干细胞领域知识发现大数据平台建设与应用

张志强[1] 胡正银[1] 杨 宁[1] 文 奕[1] 覃筱楚[2] 宋亦兵[2] 潘光锦[2]

（1. 中国科学院成都文献情报中心；2. 中国科学院广州生物医药与健康研究院）

摘 要

“面向干细胞领域知识发现的科研信息化应用”是中国科学院“十三五”信息化专项课题之一。该课题围绕干细胞领域科学研究与知识发现对科研信息化的迫切需求，综合利用科研大数据与新一代人工智能技术研发了“干细胞领域知识发现大数据平台”，为干细胞科研活动与科技管理提供专业、高效、精准的知识服务。本文从干细胞领域知识发现大数据平台的客观需求出发，系统介绍了平台建设的总体目标、技术路线、关键技术、建设成果和服务成效，并对平台的下一步建设思路和发展方向进行了展望。

关键词

科研信息化；知识发现；大数据平台；干细胞

Abstract

The project of “E-Science Application for Knowledge Discovery in Stem Cells” is an important part of the 13th Five-Year Plan e-Science Programme of Chinese Academy of Sciences. Focusing on the critical needs of science and technology innovation and subject knowledge discovery in stem cell research for e-Science, the project makes comprehensive use of big data and new generation of artificial intelligence technologies to develop a platform of “Stem Cell Subject Knowledge Discovery” (SCKD), which can provide professional, efficient and accurate subject knowledge discovery services for the research and S&T management in stem cell. This paper introduces the project’s objectives, tasks, key technologies, achievements and service effect, and discusses the future work of SCKD.

Keywords

e-Science; Knowledge Discovery; Big Data Platform; Stem Cell

1 引言

干细胞是当今生命科学研究的热点和前沿，以干细胞技术为核心的再生医学有望成为继药物治疗、手术治疗之后的第三种疾病治疗途径，正孕育着重大的科学突破与巨大的产业带动[1]。在干细胞领域，以科技文献、科学数据、临床试验、医药产品与科技服务资源为核心的科研大数据呈“井喷式”增长，科学研究日益成为数据驱动的知识发现活动，大数据驱动的知识发现与技术突破正成为科技创新的新引擎，集成科研大数据与知识计算环境的知识发现平台已成为科研活动的重要工具。干细胞领域科研大数据具有数量巨大、类型繁多、关系复杂和来源分散等特点，如何从海量的多源异构数据中进行

知识的自动化抽取、结构化组织、语义化关联与知识化计算，以及从中高效、精准地进行有价值的知识挖掘是知识发现的关键。目前国际上类似的数据集成知识发现平台，如哈佛大学干细胞研究所研发的干细胞创新引擎（Stem Cell Commons）[1]、数据密集型生物医学研究平台 Galaxy[2] 等，主要提供学科领域科学数据管理、计算和分析服务，存在数据类型单一且缺乏关联、知识计算算法有限等不足，难以满足学科领域研究热点、研究重点与发展趋势分析，以及基于大数据知识计算的关键技术挖掘与技术预见等学科知识发现需求。

针对大数据时代科研信息化应用的新形势和新挑战，面向干细胞领域知识发现的需求，课题组综合运用信息抽取、自然语言处理、本体、知识融合、机器学习、知识图谱、知识计算与可视化分析等文本挖掘与新一代人工智能技术，结合中国科学院成都文献情报中心（以下简称成都中心）科技文献、数据资源优势与中国科学院广州生物医药与健康研究院（以下简称广州生物院）干细胞科研优势，研发了干细胞领域知识发现大数据平台。该平台通过构建干细胞领域的知识图谱，研发领域专业知识计算环境，实现了干细胞领域多源异构的科研数据融合与知识关联，有效打破了数据孤岛，为广州生物院及中国科学院其他研究单元的干细胞科研活动提供全面、专业、精准、高效的数据获取、信息推送、知识发现与情报支撑服务，推动大数据驱动的知识发现应用，推进科研活动与信息化的融合，提升科研信息化应用水平。

2　干细胞领域知识发现大数据平台的客观需求

2.1　干细胞技术与再生医学的发展需求

国家统计局 2020 年 2 月 18 日发布的《2019 年国民经济和社会发展统计公报》[3] 显示，目前我国 60 周岁及以上人口数约为 2.54 亿人，大概占总人口的 18.1%；而到 2050 年，我国老年人口总量估计将超过 4 亿人，老龄化水平将达到 30% 以上。这种严峻的老龄化趋势使得危害我国人民健康的重大疾病谱系已经发生了巨大变化，与衰老密切相关的组织器官损伤、心血管等功能脏器的衰竭、退行性疾病、癌症等已成为主要疾病类型。而针对这些疾病，传统治疗手段如药物治疗和手术治疗等往往收效甚微，难以满足现阶段人民群众日益增长的医疗需求。同时，人类在对深海、太空、高原等领域的不断开拓中也会面临着低氧、高压、辐射等巨大的健康挑战。这一系列难题仅通过常规医疗途径难以得到彻底解决。以干细胞技术为核心的再生医学研究将为以心脑血管疾病、癌症、糖尿病、帕金森综合症、阿尔兹海默症等为代表的与衰老密切相关的疾病的防治带来革命性变化，并将为新型颠覆性医疗技术的诞生奠定基础。

