National Science Library(Chengdu), Chinese Academy Of Sciences

Research and Practice on Mobile Service Models for Knowledge Discovery in the Stem Cell Domain Based on Knowledge Graph

Presenter： Yuan Xu

yuanx@clas.ac.cn

The Eleventh International Conference on Big Data, Small Data, Linked Data and Open Data (ALLDATA 2025)

Yuan Xu Engineer

National Science Library(Chengdu), Chinese Academy Of Sciences
Department of Data Resources.

Previously participated in the development of several important disciplinary knowledge discovery platforms；
Currently focusing on open science data governance and applied innovation research, aiming to explore feasible pathways for data value transformation through multidimensional practices.

1.Research Background

Stem Cell Knowledge Discovery Challenges

With the rise of data-intensive research paradigms, data-driven disciplinary knowledge discovery has become a significant characteristic and development direction in scientific research. Scientific activities increasingly rely on massive data.

However, the field of stem cell research currently faces several prominent challenges, including data overload, information fragmentation, insufficient intelligent analysis, knowledge gaps, and limited support for mobile scenarios.

2.Research Objectives

To address these critical issues, the study establishes research objectives:

This research by deeply integrating stem cell biology with information science and technology, constructing a knowledge graph-based mobile knowledge discovery service model for ubiquitous environments, establishing a complete "data-knowledge-value" transformation chain, and systematically uncovering the scientific principles and knowledge associations embedded in research data - thereby facilitating a paradigm shift from data-driven to knowledge-driven methodologies in stem cell studies.

Three levels of research tasks

- Develop knowledge graph
- Build a model
- Develop mobile tools

3.Project Introduction

This study, supported by the Chinese Academy of Sciences' "13th Five-Year Plan" research informatization project titled "Research Informatization Applications for Knowledge Discovery in the Stem Cell Field," addresses the demand for research informatization in stem cell knowledge discovery.

Leveraging the scientific and technological information data resources of the National Science Library (Chengdu), Chinese Academy of Sciences, and the research strengths in the stem cell field of the Guangzhou Institute of Biomedicine and Health, Chinese Academy of Sciences, **a stem cell knowledge graph with "multi-form, multi-granularity, and multi-dimensional" data integration was constructed.**

3.Project Introduction

Based on this, the research team developed the "Stem Cell Knowledge Discovery Platform" and expanded the research on mobile service systems. The team also created the WeChat Mini Program "Stem Cell Assistant," migrating the core functionalities and key services of the "Stem Cell Knowledge Discovery Platform" to the Mobile terminal, thereby achieving the mobile application of the stem cell knowledge graph.

