ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202306.00099

Research on Research Data Services Under the Trend of Intelligent Research

Authors: Zhang Jingrui, Sun Mengge, Han Tao, Han Tao

Date: 2023-06-12T00:00:00+00:00

Abstract

Purpose To systematically review and summarize the operational workflow of scientific research data within the research process under the trend of intelligentized scientific research, uncover the underlying demands for research data, and provide insights for the transformative development of research data services in this new trend. Methods Guided by the theory of the research data lifecycle and taking the fields of materials and chemistry as examples, this study analyzes the process by which research data is transformed into knowledge in intelligentized scientific research. It constructs an operational workflow for the research data lifecycle comprising six stages: data management planning, data generation and collection, data processing and analysis, data production and publication, data storage and sharing, and data reuse, thereby uncovering the roles and potential demands of research data. Results Intelligentized scientific research demonstrates demand characteristics for multi-source heterogeneous data integration, fine-grained data structuring, exploration of human-computer interactive language representation, relational data mining, and enrichment of research data types. Conclusion It is recommended that future development of research data services should enhance the construction of high-quality comprehensive domain data networks, deepen embedded research-oriented data services, improve librarians' domain knowledge and artificial intelligence literacy, emphasize the mining of experimental information from textual data, and focus on the exploration of human-computer interactive languages.

Full Text

Research on Scientific Research Data Services in the Era of Intelligent Scientific Research

Jingrui Zhang[1],[2], **Mengge Sun**[1],[2], **Tao Han**[1]
[1] National Science Library, Chinese Academy of Sciences, Beijing 100190, China

[2] Department of Library, Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190, China

Abstract:
Objective: This study systematically examines and summarizes the operational workflow of scientific research data within the research process under the trend of intelligent scientific research, identifies underlying data demands, and provides insights for the transformation and development of research data services in this new paradigm. **Methods:** Guided by the theory of the research data lifecycle and using materials and chemistry as exemplary fields, we analyze how research data transforms into knowledge in intelligent research contexts. We construct a six-stage research data lifecycle framework encompassing data management planning, data generation and collection, data processing and analysis, data production and publication, data storage and sharing, and data reuse, to explore the role and latent needs of research data. **Results:** Intelligent scientific research demonstrates distinct demand characteristics for multi-source heterogeneous data integration, fine-grained data structuring, exploration of human-machine interaction language representations, data association mining, and enrichment of research data types. **Conclusions:** We recommend that future research data services should: accelerate the construction of high-quality, comprehensive domain-specific data networks; deepen embedded research data services; enhance librarians' domain knowledge and AI literacy; prioritize mining experimental information from textual data; and focus on exploring human-machine interaction languages.

Keywords: Intelligent Scientific Research; Research Data; Research Data Services

---

In recent years, the application of mature artificial intelligence (AI) technologies to challenging fundamental scientific research has dramatically improved research efficiency, sparking a transformative movement in scientific inquiry. The primary agent of knowledge discovery has shifted from human researchers to intelligent scientists, while the object of study has transitioned from traditional experimental subjects to research data.

Research data refers to all types of data generated throughout the entire scientific research process across various disciplines [**?**], including published outcomes such as scientific literature and patents, as well as scientific data produced from basic research, applied research, experimental development, and observational testing. Currently, research data is presented in forms understandable and learnable by humans and stored digitally. However, AI technologies cannot accurately extract the implicit scientific laws within this data; only domain-constrained research data can be effectively utilized for AI learning.

Under the data-intensive research paradigm, research data is characterized by rapid growth, massive scale, and diverse sources and formats, posing challenges

for data management, preservation, integration, and sharing. Consequently, development efforts have emphasized the digital transformation of research paradigms, proposing the construction of open and shared scientific databases or platforms and establishing explicit associations between data to enable discovery, access, integration, and analysis of research data, thereby fostering relationships between previously unrelated domains and promoting knowledge discovery [**?**]. In response, nations worldwide have continuously invested in strategies centered on research data integration and sharing. For example, the European Union' s Seventh Framework Programme (FP7, 2007-2013) launched the Global Research Data Infrastructures (GRDI 2020) project, incorporating research data integration and sharing into its research agenda [**?**].