目前，再生医学已经成为各国政府、科技和企业高度关注和大力投入的重要研究领域，成为代表国家科技实力的战略必争领域。1999 年以来，干细胞与再生医学领域的研究成果先后 11 次入选 *Science* 年度十大科技突破。2012 年，诺贝尔生理医学奖授予细胞核重编程及诱导多功能干细胞研究领域的两位科学家，彰显科学界对干细胞和再

1　https://hsci.stemcellcommons.org/。
2　https://usegalaxy.org/。
3　https://www.gov.cn/xinwen/2020-02/28/content_5484361.htm。

生医学的高度重视。在国内，再生医学的重要性已引起相关决策部门和科技界的高度关注。2015 年，科技部启动了国家重点研发计划“干细胞与转化医学”重点专项，旨在从国家层面推动干细胞与再生医学方面的研究，整体提升我国在该领域的核心竞争力。中国科学院于 2011 年开始实施国家“干细胞与再生医学研究”战略性先导科技专项计划，取得了一批创新性研究成果。在此基础上，中国科学院正在筹建“干细胞与再生医学创新研究院”，将围绕干细胞与再生医学来推动第三次医学健康革命。

2.2 面向干细胞领域知识发现的科研信息化需求

科研信息化实质为科学研究活动本身的信息化，其特征是充分利用网络信息基础设施与信息化技术，促进科技资源交流、汇集与共享，变革科研组织与活动模式。随着信息化的发展，科研仪器、科学实验及科学交流等一系列科研活动每时每刻都产生着海量、异构、多元化的数据信息，科学研究日益成为数据驱动的知识发现活动，d-Science（数据驱动的科学）时代来临，并步入了以数据为中心来思考、设计和实施科研活动，通过对海量数据的处理和分析获得科学发现的第四范式——“数据密集型科学发现”范式[2]。各学科领域的研究对象已不再是单一的孤立系统，而是涵盖更大范围、涉及多个学科的复杂创新系统。各学科领域数据大量、快速增长，使得每个学科都出现了二元发展的态势——“X 信息学”（X-Informatics）[2]。以生物医药领域为例，出现了用于理解大量数据所包含的生物学意义的生物信息学（Bio-informatics）[3] 和分析病人健康与社会卫生信息的医学信息学（Medical Informatics）[4]。

第四范式在生命与大健康领域已得到广泛应用，其核心是多源异构数据的集成与海量数据的分析。数据集成方面，生物大分子序列测定与“人类基因组计划”得到的相关生物学数据越来越多，快速推动了生物信息学的发展。序列数据库 Swissprot、GenBank、PharmGKB、IPA 等多种生物医学数据库与平台的发展，为多种知识挖掘技术与工具的开发提供了丰富的资源支持。数据分析方面，也出现了众多大数据分析软件和平台，如 IBM 公司推出了基于大数据分析与人工智能的 Watson 医疗系统[4]。欧盟第七框架计划支持研发了用于支持小分子筛选、新药设计的生物医药知识发现系统 Open Phacts[5]。Open Phacts 利用知识图谱技术将从分子到基因组，再到患者的各种数据集关联起来，并利用深度学习算法发现潜在的知识与隐含的知识关联[5]。英国生物科技公司 Benevolent Bio 利用大数据知识发现平台 JACS（Judgment Augmented Cognition System），从全球范围内海量的学术论文、专利、临床试验、患者记录等数据中，提取出有用的信息，发现新药研发的蛛丝马迹[6]。借助 JACS 的分析能力，Benevolent Bio 发现多个可用于治疗肌萎缩性侧索硬化症的潜在化合物，极大地提升了药物研发的效率[6]。

结合广州生物院的干细胞科研需要，面向干细胞领域知识发现的科研信息化应用需求主要集中在三个方面：

（1）干细胞研究热点、研究重点与发展趋势分析。科研管理人员希望通过文献计量、科学计量、知识计量等方法，对包括科技政策、科研项目、科技论文、专利、临床试验、新药说明书等在内大量干细胞领域科技文献与科技信息进行统计、分析与挖掘，

4 https://www.ibm.com/watson。

能动态发现干细胞领域具体的研究热点、研究重点与发展趋势，为“定标赶超”提供科学依据与参考。

（2）基于大数据知识计算的干细胞关键技术分析与技术预见。科研人员希望通过对多源异构的干细胞科研大数据进行数据挖掘与知识计算，结合知识图谱、机器学习和深度学习等技术，构建可视化的细胞组织预测模型和其他知识发现工具。这些工具可帮助预测癌症和其他一些疾病的细胞布局变化，可以解决一些复杂的干细胞关键技术问题和揭示大量丰富的细胞生物学基础信息，加速干细胞研究、癌症研究和药物开发方面的进展，为更精准的技术预见提供支撑。