Stem Cell Knowledge Discovery Platform

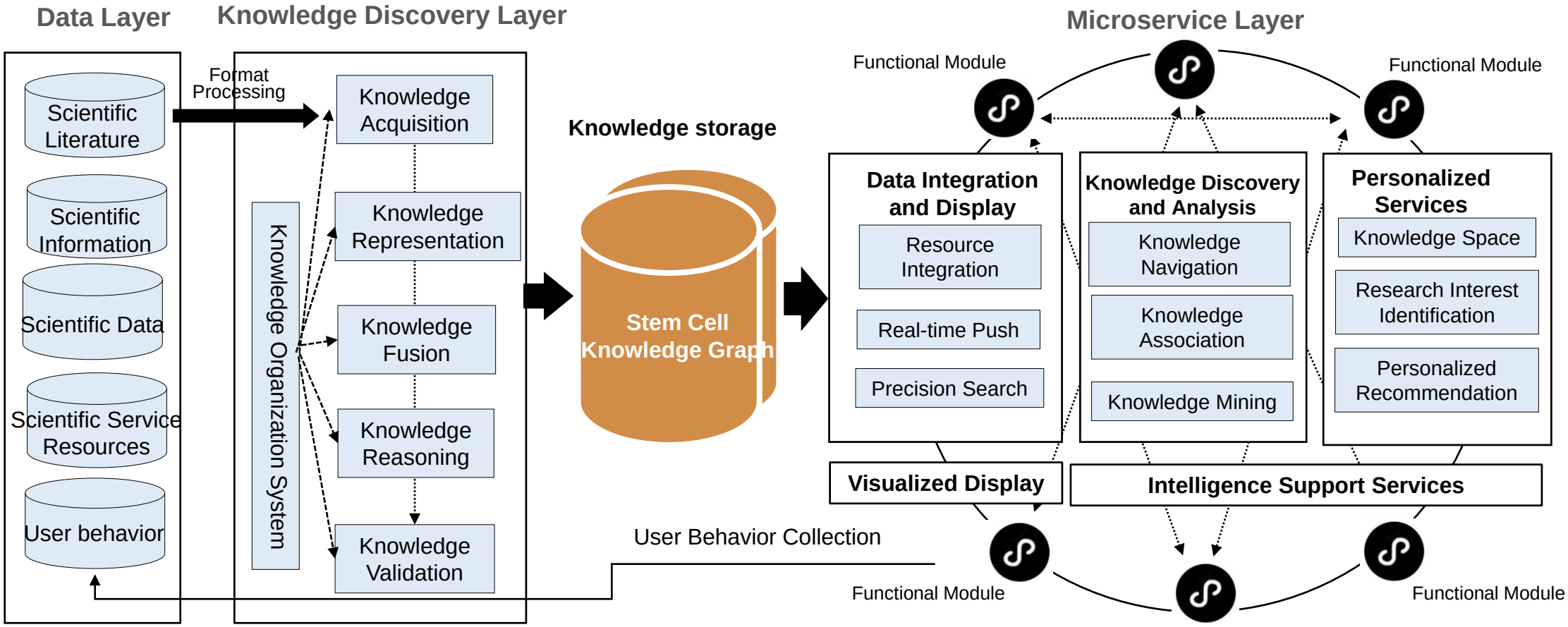
Stem Cell Knowledge Graph

"Stem cell assistant" wechat mini program

4.Research Content

Mobile Service Models for Knowledge Discovery in the Stem Cell Domain Based on Knowledge Graph

The model, driven by the stem cell knowledge graph, constructs a complete service process of aggregating massive scientific data, conducting in-depth knowledge discovery, and providing mobile services.



4.Research Content

Mobile Service Models for Knowledge Discovery in the Stem Cell Domain Based on Knowledge Graph

Data Layer

Based on distributed data collection and governance technologies, this layer integrates and standardizes multi-source, heterogeneous scientific data in the stem cell field, providing a high-quality data foundation for knowledge graph construction.

Knowledge Discovery Layer

Using NLP techniques like entity recognition, relation extraction, and semantic reasoning, this layer creates complex links between stem cell entities. It converts large data volumes into a structured knowledge network and builds a stem cell knowledge graph.