The accumulation of massive research data and deep integration of AI technologies in scientific research have given rise to the intelligent scientific research paradigm. Under this new paradigm, research data exhibits complex characteristics of multi-source heterogeneity, multi-dimensionality, and interrelatedness. However, the shared and integrated research data from the data-intensive paradigm suffers from dispersion, significant domain-specific quantity disparities, uneven quality, and inconsistent standards and formats, challenging the readability, usability, and comprehensibility of AI models. Addressing these needs, the United States has deployed relevant policy mechanisms to maintain leadership in research data under this new trend. For instance, in May 2020, the NIH Common Fund launched the "Bridge2AI" program, aiming to generate machine-understandable, unified, and standardized biomedical and behavioral datasets and develop automated tools to accelerate standardized dataset creation [**?**]. In 2021, the U.S. Materials Genome Initiative established strategic goals for intelligent research trends, including unified and standardized materials data infrastructure, promotion of open materials data sharing, and unification of metadata standards to fully leverage the power of materials data in AI research and development [**?**].

National research data policies and programmatic deployments for this paradigm shift provide data service guarantees for the development of intelligent scientific research. In practice, intelligent research data services comprise domain researchers, data scientists, and information service personnel. Domain researchers and data scientists continuously explore and participate in research data work under intelligent research trends from the perspectives of research needs and cutting-edge AI technologies, constructing high-quality, fine-grained, multimodal research databases to meet AI model data requirements. In contrast, information service personnel remain focused on traditional literature metadata organization, consultation, and training services for emerging technologies. Although preliminary explorations into fine-grained data extraction and knowledge service transformation have begun, their engagement and exploration in intelligent research data services lag behind domain researchers and data scientists, failing to highlight the advantages of library and information institutions in combining data and technology.

In summary, intelligent scientific research has recognized the importance of research data and implemented relevant policies and programs. Research data undergoes a lifecycle process in scientific research, yet existing studies lack exploration of the operational workflow, role, and demands of the research data lifecycle under new trends. Therefore, this paper examines intelligent research practices to summarize the lifecycle operational process, role, and latent needs of research data in scientific research from a lifecycle perspective, and subsequently provides insights for developing research data services in library and information institutions under this new trend.

2 Analysis of the Research Data Lifecycle Operational Process in the Era of Intelligent Scientific Research

The operational workflow of research data in scientific research is a goal-driven DIKW (Data-Information-Knowledge-Wisdom) model, representing the transformation of research data into information, knowledge, and wisdom under research objectives to support goal achievement. This section uses frontier fields in intelligent research—materials and chemistry—to explore the research data lifecycle operational process, focusing on how data transforms into ultimate wisdom (Figure 1

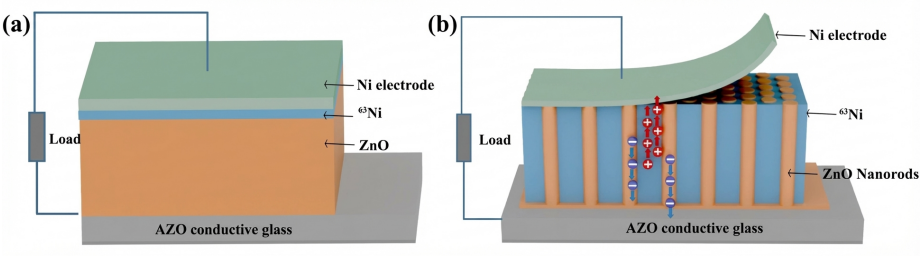

Figure 1: Figure 1

). This framework is constructed from a practical research perspective, with detailed elaboration and analysis of each stage and process based on data lifecycle theory.

2.1 Data Management Plan (DMP)

The determination of a research topic establishes a "navigation tower" for the project. Following topic selection, a plan to assist in research data management must be developed. A Data Management Plan (DMP) is a dynamic document describing project data collection, documentation, management, and publication throughout the research project lifecycle, including techniques, methods, and policies for creation, documentation, access, storage, and sharing [**?**]. DMPs play a critical role in managing and reviewing research data and outcomes. Against the backdrop of open science, research funding agencies, as the main drivers

of open access, have implemented policies requiring DMP submission during project applications to promote research data sharing [**?**].