（3）干细胞研究科研数据一体化管理。在干细胞科研活动过程中，常常需要利用、分析大量的“细粒度”数据类对象，包括科技文献中的知识单元、基金信息等，科学仪器中的实验数据从产生到分析、结果利用，科学研究用实验动物的进站检疫实验过程及实验数据的连续追踪等。通过构建一体化的干细胞研究科研数据管理平台，可以打通不同来源数据“孤岛”，有效提升科研效率与信息化服务水平。

3　干细胞领域知识发现大数据平台技术框架

3.1　总体目标

围绕干细胞领域知识发现对科研信息化的需求，综合利用大数据与新一代人工智能技术研发“干细胞领域知识发现大数据平台”。通过系统的数据搜集与数据标准化体系建设实现领域多源异构数据的集成与整合，通过专业的知识管理与知识计算推进领域知识的精准化表达和智能化关联，通过丰富的可视化界面提供多种“一站式”智能检索模式的科技情报与知识服务，助力领域科研人员开展知识发现研究，推进国家在干细胞领域的技术预见和战略布局工作。

3.2　技术路线

围绕以上总目标，课题组从干细胞知识图谱、知识计算环境与知识发现服务平台三方面来建设干细胞领域知识发现大数据平台。其中，知识图谱是数据基础，知识计算环境是分析工具，知识发现服务平台则是服务窗口，其总体架构如图 1 所示。

3.2.1　构建干细胞知识图谱

课题组综合运用知识抽取、知识挖掘、知识融合与可视化技术，及时、准确、规范地对分散在不同数据源中的干细胞“政、产、学、研、医、用”科技文献、科技信息、科学数据与科技服务资源等科研大数据进行集成，对其中蕴含的知识点进行自动化抽取、结构化组织与知识化关联，以构建干细胞知识图谱，实现“多形态－多粒度－多维度”数据、信息与知识的有效融合，为知识发现提供高质量数据支撑。干细胞知识图谱构建可分为基础数据汇聚、知识内涵挖掘与知识语义关联三个步骤，具体流程如图 2 所示。

