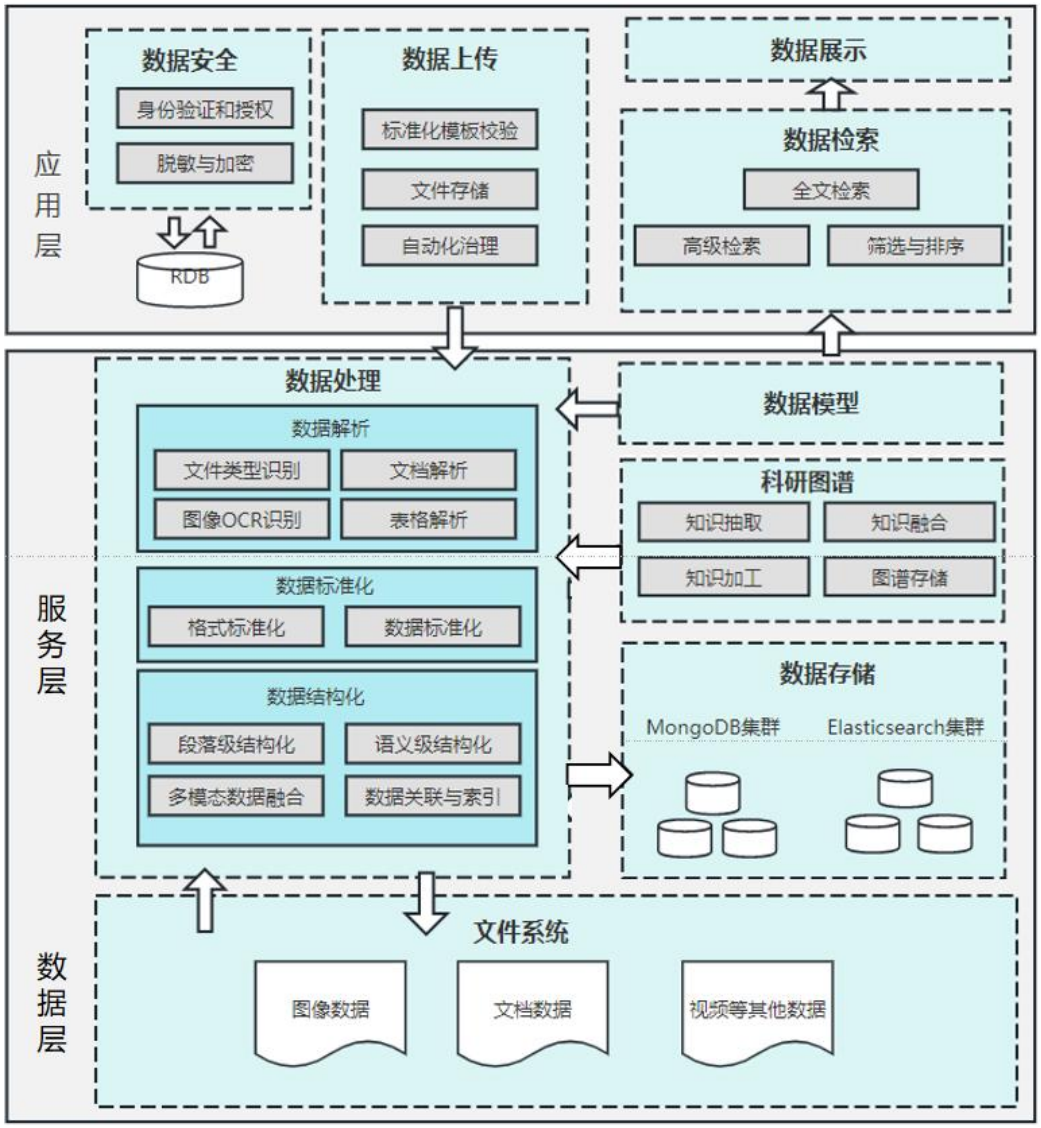

图 1 总体架构设计图

5.2 技术架构要求

系统采用混合型数据存储架构，以适应多模态医疗科研多模态数据的多样化存储与高效检索需求。

文档型数据库存储用于管理非结构化与半结构化数据。文档型数据库基于JSON的非关系型数据模型，使用BSON（Binary JSON）格式存储文档，适用于存储文献全文、图像、表格、视频、音频等异构科研数据。其文档模型支持动态字段扩展，便于直接映射多模态数据解析后的结构化结果。通过数据规范化与合理的文档结构设计，实现数据的高效存储与查询。系统支持自动索引优化，根据查询模式自动创建索引，提升查询性能。