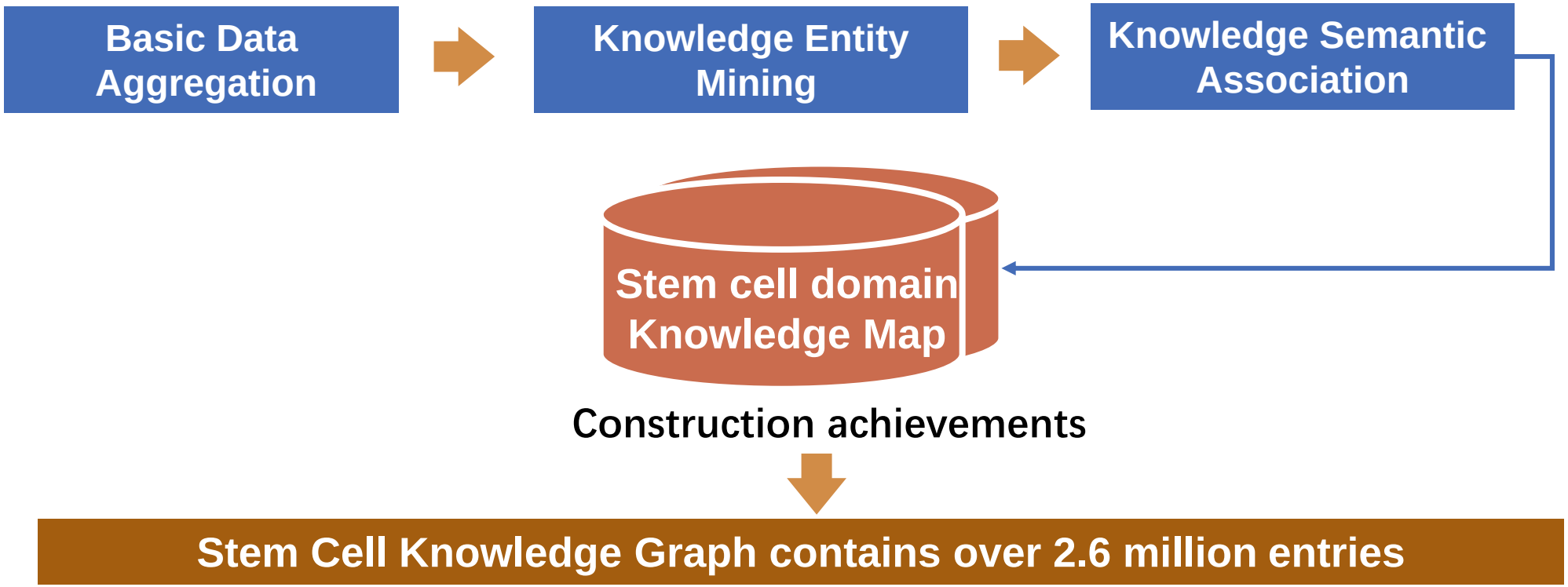
Microservice Layer

Microservices architecture encapsulates knowledge discovery functions, containerized deployment builds a scalable system, and WeChat Mini Program efficiently transmits stem cell knowledge graph data.

4.Research Content

Stem Cell Knowledge Graph

The project constructs a stem cell knowledge graph in the field of stem cells from three levels: foundational data aggregation, knowledge connotation mining, and knowledge semantic association.



4.Research Content

— Basic data aggregation—

The foundational data layer efficiently integrates 16 categories of basic scientific and technological data in the field of stem cells, including research papers, patents, funded projects, clinical trials, product regulations, experts, and institutions. It incorporates 103 authoritative core data sources such as PubMed, Cortellis, and StemCell Commons. To address the complexity of multi-source heterogeneous data in the stem cell field in terms of format, semantics, and structure, a standardized data processing workflow is established to achieve in-depth data governance.

16 Types of Scientific and Technological Data

Scientific Literature	Papers, patents, reports, journals, monographs, etc.;
Scientific Information	News, R&D dynamics, experts, projects, policies and regulations, etc.;
Scientific Data	Clinical trials, pharmaceutical products, scientific experiments, etc.;
Scientific Service Resources	Scientific instruments, experimental animals, experimental reagents, etc.

- **Unified Data Format**
- **Data cleaning and filtering**
- **Word segmentation and part-of-speech tagging**
- **Big Data storage**