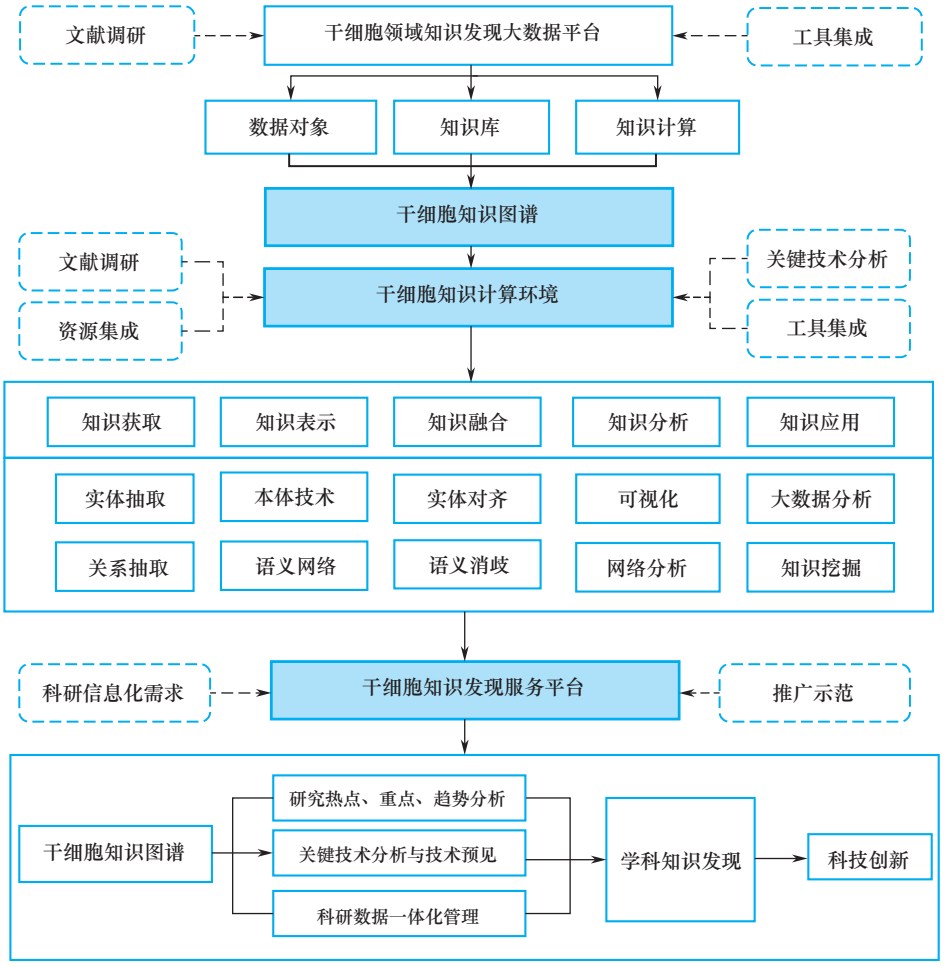


图 1　干细胞领域知识发现大数据平台总体架构

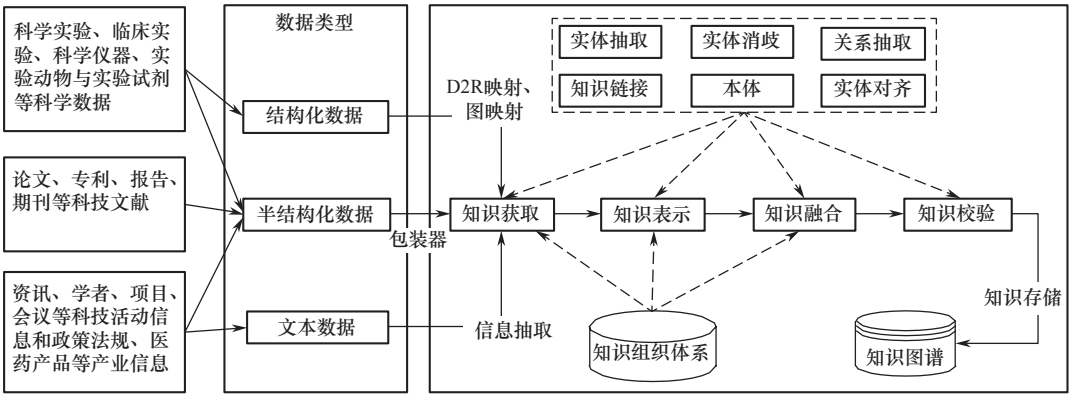


图 2　干细胞知识图谱构建流程

（1）基础数据汇聚。汇聚干细胞领域的论文专利、基金项目、临床试验、产品法规、专家机构等多源异构科研数据，并建立长期更新机制。分别采用关系型数据库、图

数据库与 Solr 索引技术对科研大数据中的基础数据、知识图谱三元组数据及面向知识发现平台的综合服务数据进行有效存储与管理。

（2）知识内涵挖掘。知识内涵挖掘是构建知识图谱的关键。课题组采用基于本体的知识抽取技术，从科学仪器、动物模型、实验技术、细胞器官、疾病基因等科研人员关注的视角挖掘干细胞领域的知识内涵。首先，参考 UMLS 超级词表[7]、Stem Cell Commons 等领域知识本体和知识服务平台，利用生物医药领域知识抽取工具 SemRep[8] 对部分核心科技文献进行文本挖掘，获得知识实体及实体之间的关系，以形成干细胞知识图谱的核心知识组织体系。然后，按照知识组织体系对干细胞领域的其他科研数据进行知识实体抽取、语义标注和数据融合，以进一步丰富干细胞知识图谱的实例数据。

（3）知识语义关联。综合科学计量学指标与文本挖掘技术，基于引用、致谢、合作网络、知识实体共现等关系，建立知识图谱中各类科技信息、知识实体等之间的语义关联。目前，干细胞知识图谱定义了“文献－文献、文献－知识实体、知识实体－知识实体”三大类共 58 种语义关系，其核心是以 Subject-Predicate-Object（SPO）语义网形式呈现的“知识实体－知识实体”之间的语义关系。

3.2.2　研发干细胞知识计算环境

干细胞知识计算环境集成了干细胞及相关领域的算法、模型、软件与工具，实现了数据管理、数据清洗、数据挖掘及报告输出等功能，可提供数据可视化分析、知识推理等知识计算服务。从流程上，知识计算环境分为数据管理、数据处理与数据分析三个部分。

（1）数据管理模块。数据管理模块是知识计算环境的基础，主要提供数据接引、整合及存储等情报分析与知识发现所必需的数据管理功能。除了可直接利用的干细胞知识图谱数据外，数据管理模块还可以摄入外部数据，对外部数据按规则进行清洗、转换和整合，还可将数据集关联后进行合并处理。数据存储模块支持对数据进行预览、导出、追加和共享等操作。共享分区的数据可被其他用户使用。

（2）数据处理模块。数据处理模块是知识计算环境的核心，主要提供算法管理、数据管理、任务管理等功能。算法管理模块实现对系统算法的统一管理。目前，知识环境集成了机器学习、推荐系统、自然语言处理等 30 种算法，正在集成干细胞领域专用计算模型。

（3）数据分析模块。数据分析模块是对知识计算结果进行多维度、细粒度、多类型的可视化分析与展示。数据可视化支持多种图表效果展示，如柱状图、折线图、饼状图、散点图、热力图、地图、雷达图、漏斗图、词云、关系图等。此外，还可通过仪表盘对图表进行集中展示。仪表盘数据会随着干细胞领域知识发现数据的变化而动态更新，并支持微信、微博、QQ 空间等多种方式分享。干细胞领域知识发现大数据平台还支持直接将分析结果生成分析报告。分析报告模块实现了分析报告分组管理，以及分析报告的新建、分享、导出等操作功能。

