数据管理平台项目技术解决方案

目　录

1.1.1

1.1.2 项目背景

数据科技发展水平是国家的核心竞争力，建设以知识服务为目标的科技资源已成为国家软实力的重要标志。当今世界，各国科技资源在影响区域决策、引导社会舆论、服务公共事务、体现国家软实力等方面发挥着重要作用。高端科技资源建设，不仅是国家进行宏观决策的有力支撑，也是推进国家治理体系和治理能力现代化的重要内容，加快以计算机科学与人工智能为代表的科技领域知识管理与服务能力建设，是科技强国的紧迫需求。在产业数字化发展的背景下，国家大力支持数据融合应用在产业创新发展中发挥更大作用。

为深入实施创新驱动发展战略，规范管理科技资源共享服务平台，推进科技资源开放共享，依据《国家科技资源共享服务平台管理办法》（国科发基〔2018〕48 号），《吉林省科技资源共享服务平台管理办法》，规范管理吉林省科技资源共享平台，推进科技资源开放共享，提高科技资源利用效率，促进创新创业，为加速吉林经济振兴提供科技支撑。

科学技术数据研究所是中国科学技术工作者的群众组织，是中国共产党领导下的人民团体，是党和政府联系科学技术工作者的桥梁和纽带，是国家推动科学技术事业发展的重要力量。汇聚科学技术数据研究内外部数据，引领数据资源的有效治理和共享融合，开展以数据的深度挖掘与融合应用为特征的智能化应用，打造动态感知、互联、智能的数据管理平台， 是科学技术数据研究数据化建设的重要内容。

1.1.3 项目概述

数据管理平台是基础支撑与条件保障类科技创新基地，平台面向全省科技创新、经济社会发展和创新社会治理，加强优质科技资源有效集成，提升科技资源使用效率，为科学研究、技术进步和社会发展提供数据化、社会化的科技资源共享服务，遵循合理布局、整合共享、分级分类、动态调整的基本原则，加强能力建设，规范责任主体，促进开放共享。

平台依托科学技术数据研究所学科门类齐全、领域交叉充分、智力资源密集的独特优势，聚焦科技领域，坚持问题导向，以全球视野动态汇聚、融合关联中国科协内外资源，构建面向全球科技领域的覆盖面广、权威性高、实时性强的知识数据资源池，形成“科技领域——专家人才——科研成果”的科技资源知识图谱，建成“研究兴趣/学术影响/研究方向”等立体、多维、高精度的专家画像标签体系，建成数据知识领域研究热点、趋势、人才态势感知服务，利用复杂网络关系分析、交互学习等挖掘技术为宏观数据管理与决策提供支持服务。

通过平台的建设，整理省内数据拥有单位的科学研究数据、检测数据、勘查数据等，建立起若干数据中心和主体数据库，搭建吉林省科学数据平台门户网站，为吉林省各行各业，特别是政府部门开展科技管理、决策，企业、高校、研究院所开展研发及横向联合、数据沟通，为发挥吉林省科教优势，促进经济发展提供及时有效的服务和支持。它是吉林省创新体系的重要组成部分，具有投入稳定、社会共享、

公益性和持续性等特点，对全省经济、社会和科技快速发展具有重要意义。

1.1.4 建设目标

数据科技发展水平是国家的核心竞争力，建设以知识服务为目标的科技资源已成为国家软实力的重要标志。本项目以科学技术数据研究所数据中心的大数据为支撑，构建大规模实体要素之间的知识网络图谱，形成立体全景科技态势:感知服务能力，为宏观科技管理与决策提供支持服务。进一步吸收、融合多来源异构数据，通过持续的数据治理，不断提高数据质量、扩大数据范围、提升数据服务能力；强化数据处理、数据管控和数据挖掘能力，建设更为丰富、更加精准的科技管理大数据服务，为不断提升科技管理现代化创新能力的需求提供全面的技术和数据支撑。