4.Research Content

—Basic data aggregation—

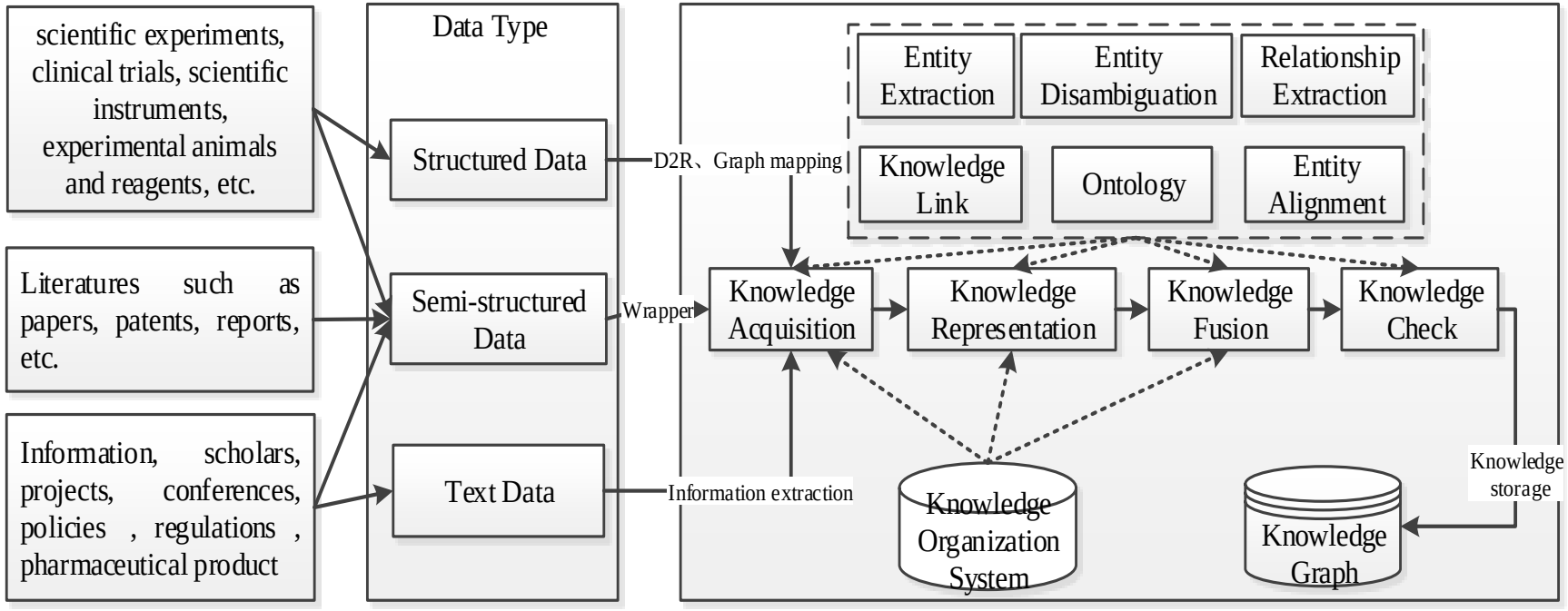
103 Authoritative Core Data Sources (Partial)

Data Types	Data content
Scientific and Technological Data	StemCell Commons, Stem Cell Standardisation, HMDB, StemCellDB, CSCdb, SHOGoiN细胞数据库, BioPortal等
Fund Projects	美国联邦项目库（Federal Reporter），包括美国NIH、NSF等2011年度起的资助项目数据（干细胞）
	欧盟 Horizon 2020的资助项目数据（干细胞）
	英国RCUK 2011年度起的资助项目数据（干细胞）
	国家自然科学基金 2011年度起的资助项目数据（干细胞）
	国家重点研发计划重点专项(生物医学类) 2016年起（干细胞）
Research Papers	PubMed干细胞论文、ESI高被引论文、Nature Index期刊论文数据
Clinical Trials	NIH Clincialtrials.gov、WHO clinical trials干细胞临床试验数据
Patents	中国国家知识产权局年度授权的有效发明专利（干细胞）
	全球其他国家或地区发明专利（干细胞）
Products	CFDA中国新药研发监测数据库（CPM）、Cortellis数据库干细胞药物
	美国FDA、欧盟EMA、日本PMDA批准上市干细胞药物数据
Scientific and Technological Talent	干细胞科技人才数据库，杰青、优青、长江学者、千人计划、高被引科学家（爱思唯尔）、全球top 50影响力干细胞科学家

Knowledge Connotation Mining and Semantic Association

The project employs technologies such as entity extraction, entity disambiguation, relation extraction, knowledge linking, ontology, and entity alignment to extract data objects, knowledge objects, and semantic associations from high-quality data sources in the stem cell field, including scientific literature, scientific data, scientific activities, and industrial information. These elements are formed into knowledge triples, establishing data models such as "concept-entity-attribute-relationship," thereby constructing a domain-specific knowledge graph.

Construction process of cell knowledge map



Knowledge Connotation Mining

By referencing domain knowledge ontologies and knowledge organization systems such as the UMLS Metathesaurus, DisGeNET, and StemCell Commons (including thesauri, classification systems, and semantic dictionaries), the process involves text mining of core scientific literature from perspectives of interest to researchers, such as scientific instruments, animal models, experimental techniques, cellular organs, and disease genes. This mining yields knowledge entities and the relationships between them, thereby forming the knowledge organization system for the stem cell knowledge graph.