3.2.3 建设干细胞知识发现服务平台

干细胞知识发现服务平台主要提供干细胞科研大数据集成化管理、干细胞科技信息“一站式”智能检索与基于知识图谱的精准知识检索、干细胞科研热点前沿探测与干细胞科研画像四类知识发现服务。其技术路线介绍如下：

（1）干细胞科研大数据集成化管理。分别利用关系型数据库 MySQL 与图数据库 Neo4j 对干细胞科研大数据进行集成化管理。关系型数据库存储原始的基础数据，图数据库则存储经过知识加工的知识图谱数据。Neo4j 是一种 NoSQL 类型的数据库，可以灵活扩展数据结构与类型，符合以三元组为核心的知识图谱数据管理的需求。

（2）干细胞科技信息“一站式”智能检索与基于知识图谱的精准知识检索。利用 Solr 分面检索技术，对干细胞科研大数据与知识图谱中的知识组织体系进行分面索引。用户可通过 Solr 索引对所有类型干细胞科技信息资源进行“一站式”智能检索，以及基于干细胞知识组织体系来进行精准的知识检索与知识导航。

（3）干细胞科研热点前沿探测。从国际研发重点、中国研发重点、中国科学院重点突破方向等不同层面，归纳出一系列干细胞领域的热点前沿研究主题。这些主题包括相关的论文、专利、项目、新闻、专家与机构等信息。

（4）干细胞科研画像。采用知识图谱与画像技术对干细胞科研热点前沿主题、科研人员及科研机构，从论文、专利、项目、新闻及研究热点等多个角度进行科研画像。

3.3 关键技术

平台建设过程中所用到的关键技术主要包括知识图谱实体对齐、面向知识发现的知识计算和基于多维索引的统一数据视图技术。

3.3.1 干细胞知识图谱实体对齐

实体对齐是知识融合的基础，是知识图谱是否规范及具备可扩展性的前提。实体对齐是指发现相同或不同数据源中两个或多个实体是否指向真实世界同一知识对象的过程 [9]。知识图谱实体对齐的目标是能够高质量链接多个不同来源的数据，从顶层创建统一的知识表示规范，从而帮助计算机更好地理解数据，为知识计算与知识发现提供高质量数据 [9]。课题组主要通过两个方面进行干细胞知识图谱实体对齐。

（1）基于唯一标识符的数据规范化。通过对不同数据来源的知识对象进行数据清洗、筛选和规范化，将不同数据来源中表示同一对象的实体归并为一个具有统一标识符的知识实体添加到知识图谱中。例如，使用 DOI、专利号、ORCID、项目编号、概念编号等唯一标识符分别对期刊论文、专利、研究人员、科研项目及知识概念等科研数据进行实体对齐 [10]。

（2）基于标准的知识概念对齐。对概念、术语等知识概念进行准确的实体对齐是保证干细胞知识图谱质量的关键。该部分工作主要利用 UMLS[7] 中标准化的生物医药超级叙词表与 SemRep[8] 知识抽取工具完成。SemRep 是美国国家医学图书馆的语义知识表示项目（Semantic Knowledge Representation，SKR）的重要成果之一，是一款基于自然

语言处理技术和 UMLS 的语义知识抽取与表示工具。SemRep 以 UMLS 中的超级词表、语义网络和专家辞典为基础，专指性较强，反映学科知识也较具体，可以高效、精准地从生物医药科技文本中抽取知识实体及知识实体之间的语义关系。首先，利用 SemRep 从生物医学文本中自动抽取 Subject-Predicate-Object 三元组结构。其中 Subject、Object 是 UMLS 超级词表中的规范化概念，Predicate 则是 UMLS 语义网络中的标准化语义类型。然后，对所获得的知识实体参照干细胞知识组织体系进行映射。为保证知识图谱数据的质量，还采用基于规则的自动检测和人工辅助的方式对知识图谱中的实例数据进行校验。

3.3.2　面向知识发现的知识计算技术

面向知识发现的计算分析面临的首要问题是多源异构数据问题。在传统数据计算中，数据来源较单一、格式较规整、关联较简单，数据计算技术较成熟。但是在面向知识发现的计算环境下，不仅数据源多种多样，而且数据格式复杂，关联较多，是典型的多源异构数据。传统数据挖掘算法不能直接应用于多源异构数据。

知识计算环境通过采用开放计算接口方式解决上述多源异构数据的知识计算问题。除了提供内置的数据清洗、文本挖掘、机器学习等算法，以便用户直接使用外，知识计算环境还具备良好的扩展性，支持用户自定义开发新算法，并且可以将新算法和内置算法结合在一起去创建新的业务模型，满足特定知识发现需求。