分布式全文检索引擎专门用于实现高效、实时的数据检索功能。其倒排索引技术支持全文检索、结构化检索与多条件组合检索，满足科研用户对疾病名称、研究项目、关键指标等内容的快速检索需求。分布式架构支持水平扩展，通过分片与副本机制保障高可用性与数据可靠性。分布式全文检索引擎支持多租户管理，可实现不同研究项目或用户组的数据隔离与安全访问。

此外，系统通过数据输入与连接管理模块，实现多模态数据的安全接入与存储。数据输入阶段，系统接收经过解析与标准化处理的结构化数据，将其转换为JSON等适用格式。数据库连接采用加密传输与身份验证机制，确保数据安全。在存储过程中，系统执行数据校验与异常处理，保障数据完整性与一致性。

6 系统组成

6.1 总体功能

多模态医疗科研数据与共享平台集数据管理、知识发现、科研协作与安全管理于一体，通过数据治理融合，建立以“项目-课题-文件”为主线的多模态科研数据库。总体功能涵盖多模态数据治理、科研知识图谱构建、系统操作功能及系统管理四个方面。平台支持心脑血管疾病科研数据的全生命周期管理，从数据采集、存储、处理、分析到共享应用，为科研人员提供一站式数据支持与决策服务。通过模块化设计与分层架构，平台具备高度的可扩展性与灵活性，能够适应不同规模与研究方向的科研需求。

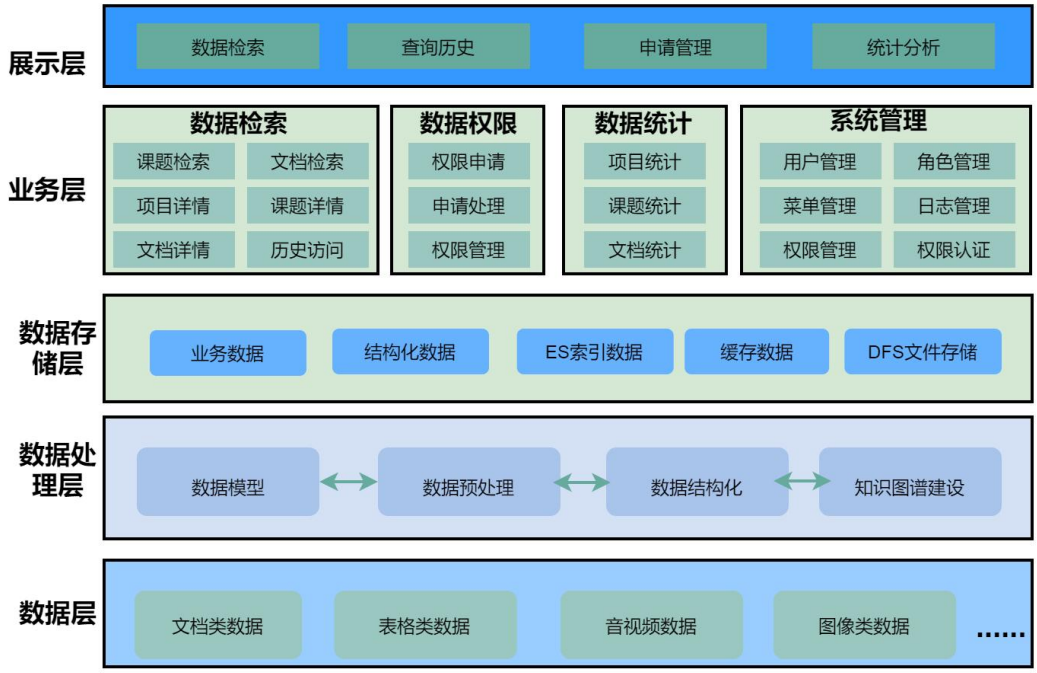

图 2 软件整体功能架构图

多模态医疗科研数据平台系统功能架构由五层构成。整体上，该架构通过各层的协同工作，实现医疗科研数据的存储、处理、管理和展示，为科研人员提供全方位的共享科研平台支持。