Knowledge Semantic Association

By integrating scientometric indicators with text mining technologies, semantic associations among various types of scientific information and knowledge entities within the knowledge graph are established based on relationships such as citations, acknowledgments, collaboration networks, and co-occurrence of knowledge entities. The project utilizes SemRep to extract semantic relationships between knowledge entities from stem cell research data, presenting these semantic associations in the form of an SPO semantic network.

SPO Triple Semantic Association Example [Functional Relation]

Subject: Stem Cell Therapy for Cardiac Regeneration
Predicate: TREATS
Object: Heart Failure

_____

The complete SPO triple is represented as: (Stem Cell Therapy for Cardiac Regeneration, TREATS, Heart Failure)
This triple indicates the treatment relationship between "Stem Cell Therapy for Cardiac Regeneration" and "Heart Failure".

4.Research Content

Stem Cell Knowledge Graph

The project integrates big data in scientific research, which includes scientific literature, technological information, scientific data, and resources for scientific and technological services, all dispersed across various data sources.

Data Aggregation

16 types of scientific data,
research data over **400,000 entries**

Basic Data Statistics Table

Category	Quantity
Scientific Papers	302,542
Patents	92,354
Research Reports	1,025
Core Journals	26
Classic Monographs	196
News and Information	612
Experts	235
Research Institutions	48
Projects	17,528
Policies and Regulations	133
Clinical Trials	4,806
Medical Products	1,211
Scientific Experiments	137
Open Scientific Data	>200 million
Scientific Instruments	152
Laboratory Animals	68
Laboratory Reagents	1,021

4.Research Content

Stem Cell Knowledge Graph

The project centers on the SPO semantic network of stem cell scientific literature and, from eight perspectives including scientific instruments,laboratory animals, experimental protocols, experimental reagents, methodological techniques, cells, organs, diseases, and genes, utilizes knowledge extraction technology to mine over 20,000 entities from domain-specific research data.

Entity Mining

8 perspectives, 21,142 knowledge entities

Knowledge Entity Data Table

Entity Type	Quantity	Example
Scientific Instruments	8009	Protein purification system，peptide synthesizer，Pick level nucleic acid detection system，etc
Experimental Animals	159	mouse，rat，drosophila，zebrafish，rabbit，pig，monkey，etc
Experimental Protocol	68	DNA Extraction Protocols，Protein Purification Protocols，mRNA Protocols，etc
Experimental Reagents	14041	Transcription Factors，Biomarkers，RNA Messenger，DNA-Binding Proteins，etc
Gene	2628	Runx2 gene，CD31 gene，SOX9 gene，TNF-alpha gene，ARF gene,etc
Disease	689	Leukemia，Vascular Diseases，Sclerosis，Diabetes Mellitus，etc
Cell Organs	274	progenitor cell，neuron，liver embryonic stem cell，neural stem cell，etc
Methods and Techniques	2588	cell culture，bone marrow transplantation，stem cell therapy，flow cytometry，etc
Total	21142	

Semantic Relationships

58 types, 733 semantic associations, 2,224,733 knowledge associations

By integrating scientometric indicators with text mining technologies, the project establishes semantic associations among various types of scientific information and knowledge entities within the knowledge graph, based on relationships such as citations, acknowledgments, collaboration networks, and co-occurrence of knowledge entities.

It defines a total of 58 types of semantic relationships across three major categories: "document-document, document-knowledge entity, and knowledge entity-knowledge entity," with the core focus being the semantic associations between "knowledge entity-knowledge entity."

Semantic Relationships

Semantic Object	Semantic Group	Semantic Relationship
Knowledge entities – Knowledge entities	Interaction relation	ASSOCIATED_WITH; INTERACT_WITH; PRODUCE; CONVERT_TO; COEXIST_WITH
	Functional relation	TREAT; PREVENT; USE; METHOD_OF; ADMINISTERED_TO
	Representation relation	DIAGNOSE; MANIFESTATION_OF
	Comparative relation	HIGH_THAN; LOWER_THAN; SAME_AS; ISA
	Position relation	PART_OF; LOCATION_OF; OCCUR_IN
	Influence relationship	AFFECT; PROMOTE; DISRUPT; CAUSE; INHITBIT; PREDISPOSE; AUGMENT; PRECEDE; PROCESS_OF; COMPLICATE; STIMULATE;
	Co-occurrence relation	co-occurrence
	Hierarchical relation	subclass_of
Knowledge entities – Literatures	Affiliation relationship	belong_to_PMID (Associated Paper) belong_to_PN (Associated Patent) belong_to_CT (Associated Clinical Trial)
Literatures – Literatures	Co-occurrence relation	co-author co-topics co-keywords