3.3.3　基于多维索引的统一数据视图

面向用户知识服务的数据对象既包括各类科技文献、临床试验、科学数据与科技服务资源等科研数据，还包括知识图谱中的知识实体、知识计算环境新生成的显性知识等。其存储形式则分为关系型数据表和 RDF 三元组，需要采用统一的数据视图来屏蔽这些数据对象形式上的差异。多维索引是一种对多形态、复杂数据进行多层次、多角度索引的技术，它可用于整合多源异构信息，提供统一的数据视图，在数据集成与知识融合等信息系统中得到了广泛的应用[11]。

课题组采用多维集成索引技术对干细胞科研大数据进行索引，以向干细胞知识发现服务平台提供统一的数据检索接口。具体而言，多维集成索引采用 Apache Solr 索引技术，对不同数据类型的元数据、知识实体、知识组织体系等进行单独索引和集成。索引字段包括知识发现所有常用字段，如知识资源类型、资源题名、作者、发明人、机构、申请人、机构类型、出版年代、来源、关键词、分类号、科学仪器、实验动物、实验方案、实验试剂、方法技术、细胞、器官、疾病、基因、科研活动、科研产出等[11]。根据不同数据源的数据更新频率采用相应的索引更新机制，包括日更新、周更新、月更新等。

4　平台建设成果和服务成效

“干细胞领域知识发现大数据平台”（https://stemcell.kmcloud.ac.cn/）秉承“边建设，边服务，边完善”的原则，取得了较丰富的成果和较显著的服务效益，凝聚了一批稳定

的学科社区用户。

4.1 建设成果

干细胞知识图谱集成了大量包括科技文献、科技信息、科学数据与科技服务资源在内的干细胞领域科研数据，知识计算环境具有高性能的数据处理能力和丰富的可视化方式，干细胞知识发现服务平台则实现了四项核心功能并提供相应的知识服务，可有效满足本领域科技人员、管理人员和决策人员的不同需求。

4.1.1 干细胞知识图谱

从 103 个权威核心数据源中集成了四大类 16 小类的干细胞领域科研大数据 40 余万条，包括：①论文、专利、报告、期刊、专著等科技文献；②新闻资讯、干细胞研发动态、专家、项目、政策法规等科技信息；③临床试验、医药产品、科学实验等科学数据；④科学仪器、实验动物与实验试剂等科技服务资源。在此基础上，从科学仪器、实验动物、实验方案、实验试剂、方法技术、细胞、器官、疾病、基因等视角，利用知识抽取技术从中挖掘出 2 万多条知识实体。利用这些知识实体对干细胞科研数据进行了多维度、细粒度的语义标注以形成干细胞知识图谱，实现了领域知识内涵的深度挖掘与知识关联。目前，干细胞知识图谱总数据超过 220 万条。

4.1.2 干细胞知识计算环境

干细胞知识计算环境基于私有云和 Spark 进行构建，其知识计算框架集成了 26 种数据清洗、加工与融合规则，以及自然语言处理、分类回归、推荐、结果评价等 30 种通用数据挖掘算法。该环境还集成了 20 个干细胞相关知识计算模型，可提供柱状图、折线图、饼图、漏斗图、雷达图、词云图等 12 种可视化方式和 1 个基于模板的报告自动生成工具。

4.1.3 干细胞知识发现服务平台

基于学科知识发现的具体需求，干细胞知识发现服务平台实现了 4 项核心功能：①干细胞科研大数据集成化管理。集中管理各种类型的干细胞科研数据，可快速向用户提供个性化数据服务。②“一站式”智能检索与精准知识检索。通过一键检索，可获取干细胞新闻资讯、论文专利、基金项目、医药产品、政策法规、产业情报等多类型科技信息。此外，用户还可依据 2 万余条干细胞知识点进行检索与查阅，无须详读全文即可快速、全面地掌握科技文献内容。③热点前沿探测。从国际、国家研发重点及中国科学院重点突破方向等不同层面，挖掘出 22 个干细胞热点前沿主题。④科研画像。利用知识图谱数据，对科研机构、科学家等科研创新主体进行画像。此外，课题组还研发了平台对应的微信小程序“干细胞助手”，将服务扩展到移动端。知识发现服务平台的部分界面截图如图 3 和图 4 所示。

平台还支持开展了干细胞领域创新主题识别与演化分析、科研合作社区识别、高质量专利挖掘、疾病与基因关联关系挖掘、学科知识结构画像等知识发现应用。

图 3　干细胞领域科研数据“一站式”检索

图 4　基于知识图谱的精准知识检索与知识导航

4.2 服务成效

干细胞领域知识发现大数据平台于 2018 年 11 月正式发布，采用签约用户方式提供服务。该平台受到了中国科学院相关研究所、中国医学科学院、清华大学及剑桥大学、德国癌症研究中心等 20 多家机构的 200 余名科研人员的广泛关注，平台用户数稳步增加（见表 1）。

表 1　干细胞领域知识发现平台用户访问情况

月份	总访问量（人次）	网页访问（点击数）	访问流量（GB）	平均在线时间（s）
2018.12	383	10296	1.31	17.26
2019.01	470	11187	1.49	18.19
2019.02	416	8908	0.94	18.26
2019.03	435	12039	1.74	18.11
2019.04	786	16954	2.48	17.12
2019.05	1123	35773	5.82	17.90
2019.06	3982	87365	8.19	39.90