项目主要建设目的如下：

1.结合国家战略和吉林省经济社会发展的需求，持续开展重要科技资源的收集、整理、保存工作；

2.承接科技计划项目实施形成的科技资源汇交、整理和保存任务；

3.开展科技资源的社会共享，面向各类科技创新活动提供公共服务，开展科学普及，根据创新需求整合资源开展定制服务；

4.建设和维护在线服务系统，开展科技资源管理与共享服务技术研究和应用。

最终，实现加强优质科技资源有效集成，提升科技资源使用效率，为科学研究、技术进步和社会发展提供数据化、社会化的科技资源共享服务平台，推进科技资源开放共享，提高科技资源利用效率，促进创新创业，为加速吉林经济振兴提供科技支撑。

1.1.5 建设必要性

为进一步加强优质科技资源有效集成，提升科技资源使用效率，科学技术数据研究所依据“盘活数据资产、发挥数据效能，科学性、可行性、创新性、前瞻性相结合”的原则，统筹开展了数据管理平台建设工作，尝试在科技人才精准服务、科技人才成长规律以及科技人才区域流动等方面提供大数据决策支撑服务。系统以人、机构、成果为纽带和数据组织核心，对所有类型实体数据资源进行全面融合，形成融会贯通的大规模关系网络，并基于此实现了多类深层知识分析挖掘，在一定程度上，实现了科学技术数据研究现有业务数据资源与互联网数据资源的消歧与融合，在资源共享、业务协同、决策支持等方面取得一定效果。

数据作为生产要素的属性表明，其未来必将走向市场。数据应用范围将从传统的组织内部应用为主，发展为支撑内部和服务外部并重，数据资产应用和服务范围的扩大，将成为组织战略发展的一部分。今后一段时期，组织能否树立数据作为生产要素的战略意识，挖掘和利用数据价值、盘活数据资源，实现数据资产保值到增值，决定了组织能否迈出生产要素到生产力转化的重要一步。

1.1.6 建设依据

为深入实施创新驱动发展战略，规范管理科技资源共享服务平台，推进科技资源开放共享，依据《国家科技资源共享服务平台管理办法》（国科发基〔2018〕48 号），本平台的建设围绕吉林省深入实施创新驱动发展战略，重点利用科研设备设施、科学数据、生物种质、实验材料等科技资源而设立的专业化、综合性公共服务平台，构建大规模实体要素之间的知识网络图谱，形成立体全景科技态势感知服务能力。

1.1.7 项目现状

近年来，随着“科教兴省”战略的实施，尤其是党的十六届五中全会提出把增强自主创新能力作为科学技术发展的战略基点和调整产业结构、转变增长方式的中心环节以来，吉林省对科技的投入不断增加。到2021年全省科学研究与技术开发机构422个，其中政府部门所属独立研究与开发机构135个，高等院校所属科研机构170个,大中型工业企业办科研机构117个。从事科技活动人员8.2万人,其中研究与发展人员2.8万人。拥有中国科学院和中国工程院院士29人。全省已建国家及省级高技术研究重点实验室、工程技术研究中心(创新中心)等科技公共服务平台93个，经国家认定企业技术中心23个。全社会科技创新投入大幅度增长,2021年研究与发展活动经费(内部)支出50.9亿元，占全省生产总值的0.96%。