6.2 展示层

采用负责前端界面展示，为用户提供直观的数据可视化界面和交互操作入口。

6.3 业务层

分为数据检索、数据权限、数据统计和系统管理四大模块。数据检索包括课题检索、文档检索等功能；数据权限涉及权限申请、处理和管理等；数据统计涵盖项目、课题和文档统

计；系统管理包括用户、角色、菜单、日志管理以及权限认证等，实现对平台各项业务功能的管理与操作。

6.3.1 系统操作

系统操作功能面向科研用户，提供便捷、安全的操作界面与工具：

登录与身份验证：提供用户登录与身份验证功能，确保授权用户访问系统，保障系统与数据安全。

概览与统计分析：对文档、课题、项目进行统计分析，以图表形式展示文档数量、课题分布、项目进度等信息，支持科研决策与管理。

文档检索：支持疾病名称、研究项目等关键词搜索，以及多条件高级检索，帮助用户快速定位相关文档。

关键字与高级检索：支持全文检索与字段检索，满足用户精准、快速的检索需求。

数据导出与导入：支持原始文档导出为PDF、Word、Excel等格式，同时允许用户按规范格式导入外部科研数据，实现数据快速录入与整合。

项目管理：支持科研项目的创建与全生命周期管理，确保项目按计划执行。

数据分析：提供描述性分析、交叉表、相关性分析等功能，实时保存观测指标数据，支持数据质量检查与关联分析。

数据申请与审批：提供在线数据申请与审批功能，确保数据合法使用与安全共享。

历史追溯：支持历史查看项目、申请项目与搜索条件的完整追溯，提高科研工作效率。

个人中心：允许用户查看与编辑个人信息、管理数据申请记录与修改密码等。

退出登录：提供安全退出机制，清除用户会话信息，保障系统安全。

6.3.2 系统管理

系统管理功能面向平台管理员，实现系统配置、用户管理与安全监控：

用户管理：提供用户新增、修改、删除等操作，支持用户信息与权限的集中管理。

角色管理：支持内置角色与功能权限设置，可根据用户职责分配不同级别的数据访问与使用权限。

菜单管理：根据不同角色动态配置菜单显示，确保用户仅访问权限范围内的功能模块。

日志管理：支持系统日志与用户操作日志的查询与分析，用于安全审计、故障排查与系统优化，保障系统合规性与安全性。

6.4 数据存储层

根据数据形态采用相应的存储架构，如MySQL存储业务数据，MongoDB 存储结构化数据，ES 存储索引数据，Redis用于缓存数据，DFS实现文件存储，满足不同类型数据的存储需求，保障数据的高效存储与读取。

6.5 数据处理层

包括数据模型、数据预处理、数据结构化和知识图谱建设，对原始数据进行加工处理和分析，挖掘数据价值，为上层业务提供数据支持。

6.5.1 数据处理模块

数据处理功能模块致力于实现多模态科研数据的标准化、结构化与可用化，具体包括以下核心功能：

多模态数据处理：建立统一数据模型标准，整合结构化、半结构化和非结构化数据，实现数据的标准化管理与存储。例如，将科研文献中的文本、图像、表格等数据按统一标准整合存储，支持综合分析与利用。

多模态数据解析引擎：采用分层模块化架构，实现PDF、OCR等多模态数据的自动识别与解析，准确提取文本、图像、表格中的关键信息，并转换为可处理的数据格式，为后续分析提供支持。

自然语言分词及结构化：利用结构化引擎实现文本段落及语义级结构化处理，包括分词、词性标注、命名实体识别等，将文本信息转化为结构化数据，支持文本分析、信息检索与知识挖掘。

多模态语义对齐：通过格式转换与编码标准化，采用动态时间规整、关键点检测与空间对齐、跨模态对比学习等技术，将不同模态数据映射到统一语义空间，实现多模态数据的语义对齐，提升数据关联性与一致性。