与国际上同类平台如 Stem Cell Commons 相比，干细胞领域知识发现大数据平台具有“数据全面”与“服务精准”两大优势。该平台充分发挥成都中心的科技信息资源与专业知识组织优势及广州生物院的干细胞科研优势，构建了高质量的干细胞知识图谱，可提供从数据、算法、软件到应用的一站式服务。除普惠信息服务外，该平台还为广州生物院的科研工作、知识发现与决策咨询提供了个性化知识支撑服务，可以有效支撑重大科研项目申报和干细胞领域战略布局，典型案例如表 2 所示。

表 2　个性化知识支撑服务典型案例

类型	内容	效果
支撑干细胞战略布局	为撰写《广州再生医学与健康广东省实验室建设方案》提供数据支撑	2018 年实验室获批启动
	为撰写《粤港澳大湾区人类细胞谱系大科学设施建设方案》提供数据支撑	启动前期预研
支撑重大科研项目	为国家重点研发计划项目“人多能干细胞分化过程中谱系命运决定的调控及异质性机制研究”知识产权分析提供数据支撑	为项目申请专利提供知识产权分析
	为成都中心参与的国家重点研发计划项目“成渝城市群综合科技服务平台研发及应用示范”提供理论、方法与技术支撑	支持项目顺利开展
	为广州生物院参与申报国家重点研发计划“珠三角城市群典型产业综合科技服务应用示范”提供数据支撑	项目立项
支撑信息情报与决策咨询工作	撰写决策咨询建议 4 份	决策参考
	完成干细胞研究报告 6 份	直接服务于广州生物院的科研管理工作
	编辑《干细胞研发动态》快报 16 期	发送相关课题组，同时寄送管理部门领导参阅
	支持建立全球首个“人胚胎基因编辑”法律法规数据库	向相关领导提供“人类胚胎实验伦理规范”应急数据和情报服务

自新冠肺炎疫情发生以来，广州生物院高度重视，紧急启动应急反应机制，组织研究院优势力量进行科技攻关，围绕抗病毒药物研发、肺炎临床救治方案、病毒快速检测、疫苗研发、动物感染模型建立及致病机理研究等方面开展了相关工作，并取得重要成果。根据科研应急需求，课题组通过平台进行文献大数据挖掘，分析了特定基因与病毒、免疫共现关系，为科研人员筛选可能与病毒相关的基因提供了有效的科研信息化支撑。

5　总结与展望

“干细胞领域知识发现大数据平台”建设与服务的宗旨是融合多源信息、打通数据孤岛，挖掘知识关联、放大数据价值，集成知识计算、促进知识发现，推进科研活动与信息化的融合，支撑研究所重大创新，实现国际先进的科研信息化应用。在中国科学院“十三五”科研信息化专项的支持下，该平台构建了干细胞知识图谱，集成了专业知识计算工具，初步实现了基于科研大数据的知识发现服务，打造了面向干细胞领域知识发现的科研信息化应用示范。未来，课题组将继续遵循干细胞大数据应用发展的长远目标，着力于从领域数据资源、典型应用示范和领域知识发现三方面继续推进干细胞领域知识发现大数据平台的建设。

（1）推进干细胞与再生医学科研大数据中心建设，形成领域数据标准体系。提前谋划，集成未来可能的“卡脖子”科研数据资源，包括领域相关的通路、蛋白质结构、代谢组学及临床试验数据等。进一步集成多类型、多来源、多形态的科技信息，如实验视频、科普、微信、博客数据及评论信息等。进一步梳理干细胞领域科研大数据的规范与标准，制定规范统一、灵活、可扩展的领域元数据规范与数据质量标准体系。

（2）进一步优化知识计算环境，拓展应用示范场景。将国内外主流的干细胞与再生医学领域的知识计算模型、方法工具整合到知识计算环境中，构建领域知识计算工具导航。面向干细胞知识发现的具体需求，针对干细胞领域创新主题演化分析、领域专家合作态势与科研社区识别等场景，设计相应的标准化、规范化流程方法，构建“一站式”的知识计算模型。

（3）深入开展定制化知识发现研究与应用，助力干细胞科学研究。进一步征求领域专家的意见，按照专家基于大数据知识发现的需求，深入开展定制化科学大数据分析与挖掘服务，支撑科学家在干细胞前沿方向的知识发现工作，使平台成为干细胞领域科学研究强有力的科研助手。

致谢

中国科学院成都文献情报中心许海云、刘春江、张鑫、彭霖、陈文杰、赵爽、徐源以及中国科学院广州生物医药与健康研究院郭晨参加了干细胞领域知识发现大数据平台的建设工作，特此致谢！

参 考 文 献

[1]　周琪 . 体细胞重编程，挑战与希望 [C]. 中国细胞生物学学会 2013 年全国学术大会 • 武汉论文摘要集 , 2013:7-8.