如此庞大的科技数据资源在管理方面，主要存在以下现象：

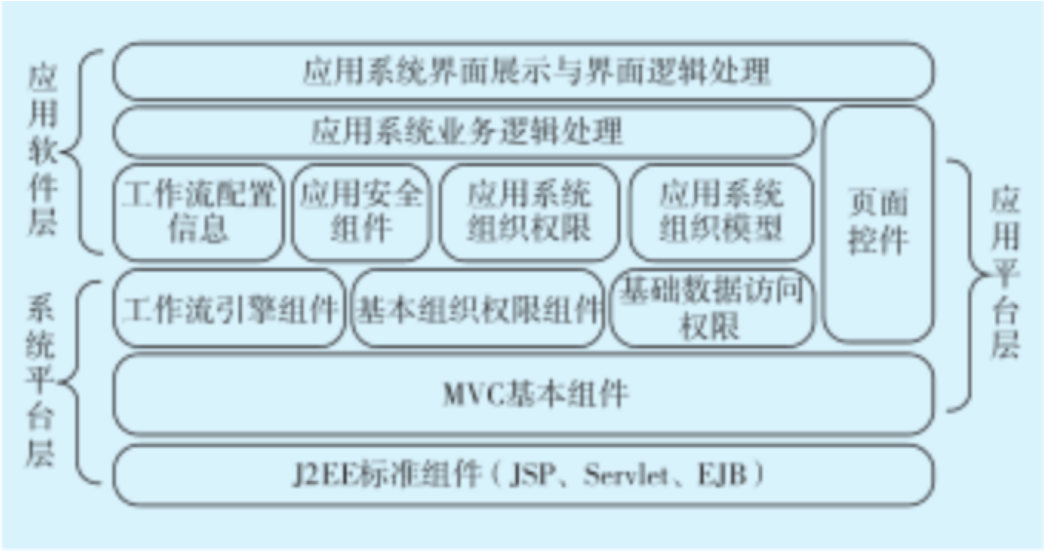
1. **海量“孤岛”科技数据难以共享**

。科技数据的数据源载体多、存储形式多样、数据类型广泛，数据资源孤立分散，给科研人员的数据分析、共享及管理过程造成了比较大的麻烦。在数据驱动的研究背景下，海量数据通过多种途径和方式获取，并存储在硬盘、数据库或其他存储介质中，则研究者每次进行数据分析时都要采用不同的方式调取数据分别管理。与此同时，数据共享的方式也比较局限，若是使用网盘共享，数据上传、下载耗时耗力；移动硬盘共享倘若后续发生数据修改也很难再同步给相关共享人；云计算虽然可以调用公开数据，但有些无法提供本地上传数据集的接口，也并不方便。

2. **科技数据数据资源配置不平衡。**吉林省科技数据资源配置集中度较高，少数的科研机构、高等院校占有大量的科技数据资源，对于多数企业，特别是中小企业而言一方面自身对于科技数据投入的认识不足，而且企业应用数据技术的水平偏低，应用范围只停留在设立企业网站上；另一方而购买大量的网络数据库资源需要雄厚的资金支持,往往大大超出企业的投资成本。
3. **科技数据存储安全性都没有保障。**传统的有限防护机制不一定能保障数据权益和数据安全，数据共享者将面临风险责任与权利受益的矛盾。一方面，科学数据本身具有可复制性，在共享中易被窃取，造成数据贡献者自身产权受到侵犯；另一方面，数据的集中化共享很有可能导致数据使用边界模糊，增加了数据误用、数据滥用等多重风险。现有大部分共享平台可追溯性差，即使数据泄露，参与用户也很难追究。

1.1.8 技术路线

系统应用软件采用基于组件的多层架构。最底层是系统平台层，主要基于标准的 J2EE 组件。上层是应用平台层，包括工作流引擎组织权限框架、基础数据访问组件等。这些组件分别封装了工作流、组织权限、数据访问等方面的基本功能部件，是应用系统构建业务逻辑的基础。在应用平台层之上,是由各种业务数据模型、配置数据、组织权限定义、应用系统的业务处理逻辑和界面控制逻辑等组成的软件系统。通过组件化拼装，形成了整个应用软件系统，并通过内部息互联确保整个系统稳定、有效地运行。同时这种架构已经充分考虑到未来系统的扩展性及集成性,为未来系统的扩容和与其他相关应用系统的整合提供技术保障。