6.5.2 科研知识图谱

科研知识图谱模块实现科研知识的获取、融合、加工与存储，支持深层次知识发现与推理：

知识获取：通过多模态数据智能采集与预处理，对科研报告、文献等文本进行清洗、分词、词性标注，去除噪声数据，提取关键信息与特征，为知识图谱构建提供高质量数据源。

知识融合：整合多源数据，通过数据关联、实体对齐、关系融合等技术，消除冗余与异构性，建立统一的知识表示，形成完整、准确的科研知识体系。

知识加工：构建科研数据中实验、材料等实体间的关联，挖掘潜在知识与规律，利用推理算法与机器学习技术优化知识图谱结构与语义信息，支持深度知识发现与决策。

图谱存储：采用高性能数据库存储实体、关系与属性数据，确保大规模知识数据的高效存储与访问。建立完善的安全体系，实现数据访问控制、权限管理与加密，保障数据安全与隐私。

6.6 数据层

是整个架构的基础，包含文档类、表格类、图像类等多种数据，为平台提供丰富的数据资源。

7 接入规范

7.1 平台节点

本系统采用分布式架构设计，由以下核心节点组成，各节点通过标准化接口实现互联互通：

7.1.1 数据申请与审批节点

功能：负责科研数据使用申请的提交、审核与授权管理

接入地址：http://{domain}/data/apply/v1/apply/（如：新增申请接口/add，申请处理接口/examine）

技术特征：采用RESTful API设计，支持JSON格式数据交换

7.1.2 数据检索与服务节点

功能：提供多维度数据检索服务，包括文档搜索、课题查询等

接入地址：http://{domain}/data/search/v1/（如：文档搜索接口/docAll/docSearch，课题查询接口/topic/topicSearch）

技术特征：支持关键字检索、高级检索和多条件组合查询

7.1.3 文件传输节点

功能：负责大数据文件的分片上传、下载传输管理

接入地址：http://{domain}/shardUpload/（如：初始化接口/init，分片上传接口/uploadPart）

技术特征：支持大文件分片传输、断点续传和MD5校验

7.1.4 数据管理节点

功能：提供统计分析和系统管理功能

接入地址：http://{domain}/search/（如：统计接口/docStaticCount，历史查询接口/getHistorySearchHistory）

技术特征：支持多维数据聚合分析和历史操作追溯

7.2 数据交换模式和途径

7.2.1 数据交换模式

平台支持以下两种数据交换模式：

a) API实时交互模式

适用场景：适用于结构化数据的实时查询和交易型操作

技术实现：采用HTTPS协议保障传输安全，JSON格式数据交换

b) 文件异步传输模式

适用场景：适用于非结构化大数据文件的传输

技术实现：采用分片传输机制，支持断点续传

7.2.2 数据交换途径

a) 直接接入方式

适用对象：平台直接用户（科研人员、管理员）

数据格式：遵循平台统一数据元标准（见第2部分）

b) 系统对接方式

适用对象：第三方系统接入

数据格式：支持JSON、XML等多种数据格式

c) 批量文件接入方式

适用对象：大数据量批量交换

数据格式：支持CSV、JSON、XML等格式

7.3 共享规范

7.3.1 数据申请与审批规范

a) 申请流程规范

申请提交：通过/apply/add接口提交申请，必须包含申请类型、描述、项目ID等必填字段

审批流程：采用分级审批机制，支持在线审批和结果反馈
状态管理：申请状态包括01-待处理、02-已通过、03-已拒绝等
b） 权限控制规范
访问权限：基于角色访问控制（RBAC），分数据管理员、科研用户、普通用户等角色
数据分级：按敏感程度分为公开共享、协议共享和受限共享三级
时效控制：数据访问权限设定期限，到期自动失效

7.3.2 数据访问控制规范

a） 认证授权规范
身份认证：采用多因素认证机制，支持密码、令牌等多种方式
权限验证：每次数据访问前进行权限验证，确保数据安全
会话管理：支持会话超时自动退出和安全日志记录
b） 安全审计规范
操作日志：记录所有数据访问和操作行为，包括查询、下载等
审计追踪：支持操作行为的全程追踪和事后审计
异常监测：实时监测异常访问行为并发出警报

7.3.3 数据质量保障规范

a） 数据质量标准
完整性要求：数据元完整度≥95%
准确性要求：数据准确率≥99%
时效性要求：数据更新延时≤24小时
b） 质量监控机制
质量检查：建立数据质量检查规则和自动检查机制
问题处理：设立数据质量问题反馈和处理流程
持续改进：定期评估数据质量并实施改进措施