[2] 张志强 , 范少萍 . 论学科信息学的兴起与发展 [J]. 情报学报 ,2015(10):1011-1023.

[3] 欧阳曙光 , 贺福初 . 生物信息学：生物实验数据和计算技术结合的新领域 [J]. 科学通报 ,1999,44(14):1457.

[4] 董建成 . 医学信息学的现状与未来 [J]. 中华医院管理杂志 ,2004,20(4):232-235.

[5] Williams A J, Harland L, Groth P, et al. Open phacts: semantic interoperability for drug discovery [J]. Drug Discovery Today, 2012, 17(21-22): 1188-1198.

[6] Nathan B, Jean C, Peter J, et al. Big Data in Drug Discovery [M]// Progress in Medicinal Chemistry, Amsterdam: Elsevier,2018, 57: 277-356.

[7] Kashyap V, Borgida A. Representing the UMLS® Semantic Network Using OWL[C]. The Semantic Web - ISWC 2003. Lecture Notes in Computer Science,2003.

[8] Rindflesch T C, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text [J]. Journal of Biomedical Informatics, 2003, 36(6):462-477.

[9] 庄严 , 李国良 , 冯建华 . 知识库实体对齐技术综述 [J]. 计算机研究与发展 , 2016, 53(1): 165-192.

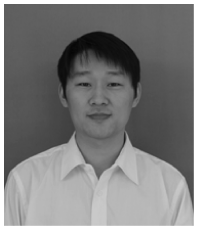
[10] 王颖 , 钱力 , 谢靖 , 等 . 科技大数据知识图谱构建模型与方法研究 [J]. 数据分析与知识发现 , 2019, 3(1) : 15-26.

[11] Wen Yi,Fu Hongguang, Shu Fang, et al. Amino acids industry knowledge service platform[J].Chinese Journal of Library and Information Science,2015,8(4):78-89.

作者简介

张志强，中国科学院成都文献情报中心主任，研究员、博士、中国科学院大学博士生导师，中国科学院特聘核心研究员。“新世纪百千万人才工程”国家级人选，四川省千人计划专家，四川省委省政府决策咨询委员会委员。独立和合作出版专著（编著）20 部、出版译著 13 部、发表论文 400 余篇。获得省部级科技进步奖、社会科学优秀成果奖等科技成果奖励 18 项。主要研究领域包括科技战略与规划、科技政策与管理、科学计量与评价、科学学、情报学理论方法与应用、生态经济学与可持续发展等。

胡正银，中国科学院成都文献情报中心副研究馆员、博士、中国科学院大学硕士生导师、中国科学院西部之光人才培养计划人选。合作出版专著（编著）1 部、发表论文 60 余篇、申请计算机软件著作权 6 项。从 2014 年开始，陆续担任 *Scientometrics*、*Technological Forecasting and Social Change*、*IEEE Transactions on Engineering Management* 和《图书情报工作》《数字图书馆论坛》等期刊审稿人。主要研究领域包括科技大数据、学科知识发现与知识挖掘、智能情报方法与技术。



杨宁，中国科学院成都文献情报中心副研究馆员。中国科学院西部之光人才培养计划人选，担任多个期刊审稿人，主持和参与国家和中国科学院项目 10 余项，在国内外重要核心期刊发表论文 20 余篇。主要研究领域包括数字图书馆方法与技术、情报理论方法与应用、科学大数据开发与应用等。

文奕，中国科学院成都文献情报中心研究员。在相关领域发表学术论文 50 余篇，作为项目负责人承担了中国科学院知识产权网等多个项目，研究成果获四川省科技进步三等奖。主要研究领域包括知识管理与知识计算。

覃筱楚，硕士，中国科学院广州生物医药与健康研究院信息情报中心工程师。参与国家、省、市级项目 20 余项，其中参与主持项目 5 项。参与发表论文 39 篇，其中 SCI 论文 28 篇。合作申请发明专利 3 件。主要研究领域包括生物医药产业情报研究、生物医药产业专利战略分析、再生医学高价值专利挖掘与培育。

宋亦兵，中国科学院广州生物医药与健康研究院信息情报中心主任，高级工程师。参与国家、广东省、市级项目 28 项，作为项目负责人主持项目 6 项，参与发表论文 12 篇。主要研究领域包括学科情报分析、产业情报分析。

潘光锦，中国科学院广州生物医药与健康研究院副院长，华南干细胞与再生医学研究所研究员、博士、博士生导师。中国科学院“百人计划”入选者，“广东特支计划”科技创新领军人才，国家重点研发计划首席科学家。迄今共发表第一或通讯作者论文 31 篇，其中以通讯（含共同）作者在 *Nature Methods*、*Nature Communications* 等杂志发表论文 23 篇。获得授权专利 3 项（含国际专利 1 项）；获批国家自然科学奖二等 2 项、广东省自然科学奖一等奖 1 项、中国科学院杰出科技成就奖（突出贡献者）等。主要研究领域包括高效获得具有潜在应用价值的功能性细胞、人多能干细胞命运转变调控。