技术架构

1)分布式缓存。

分布式缓存技术四用于动态 Web 应用以减轻数据库负担。它是通过在内存中缓存数据对象来减少读取数据库的次数,从而提高数

据库响应速度。

2)网页 HTML 静态化。

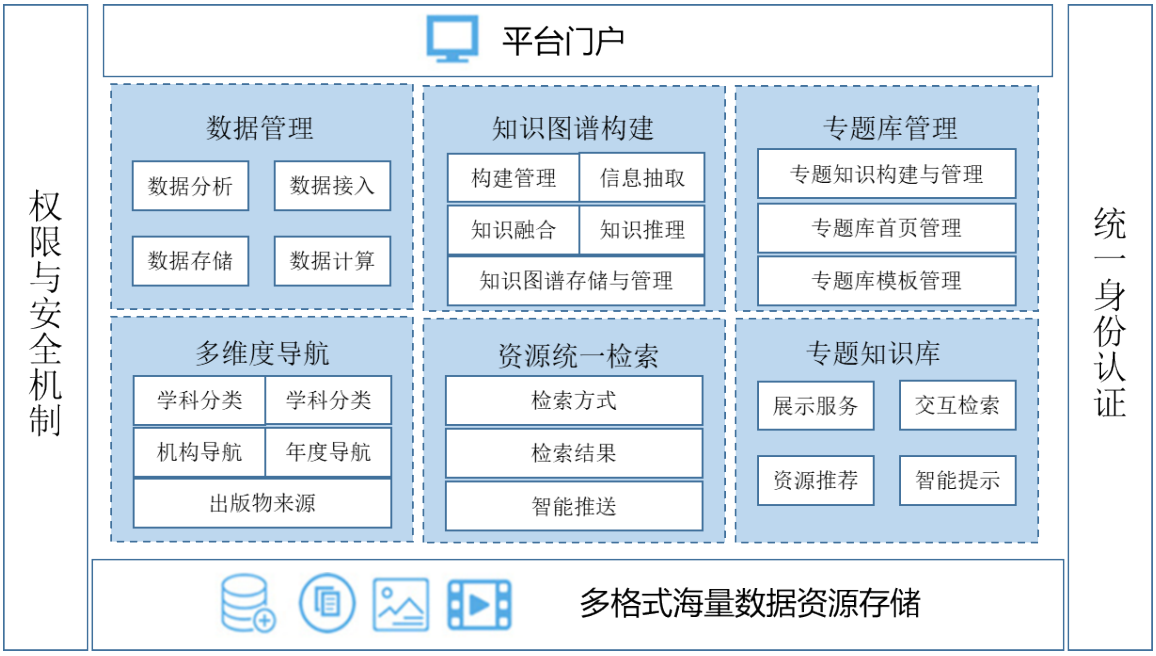
效率最高、消耗最小的就是纯静态化的 HTML 页面叫,因此本系统尽可能多地使网站上的页面采用静态页面来实现。由于本系统网页内容需频繁更新,采用了数据发布系统实现最简单的数据录人并自动生成静态页面,同时具备频道管理、权限管理和自动抓取等功能,避免了大量数据被前台程序调用,从而减少大量的数据库访问请求。.

3)数据库集群和负载均衡。

本系统采用了数据库集群技术,解决网站面对大量访问时数据库的瓶颈问题。负载均衡解决了网站高负荷访问和大量并发请求的快速响应问题。

1.1.9 总体框架

系统采用 B/S 架构即浏览器和服务器模式，用户通过浏览器输入指定的 IP 或者网址即可访问到管理系统。与传统的 C/S 架构相比，大大简化了客户端，使得客户端机器只要能上网就可以实现开发、维护等几乎所有工作都集中在服务器端，当企业对系统应用进行维护与升级时，只需更新服务器端即可，这节省了大量的时间与成本。同时系统要求:企业内部所有人员均需要能够进行简单操作，同时少数的系统管理人员会进行稍微复杂的管理操作；系统能够进行简单部署，集中管理。因此采用 B/S 结构模式进行开发较为恰当。



1.1.10 功能设计

数据管理平台的综合集成，是查询、统计、关联、图谱及可视化等各类功能的数据基石。数据管理平台实现了各来源科技数据资源的导入和集成管理，平台支持研究院现有业务数据资源导入并支持开放数据的获取。