Comparing DMPs from domestic and international funding agencies reveals that foreign agencies provide corresponding tools and templates, specifically deconstructing and explaining DMP contents for funded projects, with emphasis on post-storage sharing and access, security and ethics, and DMP cost management. Domestic DMPs, primarily led by the Ministry of Science and Technology and the Chinese Academy of Sciences, focus on data submission and management for funded programs or projects, clarifying responsibilities of data-related entities and submission pathways, but lack attention to subsequent sharing and access and cost management, without providing corresponding DMP tools and templates.

2.2 Data Generation and Collection

The data generation and collection stage is the foundational component supporting scientific research. Researchers select appropriate data types and acquisition pathways based on research objectives and data availability to obtain required research data content, providing "fuel" for subsequent scientific research (Figure 2 [FIGURE:2]).

(1) Data Production Pathways
Data production is a crucial initial stage of scientific research, accumulating substantial usable data for intelligent research. Current data production pathways primarily include four categories: scientific literature, patents, experimental data, and computational data. Scientific literature and patent text data extract metadata, entities and attribute values, tables, and experimental paragraphs through manual, semi-automated, and automated methods. Experimental data is collected and generated by researchers during observation, experimentation, and investigation processes or digitally recorded using Electronic Laboratory Notebook (ELN) tools. Computational data is simulated or predicted data generated by researchers using high-throughput screening, computational platforms, or AI model tools.

(2) Data Acquisition Pathways
Data acquisition pathways involve researchers selecting suitable dataset construction approaches based on research topics, data quality, data structure, and data accessibility characteristics. Current pathways primarily include journal publishers, patent databases, scientific databases, open data platforms, and manually constructed datasets. From the perspective of data sources and openness, commercial journal publishers are the preferred choice for researchers due to their broad literature data collection scope and high structural quality, with commonly used databases including Wiley, Elsevier, and Scopus, and specialized databases for materials and chemistry such as ACS, the American Chemical Society, and the Royal Society of Chemistry. Open data platforms support intelligent research due to their data accessibility and large accumulation, with

typical databases including the Arxiv literature data platform and general data sharing platforms like Figshare and Zenodo.

Overall, intelligent research centers on text data, numerical data, and image data, acquiring required data from commercial and open-source platforms. The demand for high-quality data in intelligent research has shifted researchers toward journal publishers, though this presents the disadvantage of high database costs, which is unfriendly for small research institutions or teams to access data. Open-source acquisition pathways are favored by most researchers due to data accessibility but suffer from inconsistent data structures and low data quality issues.

(3) Data Collection Methods and Self-Generating Integration
Data collection methods involve researchers using certain technologies to batch collect required data after determining acquisition pathways. Intelligent research requires large volumes of data, necessitating batch crawling or downloading using collection software, crawler code, or API interfaces provided by data platforms to improve collection efficiency.

Collection software comprises pre-packaged crawler platforms belonging to "foolproof" collection modes, usable for web and literature data collection, such as Octoparse, Zotero, and SciHub Pro. Existing crawler code is written based on Python frameworks, with typical toolkits including Scrapy, BeautifulSoup, Requests-HTML, and Selenium. Self-generating integration involves storing experimental material composition, ratios, and results data into databases or tables after robot chemists or material scientists complete experiments for AI model learning.

2.3 Data Processing and Analysis

Data processing and analysis constitute the core of intelligent research, aiming to ensure consistent input data quality and structure and transform it into machine-understandable forms to establish associations between research objectives and data. Based on data types and research purposes, data processing and analysis in materials and chemistry are divided into three major modes: text-data-centric, experimental-material-centric, and numerical-data-centric.

(1) Text-Data-Centric Processing and Analysis Mode
The text-data-centric processing and analysis mode extracts key experimental elements and conditions from scientific literature and patent texts, combining and transforming them into machine-readable forms. This primarily includes six processing stages: experimental paragraph (sentence) identification, data annotation, entity and relation extraction, experimental text enhancement, data representation, and association mining/data generation.