5. Mobile Service Case

"Stem Cell Assistant" WeChat Mini Program

"The 'Stem Cell Assistant' WeChat Mini Program is built based on the WeChat Developer Tools, utilizing the lightweight Flask framework as its core architecture.

Guided by user needs, the app decouples functionalities into multiple independent microservice modules. Each microservice module provides services through standardized RESTful API interfaces, enabling seamless cross-platform and cross-device integration and invocation of stem cell knowledge graph research data and functionalities."

5. Mobile Service Case

"Stem Cell Assistant" WeChat Mini Program

5.1 Diversified Data Display Services

Integrated display according to knowledge prism

Scientific Instruments | Experimental Animals | Experimental Methods | Experimental Reagents l Cell Organs | Techniques and Methods | Immunotherapy | Genetic Diseases

Integrated display by resource category

Academic Papers | Patents | Research Funding Projects | Regulatory Affairs | Clinical Trials | Master's and Doctoral Dissertations | Information Reports | Experimental Data

Data Correlation Display

Literature - Literature
Literature - Knowledge Entity
Knowledge Entity - Knowledge Entity

Real-time Data Push Services

Time-based Push | Source-based Push

5. Mobile Service Case

5.2 Data classification and screening

Data classification statistics

Accurate data screening

Screen by Resource Type
Screen via Knowledge Prism

Knowledge prism quick annotation

Visual analysis entry

5.3 Efficient and Intelligent Retrieval Service

The "Stem Cell Assistant" WeChat Mini Program has built a comprehensive and efficient search system, including:

One-stop Search

Achieves comprehensive querying of 16 heterogeneous data sources.

Composite Search

Combines resource types (such as papers, patents) with multidimensional semantic annotations of the knowledge prism (such as methods, experimental materials, etc.) to achieve precise and intelligent search of heterogeneous data in the stem cell field.