平台功能主要包括数据管理、知识图谱构建、专题库管理、系统管理、多维度导航、科技资源统一检索。

1.1.10.1 数据管理

数据管理包含数据源分析、数据接入、数据存储与计算等。

1.1.10.1.1 数据来源

本项目中所用到的数据主要是甲方合作的商业数据：包含中国知网、万方数据、维普数据、国家科技图书文献中心、中国工程院知识中心、读秀、尚唯科技报告和产品样本库、中经数据库、万方、科慧项目数据和中国科学院计算机所的科学数据等。

所涉及到的数据通过数据库或者接口方式接入，类型包含但不限于：期刊论文、学位论文、会议论文、科技报告、产品样品、标准、科技成果、科技政策、人才数据等。

1.1.10.1.2 数据采集

(1)抓取 Internet 网络资源，可以对静态网页中的文本数据进行抓取和下载，可实现基于模板的网页数据提取和元数据抽取。

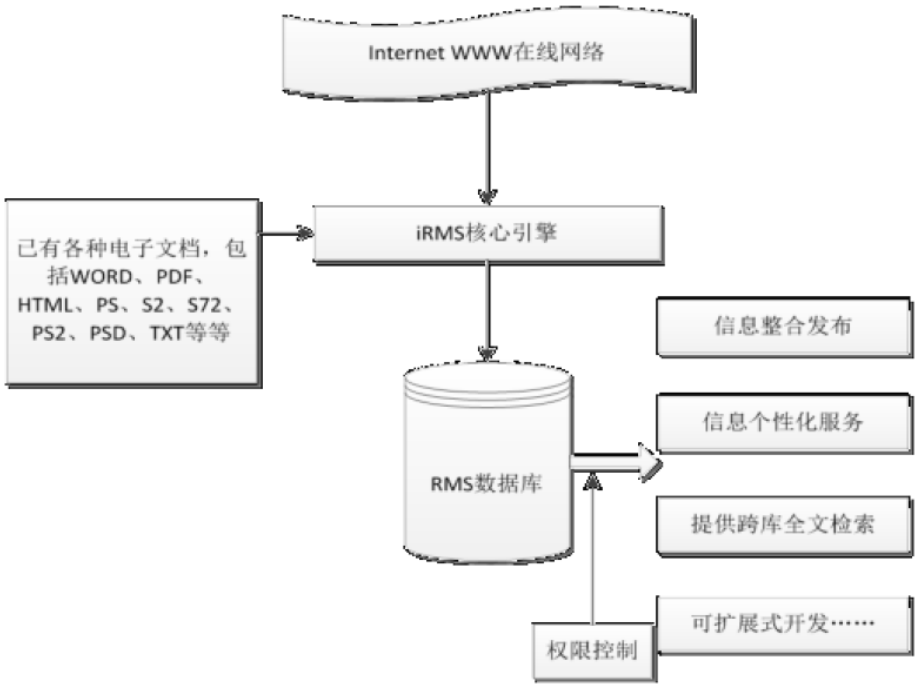
(2)自动下载网络数据库中的数据，可自动下载网络数据库的元数据以及其中的数据,并可根据数据的数据自动抽取数据库的元数据。

(3)对本地相关文档数据进行采集，可自动对本地数据资源属性进行抽取，并对资源进行相应标注。

(4)对数据进行自动过滤、归类和整理，对所采集的数据进行重复性过滤，去除重复资源，在此基础上对获取的资源进行归类，并对归类后的资源进行索引、重排等整理。

(5)根据用户需求，按照用户的个性化需求在资源中选择合适的数据提供给用户，并按照需求将相关数据加载到资源库中。

(6)对Internet进行检测和监控,为用户提供任务调度预设功能，根据用户的设定定期跟踪网络资源的变化情况,对新增资源进行识别，并在本地库中进行记录，实现实时网络监测和监控。