Scientific literature and patents contain substantial redundant text information unrelated to experimental synthesis, increasing the difficulty of extracting synthesis information. Therefore, it is necessary to identify paragraphs (sen-

tences) highly relevant to experimental synthesis to narrow the text extraction space. Identification methods include rule-based approaches and machine learning/deep learning approaches. Rule-based methods require researchers to pre-understand key characteristics of experimental paragraphs (sentences) to construct identification rules, such as domain-specific synthesis substances, property identifiers, and specified values. Simple rule construction is represented by keyword matching [**?**], while complex rule construction is represented by pattern matching [**?**] and regular expressions [**?**]. These methods are easy to understand and interpret, allowing researchers to quickly experiment and modify them, but their flexible identification capability is poor when experimental paragraphs contain numerous variables or constraints. Compared to rule-based methods, machine learning/deep learning approaches possess strong autonomous learning and flexible adaptation capabilities, classifying by learning features of experimental paragraphs (sentences). These can be divided into traditional machine learning-based methods and deep learning-based methods.

Traditional machine learning-based methods for experimental paragraph (sentence) identification center on classification approaches, requiring small amounts of manually annotated feature data, including logistic regression for high-dimensional binary classification [**?**], random forest for low-dimensional multi-classification [**?**], and Bayesian methods for high-dimensional multi-classification [**?**]. Deep learning-based methods stand out for their learning speed and accuracy, requiring large training datasets to autonomously learn experimental paragraph features, though they involve complex parameter tuning and poor model interpretability, including RNN models suitable for capturing long-sequence semantic relationships [**?**] and BERT models suitable for self-supervised parallel computation [**?**].

Annotated data forms the foundation for high-performance model learning, helping models understand contextual information. Training label data for text data extraction is scarce, requiring experimental paragraph annotation before extraction to build training data needed for extraction models. With AI model development emphasizing data annotation, specialized data annotation teams and platforms have emerged. Current experimental text annotation primarily uses crowdsourcing and manual annotation, with well-known crowdsourcing platforms such as Amazon Mechanical Turk (MTurk) abroad to avoid low team annotation efficiency, while manual annotation by domain experts suits small datasets.

Entity and relation extraction identifies entities such as products, reactants, and solvents, along with corresponding experimental conditions and relationships between entity-entity and entity-experimental condition from experimental paragraphs. Methods include integrated extraction tools and deep learning-based approaches. Integrated extraction tools include OSCAR4 [**?**], ChemicalTagger [**?**], and ChemDataExtractor [**?**]. OSCAR4 is suitable for chemical entity identification, using pure text as input and integrating regular expressions, dictionaries, and Maximum Entropy Markov Models (MEMM) to identify chemical entities.

ChemicalTagger focuses on patent experimental sections, using text strings as input, integrating OCSAR tools and regular expression-based rule methods to identify chemical entities, and combining syntax-based phrase parsers to identify relationships between operation phrases and entities to generate structured reaction pathway graphs. This tool relies on manually constructed rules, is sensitive to noise introduced by language use or preprocessing, and has poor scalability on non-patent data such as scientific literature. ChemDataExtractor is an end-to-end text mining toolkit that uses integrated Conditional Random Fields (CRF), rule-based phrase parsers, and table parsers to extract chemical entities, properties, measurements, and procedures from scientific literature text and tables to construct datasets. Deep learning-based methods include sequence-based extraction methods—Bi-LSTM-CRF [**?**] and pre-trained language model-based extraction methods—BERT-CRF [**?**]. The sequence-based Bi-LSTM-CRF method uses Bi-LSTM to capture contextual semantic relationships between words in long text sentences, combined with CRF models to predict optimal label chains for input sentences. The pre-trained language model-based BERT-CRF method uses language models to effectively obtain contextual information in text and performs supervised fine-tuning through extraction tasks, combined with CRF models for extraction.

In situations lacking training data, it is necessary to expand existing data volumes by adding negative sample data to meet model learning requirements and improve model generalization, avoiding underfitting or overfitting. Text data enhancement involves "transforming" experimental synthesis sequences by replacing entities and attributes such as compound names, quantities, time, temperature, and volume to augment experimental synthesis data.

Data representation serves as the connection point between data and algorithmic models, transforming experimental synthesis data into machine-understandable forms. Data representation methods primarily include Word2Vec, ELMo, and BERT. Word2Vec is commonly used in intelligent research due to its low-dimensional semantic vector representation characteristics, though its generated word vectors are static representations that cannot resolve synonym issues. The ELMo model uses a double-layer bidirectional LSTM to capture contextual information for encoding, representing a dynamic representation, but due to LSTM' s inherent long-distance dependency issues, it cannot capture long sequences and lacks parallel processing capabilities. The BERT model is a deep bidirectional Transformer-based pre-trained model that uses word context information for representation, representing a dynamic representation that resolves polysemy issues. This model possesses powerful parallel computing and transfer learning capabilities and is commonly used for domain model pre-training.