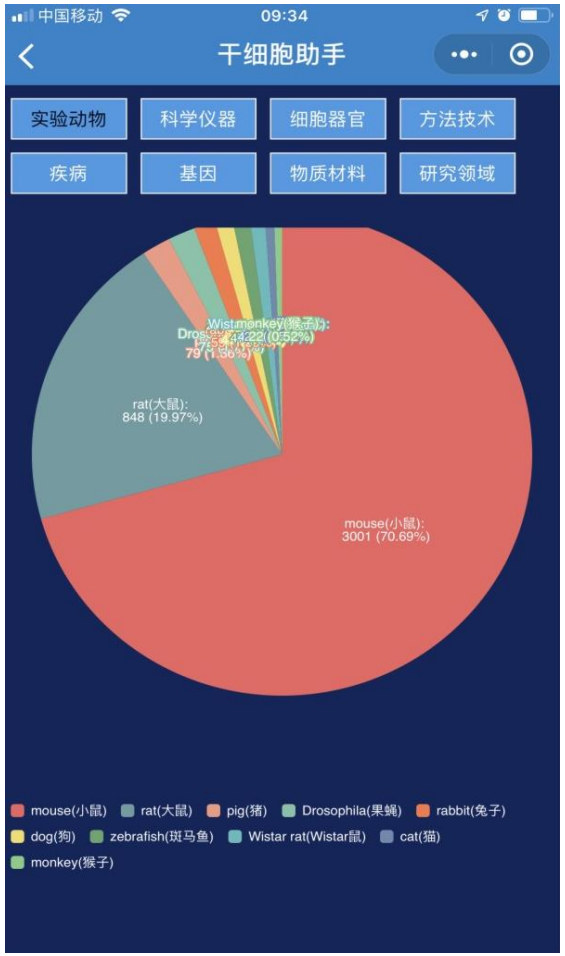
Resource Classification Search

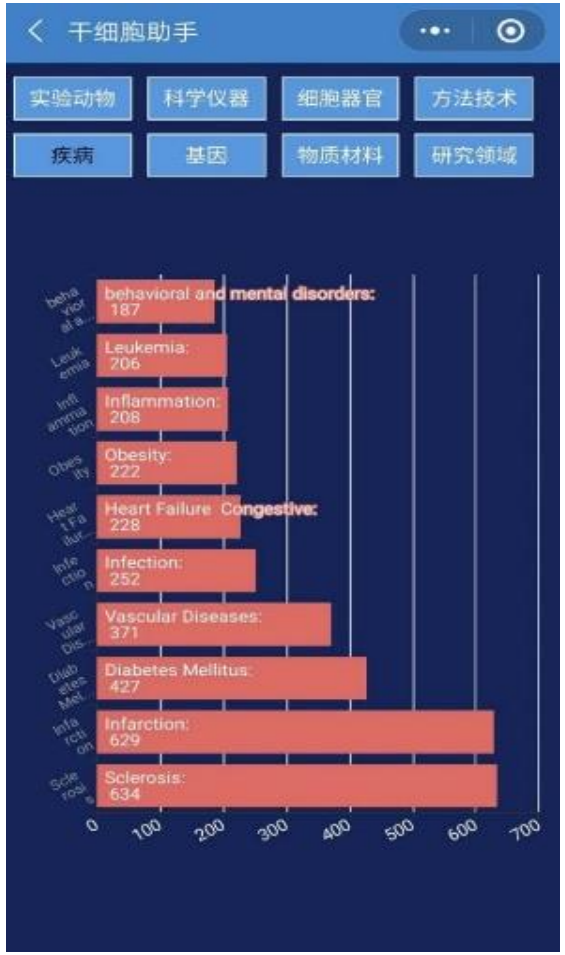
Conducts systematic classification and hierarchical organization of heterogeneous resources in the stem cell field through resource type classification search.

5. Mobile Service Case

5.4 Data Visualization

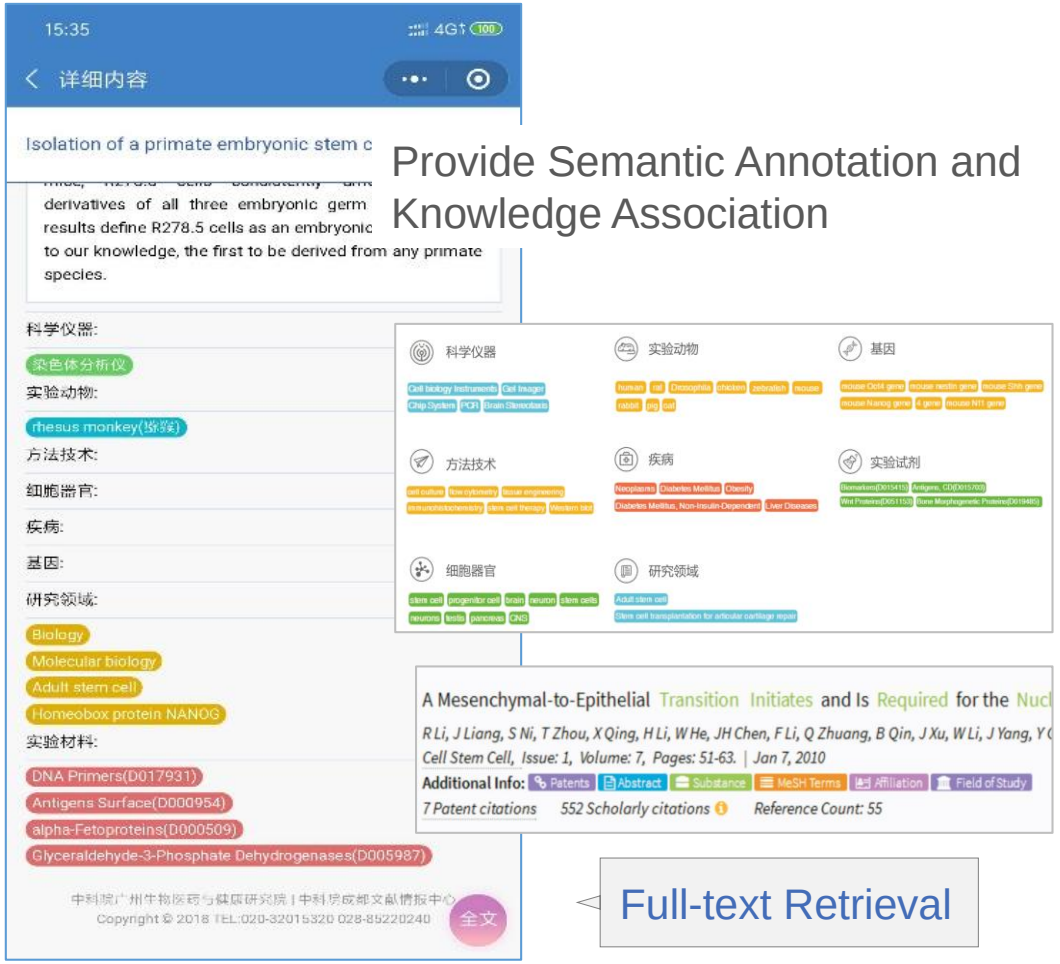
Combined with the data analysis function of stem cell knowledge map, it supports scientific research users to view data visual analysis charts, such as bar charts, pie charts, etc.