1.1.10.1.3 数据整合

根据不同数据资源所共有的标题、作者、单位、出版刊物、关键词、中英文摘要、参考文献等数据，整合到一个检索系统中，用户通过元数据对资源进行检索，系统的搜索引擎将遍览各资源数据库，最后将检索结果整合在一起将数据资源的概要和链接提交给用户。基于数据的整合，在用户提交检索请求前就已将数据资源整合到一起，因此在用户检索时期效率较高。

1.1.10.1.4 数据接入

数据管理平台提供数据源接入的功能，通过监控数据源的数据，实现实时及离线数据的同步，如果是实时的数据，会转发到数据分发服务上，由数据分发服务对数据进行实时分析，与存储。计划支持关系型数据，或者通过监控数据库的 binlog，来实现数据的同步。在数据同步方式建立好，需要通过配置的方式，将源数据的属性信息与数据平台的数据仓库的属性进行关联，这样才能完成从数据源将数据转化为数据仓库的数据结构，适应后面的数据清洗、计算、归总等处理过程，通过提供数据源，数据源的字典等信息，将数据导入到数据平台。

平台支持不同种类、不同数据源、不同目标库的数据接入。支持 Oracle、Sql-Server、My-Sql、H-base、Hive 等主流数据库,支持常用文件类型： XML、CSV、EXCEL 等。

数据库接入方式：

1. ODBC 方式联接

ODBC(Open Data Base Connectivity)翻译过来就是开放数据库互联。是由微软主导的数据库链接标准。是一种底层的访问技术,ODBC API 可以让客户应用程序能从底层设置和控制数据库,完成一些高级数据库技术无法完成的功能；但不足之处由于 ODBC 是只能用于关系型数据库，使得利用 ODBC 很难访问对象数据库及其他非关系数据库。

2. DAO 方式联接

DAO(Data Access Object)数据访问对象型。不提供远程访问功能。只提供了一种通过程序代码创建和操纵数据库的机制。最大特点是对 MICROSOFT JET 数据库的操作很方便,而且是操作 JET 数据库时性能最好的技术接口之一。并且它并不只能用于访问这种数据库，事实上,通过 DAO 技术可以访问从文本文件到大型后台数据库等多种数据格式。Microsoft Jet 为 Access 和 Visual Basic 这样的产品提供了数据引擎。

3. ADO 方式联接

ADO(ActiveX Data Object)，是 ActiveX 数据对象，是基于 OLE DB 的访问接口,它是面向对象的 OLE DB 技术,继承了 OLE DB 的优点。属于数据库访问的高层接口。是在 OLE DB 规程下开发的,基于 OLE-DB 建立连接的局部和远程数据库访问技术。同 OLE-DB 一样，它要“年轻”些。使用中，我们一般用 OLE-DB 和 ADO 替代 DAO 和 RDO。

1.1.10.1.5 数据存储与计算

（1）数据存储

分布式存储系统满足海量数字媒体资源的分布式存储，存储平台实现以下功能点：

- 数据加密（不存储裸数据，按块加密存储）；

加密系统是由明文、密文、算法和密钥组成。发送方通过加密设备或加密算法，用加密密钥将数据加密后发送出去。接收方在收到密文后，用解密密钥将密文解密，恢复为明文。在传输过程中，即使密文被非法分子偷窃获取，得到的也只是无法识别的密文，从而起到数据保密的作用。

- 海量的数据存储能力（亿级的存储能力）；

提供基于分布式文件系统和并行架构的大数据存储能力，支持PB级数据规模的高可靠和高可用存储，支持存放多种文件格式。

- 具备持续的灵活的扩容能力；

支持系统盘和本地盘扩容，弹性按需扩容。

- 支持每天百万级文件数以上写入；

利用页缓存技术 + 磁盘顺序写和零拷贝技术实现每天百万级文件数以上写入。

- 支持每天千万级文件数据读取；

通过采用开辟大块连续磁盘空间的方式来存储大量文件，也将逻辑上连续的数据尽可能地存储在磁盘阵列的连续空间上。

- 通过负载均衡能够持续提高系统吞吐量；

负载均衡提高系统的吞吐量，有效降低系统的单点故障率，让系统降低对外网端口的依赖，