Finally, based on constructed standard experimental datasets, robots automatically synthesize new data or AI algorithms learn associations between experimental substances and attributes and experimental condition parameters to serve synthesis prediction. Data association methods primarily center on prop-

erty prediction, including traditional machine learning methods such as classification and regression, and deep learning methods for data generation. Traditional classification and regression methods include Support Vector Machines (SVM) for discrete property prediction [**?**], random forest regression models [**?**], and Gaussian process regression for continuous property prediction [**?**]. Deep learning methods include Variational Autoencoders for learning joint probability distributions of experimental synthesis conditions and targets [**?**]. Traditional machine learning methods suit situations with sufficient manual feature extraction, limited data, and computational resources, offering intuitive and easy-to-implement models. For example, researchers used random forest models to model zeolite material structural characteristics based on a given set of zeolite synthesis parameter data (including numerical types related to compositional elements—values, ranges, or variables; synthesis operation actions and conditions) [**?**]. Deep learning methods directly mine implicit structures and associations from original high-dimensional data, breaking through limitations of manual feature extraction, but involve high resource consumption, complex structures, and poor interpretability. For example, researchers used variational autoencoders to compress synthesis parameters into low-dimensional representations to establish probability distribution relationships between synthesis conditions and precursor materials [**?**].

(2) Experimental-Material-Centric Processing and Analysis Mode
The experimental-material-centric processing and analysis mode uses robotic/automated experimental devices to conduct combinatorial experiments on provided experimental materials, constructing experimental result data spaces. AI models build functional associations between experimental combinations and results to learn and iteratively optimize experimental combination conditions. The core of this mode lies in synthesis condition optimization learning to find optimal experimental combination conditions.

From the perspective of whether data space construction is based on existing knowledge, this can be divided into self-generating automated synthesis and learning-based automated synthesis. Self-generating automated synthesis refers to robots establishing functional relationships between synthetic materials, conditions, and experimental results through combinatorial synthesis to supply models for iterative optimization, with iterative optimization data centered on experimental material combinations and machine experimental results. Burger, B et al. [**?**] developed a mobile AI chemist robot that conducted experiments using 16 chemical samples and analyzed results using gas chromatography, employing Bayesian optimization algorithms for iterative learning based on experimental sample and result data. Learning-based synthesis involves selecting appropriate experimental material combination conditions from existing experimental synthesis data spaces to guide robotic automated synthesis. Coley C W et al. [**?**] used neural network models to screen available target molecular structures and evaluate reaction quality based on reaction transformation rule spaces. Experimental material combination processing and analysis centers on reaction condition optimization, including Linear Discriminant Analysis (LDA)

for binary classification [?] and Bayesian optimization algorithms for independent multi-classification attributes, though LDA is unsuitable for data analysis with uneven categories, and Bayesian optimization algorithms suffer from high computational costs and complex parameter tuning.

(3) Numerical-Data-Centric Processing and Analysis Mode
The numerical-data-centric processing and analysis mode uses AI model representation learning and computational capabilities to explore complex spatial associations between material/chemical structure, composition, and properties for prediction or generation. This primarily includes five stages: data cleaning, rule extraction, data enhancement, feature engineering, and association mining.

Data cleaning removes missing, erroneous, and duplicate data from datasets or unifies different representation forms. Data cleaning methods include denoising and standardization. Denoising uses statistical methods and neural network catastrophic forgetting strategies [?] to eliminate abnormal chemical reaction data with low learning rates, improving forward prediction and retrosynthesis model performance. Additionally, due to different databases selecting different atomic starting points for describing molecular structures, resulting in different SMILES representations of molecular structures, it is necessary to transform them into unified normalized formats to remove duplicate molecules, with common toolkits including the Python toolkit RDKit and Java toolkit CDK [?]. The RDKit toolkit can further calculate molecular descriptors on the basis of molecular standardization, such as compound structure similarity calculation, molecular conformation optimization, and molecular fingerprint generation. The CDK toolkit can also search compound substructures, generate 3D images, and create molecular fingerprints on the basis of data normalization.