5.5 Deep Knowledge Discovery Service

Based on knowledge graphs and citation relationships, achieve knowledge association between papers, patents, etc., revealing stem cell scientific information from multiple perspectives and with fine granularity.



Provide Semantic Annotation and Knowledge Association

Full-text Retrieval

5. Mobile Service Case

5.6 Personalized service based on user behavior characteristics

Personal data management space

Through the systematic integration of user history collection and sharing of data, the formation of personal data management space.

Intelligent recommendation service

Collaborative filtering recommendation algorithm is adopted to find the set of users with high similarity of interests and preferences, and carry out accurate resource push.

Research interest identification

Cluster analysis was carried out on user behavior data to generate 10 scientific research labels.

Summary and Limitations

This study has systematically integrated stem cell biology with information science and technology to innovatively construct a ubiquitous environment-oriented mobile knowledge discovery service system, successfully achieving a paradigm shift in stem cell research from data-driven to knowledge-driven approaches. These research outcomes have not only provided a new generation of knowledge service tools for stem cell research, but also established a complete "data-knowledge-service" transformation chain, offering a replicable technical paradigm for knowledge discovery in the biomedical field.

Platform Dependency Limitations: Due to the API interface limitations of the WeChat Mini Program ecosystem, some advanced knowledge service functions (such as complex graph reasoning and real-time collaborative annotation) cannot be fully implemented.

Validation Limitations: The current research scope of the service is limited, and whether the knowledge service application scenarios are fully matched still needs further validation.

Prospects

The integration of large language models with knowledge discovery mobile services will transform knowledge engineering. With large model technology, we combine knowledge graphs, AI, and mobile services to shift knowledge services from passive retrieval to active recommendation, from fixed terminals to ubiquitous access, and from single information to contextual intelligence. This creates a new model of “anytime, anywhere, on-demand” knowledge services. It accelerates precise, intelligent, and ubiquitous knowledge services, offering innovative support and a reference for knowledge service systems in stem cell research and other fields. As key technologies advance, this model is expected to become a vital infrastructure for scientific innovation and knowledge dissemination, driving disciplines into a new intelligent era.

Thank you for your attention!

Research and Case Team

National Science Library (Chengdu), Chinese Academy of Sciences

Guangzhou Institute of Biomedicine and Health, Chinese Academy of Sciences

South China University of Technology

Microservices General Planning:

Zhengyin Hu, Yuan Xu, Yibing Song, Qiong Chen, Yi Wen, Xuan Wu

Knowledge Graph Construction:

Zhengyin Hu, Chunjiang Liu, Wenjie Chen, Ning Yang, Xin Zhang

Microservices Technical Implementation:

Yuan Xu, Tianjia Yao, Xiaohong Yang