8 安全设计

8.1 身份验证

8.1.1 认证体系架构

平台采用多层次身份认证体系，确保用户身份的真实性和可信度：
a） 第一层：基础凭证认证：要求用户提供用户名和符合强度要求的密码进行身份验证
b） 第二层：多因素认证：对敏感操作和管理功能启用多因素认证机制
c） 第三层：会话管理：采用令牌机制管理用户会话，支持会话超时自动退出

8.1.2 密码策略要求

密码策略应符合以下安全要求：
a） 复杂度要求：密码长度不少于12位，必须包含大写字母、小写字母、数字和特殊字符
b） 有效期：密码最长有效期为90天，到期强制更换
c） 历史记忆：系统记忆最近5次使用过的密码，禁止重复使用
d） 失败锁定：连续5次认证失败后锁定账户，需管理员解锁

8.1.3 认证技术实现

a) 采用国密算法或AES-256加密存储用户凭证
b) 传输过程中使用TLS1.2及以上协议加密认证信息
c) 实现防暴力破解机制，自动检测并阻止异常登录行为

8.2 权限控制

8.2.1 访问控制模型

平台采用基于角色的访问控制（RBAC）模型，结合属性基访问控制（ABAC）原则：
a) 角色定义：明确定义数据管理员、科研用户、审计员等系统角色
b) 权限粒度：支持功能级、数据级和字段级多粒度权限控制
c) 动态授权：根据用户属性、环境因素和资源特征动态计算访问权限

8.2.2 最小权限原则

a) 用户权限：每个用户仅授予执行其工作所需的最低权限级别
b) 时间限制：敏感权限设置时间限制，超时自动失效
c) 权限分离：实现关键权限的分离制衡，防止单点权限滥用

8.2.3 数据权限管理

a) 数据分级：根据敏感程度将数据分为公开、内部、敏感、机密四个级别
b) 访问控制：不同级别数据实施差异化的访问控制策略
c) 字段级控制：对敏感字段实现单独的访问权限管理

8.3 审计溯源

8.3.1 审计内容范围

审计系统应记录以下操作事件：
a) 用户行为：用户登录、退出、权限变更等身份相关操作
b) 数据访问：数据的查询、下载、导出等访问操作
c) 数据变更：数据的创建、修改、删除等变更操作，记录字段级修改痕迹
d) 系统事件：系统配置变更、安全策略调整等管理操作

8.3.2 审计日志规范

审计日志记录应符合以下要求：
a) 信息完整性：每条日志记录应包含时间戳、操作主体、操作对象、操作类型、操作结果等关键信息
b) 防篡改机制：采用数字签名等技术确保日志记录的完整性和不可否认性
c) 存储安全：审计日志存储在专用安全区域，与业务数据物理隔离
d) 访问控制：严格限制审计日志的访问权限，仅审计员角色可访问

8.4 备份与恢复

8.4.1 备份策略

平台实施多层次备份策略：

a） 全量备份：每周执行一次全量备份，保留最近4个全量备份版本
b） 增量备份：每天执行多次增量备份，备份间隔不超过4小时
c） 日志备份：实时备份数据库事务日志，支持任意时间点恢复

8.4.2 备份内容范围

备份操作应覆盖以下数据内容：
a） 业务数据：包括结构化数据、非结构化数据和半结构化数据
b） 系统数据：包括用户信息、权限配置、系统参数等
c） 审计日志：包括所有审计和溯源日志记录
d） 应用程序：包括应用程序代码和配置文件

8.4.3 恢复机制

a） 恢复流程：制定详细的数据恢复流程和操作规范
b） 恢复测试：每季度至少进行一次恢复演练，验证备份数据的完整性和可用性
c） 恢复时效：保证核心业务系统在4小时内恢复运行，全部数据在24小时内恢复

8.4.4 连续性保障

a） 故障切换：建立自动故障切换机制，保证系统高可用性
b） 业务连续性：制定业务连续性计划，确保极端情况下的服务持续性
c） 应急响应：建立安全事件应急响应机制，及时处置安全事件

参考文献

[1] GB/T 44109—2024 信息技术 大数据 数据治理实施指南
[2] GB/T 5271《信息技术 词汇》第1-34部分
[3] GB/T 39725-2020《信息安全技术 健康医疗数据安全指南》
[4] 《中华人民共和国个人信息保护法》[S].2021-08-20