Rule extraction identifies potential reaction centers based on learning atomic mapping information between reactants and products in chemical reactions. Current extraction technologies are unsupervised methods based on deep neural networks, with the Transformer model as a typical example. Due to its unsupervised nature not relying on annotated data and its adaptability to unbalanced reaction types, it shows great potential in rule extraction. For example, Transformer models learned atomic arrangement change patterns in chemical reactions from unannotated chemical reaction data to extract reaction rules [?].

Data enhancement uses sampling methods to expand small-sample material/chemical data. Data enhancement methods include neural network-based data enhancement, active learning, and transfer learning [?]. Neural network-based methods generate large amounts of new data through unsupervised learning sampling, including Generative Adversarial Networks [?] and Variational Autoencoders [?]. Active learning [?] uses machine learning to select valuable samples from large amounts of unannotated data for sampling to represent the unannotated data. Transfer learning improves model predictive performance on small data by transferring knowledge from related domains, such as Gupta et al. [?] pre-training the ElemNet model on the OQMD source dataset and then fine-tuning it on the target JARVIS dataset for material

properties.

Feature engineering selects data descriptors relevant to research objectives to represent material/chemical data features for AI model learning. Feature engineering includes feature selection and feature transformation. Feature selection removes redundant features from high-dimensional material task-related features to reduce feature space dimensionality and improve model prediction accuracy and generalization capability, including filter, wrapper, and embedded methods. Filter methods rank feature importance based on statistics and mutual information, offering high computational efficiency but not considering correlations between features, with common methods including correlation coefficients [**?**] and mutual information [**?**]. Wrapper methods combine supervised learning algorithms to evaluate feature subsets during feature selection, considering correlations and dependencies between features but suffering from high computational complexity in high-dimensional feature spaces [**?**], represented by Support Vector Machine (SVM)-Recursive Feature Elimination (RFE) methods [**?**]. Embedded methods are embedded within machine learning models, with no clear distinction between feature selection and model training processes, commonly including penalty-based and tree-based methods [**?**]. Feature transformation maps high-dimensional feature spaces to low-dimensional spaces to achieve dimensionality reduction, including Principal Component Analysis (PCA) and Linear Discriminant Analysis [**?**].

Association mining explores associations between structure, composition, and performance to meet prediction needs. Association mining methods include graph-structure-based association models built with Convolutional Neural Networks and Seq2Seq deep learning models built with Bi-LSTM and Transformer models. The latter leverages its advantages in learning long-sequence long-distance dependencies to construct "translation" relationships between reactants, reagents, catalysts, and products in the chemical domain. Convolutional Neural Networks excel at processing image data, enabling association of compound structure graphs by representing atoms and chemical bonds as nodes and edges in molecular graphs to identify chemical bond changes between reactant and product atom pairs and establish associations between reactants and products [**?**]. Seq2Seq models excel at processing text-based "translation" problems, with parallel computing advantages improving model efficiency. These models transform non-numerical chemical formulas into machine-recognizable forms, such as representing molecules as strings through physicochemical descriptors or molecular fingerprints, establishing associations between product strings and reactant strings to achieve retrosynthetic route prediction goals [**?**].

2.4 Data Generation and Publication

Data generation and publication refer to the production and publication of outcome data and scientific data.

(1) AI Participation in Research Data Generation

Under the intelligent research trend, AI models actively participate in research paper writing, primarily involving paper title, abstract, and full paper generation tasks. The recently released ChatGPT [**?**] serves as a typical representative, leveraging its generative AI model advantages based on massive domain text training datasets to participate in the complete paper generation process based on input topics and keywords, including paper writing perspectives and ideas, research method or tool queries, paper outlines, relevant reference resources, generating complete paper content, polishing and improving paper content, and journal selection.

Scientific data generation is not limited to experimental data and high-throughput computational data submitted by researchers but also includes relevant data contained in scientific literature and patents. Early scientific databases primarily used manual extraction for scientific literature and patents, which proved time-consuming and costly with the massive publication and accumulation of literature and patents, leading database construction to shift toward automated approaches. Computational data differs from previous simulation data generated by high-throughput screening and computation; under the intelligent research trend, computational data consists of large amounts of data generated by AI model calculations. Additionally, AI models participate in code generation work, using generative AI models for pre-training and fine-tuning to achieve generation tasks.

(2) AI Participation in Research Data Publication Review

Facing exponentially growing scientific publication submissions, the publication workload presents a tremendous challenge, with high-quality review work being a time-consuming and labor-intensive process. To address the heavy review pressure and improve research data publication speed and efficiency, machine learning models have been introduced to participate in reviewer assignment and review opinion writing tasks. For example, Charlin et al. [**?**] designed the Toronto Paper Matching System (TPMS), calculating correlations between submitted papers and reviewers' expertise by comparing text between submitted papers and reviewers' published research outcomes. Yuan et al. [**?**] used the BART pre-trained model to learn "review translation patterns" between papers from the International Conference on Learning Representations (ICLR) and NeurIPS conferences and their review opinions.

2.5 Data Storage and Sharing

Data storage involves orderly management of research outcomes and related research data produced upon project completion to enable research data discoverability, accessibility, and reusability. Data storage and sharing policies are primarily stipulated directly by funding agencies and journal publishers. Funding agencies regulate the submission and sharing of research papers and related research data derived from funded projects, represented by national-level policies including the EU Horizon 2020 policy and China' s "Technical and Management Specifications for Scientific Data Submission from Science and Technology

Programs." Journal publishers issue storage and sharing management policies for research data related to papers after obtaining paper transfer rights. The ultimate goal of both is to construct research data storage and sharing platforms adapted to intelligent research.

Regarding data storage and sharing principles, existing research data storage and sharing platforms are based on FAIR principles, focusing on data publication and sharing but lacking in utilization aspects. They primarily use data aggregation/submission as the main sharing mode, which is unsuitable for intelligent research demands for multi-source heterogeneous and associated data storage and sharing. Facing data fusion and association needs, the PARIS sharing and utilization principles have emerged, addressing distributed, isolated, and differentiated research data issues from five aspects: machine-processable analysis, online Q&A access, data security and reliability, data association and migration, and effective data supply, to achieve high-quality research data supply demands [**?**].

Regarding data storage and sharing policies, open data platform research data policies exhibit poorer standardization and fewer restrictions compared to journal publishers, with no mandatory requirements for data sharing and no development of relevant sharing standards. Compared to foreign journal publishers, Chinese journal publishers (represented by Science Press) have more generalized research data policies, especially regarding data repository guidelines, which reference relevant policies issued by foreign publishers.

Regarding data storage and sharing platform construction, platforms include general-purpose, domain-specific, and self-built databases, with general-purpose and domain-specific databases primarily being well-known open data platforms such as Figshare, GitHub, and PubChem. Additionally, intelligent research faces mismatches between existing research data and their structures with research needs, as well as demands for batch data storage and sharing of research prediction results. Therefore, researchers promote research data storage and sharing through self-built databases, with a typical case being the University of Science and Technology of China's chemical robot research work [**?**]. The USTC research team constructed a chemical reaction database storing 11.2 million chemical reactions including reactant and product structures, names, reagents, solvents, catalysts, and environmental parameters such as reaction temperature to meet the chemical robot' s learning needs from literature chemical reaction knowledge.

2.6 Data Reuse

Data reuse involves secondary utilization of research outcomes to support new rounds of data analysis and mining work, serving as the starting point for new intelligent research cycles. From the perspective of research outcome reuse, intelligent research itself represents the redevelopment and use of existing data. Textual literature reuse manifests as deep mining of new rounds of scientific liter-

ature and reuse of patent text mining data, while scientific data reuse manifests as repeated use of databases and published datasets. From the dataset reuse perspective, scientific literature datasets in intelligent research have become the core of data reuse due to their accessibility and rich content. Scientific data reuse analyzes database data reuse rates through data sharing platforms, i.e., evaluating dataset value and novelty through data browsing, download, and citation metrics, with platforms such as Figshare and Zenodo providing data statistics services on dataset interfaces.

3 The Role of Research Data in Intelligent Scientific Research

The AI for Science trend has emerged and developed under the data-intensive research paradigm, where data is the foundation and source of discovery, and AI technology is the research engine. Intelligent research centers on AI models as the core technology, with model parameters requiring training data to capture diverse domain knowledge features, i.e., mining key information from existing knowledge spaces to construct knowledge pathways. Larger data volumes enable AI models to learn more comprehensive implicit key information and more accurate association rules. Therefore, high-quality, positive-negative sample combined, multi-source heterogeneous, and structured research data play crucial roles in advancing intelligent research.

(1) High-Quality Research Data as an "Accelerator" for Intelligent Research Accuracy

Authoritative scholars in artificial intelligence and machine learning have repeatedly emphasized "data-centric AI." Intelligent scientists similarly recognize the impact of research data quality on knowledge discovery, continuously exploring methods and technologies for constructing high-quality datasets. High-quality data not only reduces complexity in data collection and pre-training stages but also improves AI model performance.

Existing well-known research databases have diverse collection pathways but still exhibit shortcomings in quality control. For example, research data collection has not fully considered AI needs, resulting in issues such as data redundancy, limited annotated data, and data consistency or standardization problems, making it unsuitable for AI model learning. Therefore, intelligent research still needs to focus on processing and constructing high-quality domain data.

(2) Positive Samples Indicate Model Learning Direction, Negative Samples Define Learning Boundaries

Positive sample data in research data trains AI models to learn common features present in the data. Negative samples refer to unsuccessful or low-quality data from scientific experiments, also called negative data, which serve a comparative distinction role, delineating boundaries for common feature learning to avoid models repeatedly capturing erroneous features and improving errors in AI model knowledge discovery.

Most research data in existing databases consists of positive samples, which cannot meet intelligent research demands for negative sample data. Negative sample mining still relies on research groups collecting their own "failure data" from research processes, which is time-consuming and yields small data volumes. Therefore, intelligent research needs to emphasize the collection, storage, and publication of negative sample data.

(3) Multi-Source Heterogeneous Data as a "Protective Cabin" for Intelligent Research Comprehensiveness

The knowledge discovery process under the intelligent research trend is a complex problem-solving process that requires decomposition into different hierarchical sub-problems to simplify solution complexity. With clear problem objectives, the problem-solving process involves data hierarchies and attribute parameters to construct solution functions. Heterogeneous data expands understanding angles or hierarchies of problems, while multi-source data enriches attribute parameter information at different hierarchies. For example, material property characteristics are related to atomic-scale atomic structures, electronic structures, and ion transport barriers, while material property responses to external environments have functional relationships with external field condition changes [**?**].

Multi-source heterogeneity refers to the diversity of research data sources and the heterogeneity of formats and presentation forms. First, scientific data is primarily extracted from experimental text, tables, and images in scientific literature and patent texts, with text, tables, and images exhibiting different structural characteristics. Second, scientific data extracted from text includes not only numerical data but also image data, three-dimensional structural data, etc. These multi-source, multi-scale data enable intelligent scientists to understand information through multiple channels and approaches and mine data associations. Taking the materials domain as an example, multi-source, multi-scale data assists in exploring micro-meso-macro scale characterization and relationships.

(4) Data Structuring as a Bridge for Human-Machine Interaction

Data structuring refers to the need to combine extracted data according to certain hierarchical and semantic structures to form standardized formats that are easy to understand and use. The ultimate purpose of standardization is to achieve machine understandability and usability.

Currently, textual literature presents the latest research data and information in semi-structured forms, with similar domain research data dispersed across large amounts of scientific literature and patent texts in forms readable and understandable by researchers. Due to different researchers' writing habits, research data representation and organization differ between literature sources. Therefore, it is necessary to construct standardized languages and organizational formats to normalize relevant data in literature, meeting the discoverability, usability, machine understandability, and reusability demands of research data under data-intensive and intelligent research trends.

4 Latent Needs for Research Data in Intelligent Scientific Research

Analysis of the above intelligent research data processing workflow reveals that AI models demonstrate demands for multi-source heterogeneous data integration, fine-grained data structuring, exploration of human-machine interaction data representation forms, data association mining, and enrichment of research data types.

(1) Multi-Source Heterogeneous Data Integration
Multi-source heterogeneous data ensures comprehensiveness in target data and attribute acquisition, depicting target data knowledge from multiple angles and benefiting comprehensive feature learning for AI models. From the above case analysis, intelligent research data originates from multiple multi-type databases, such as the AlphaFold model [**?**] combining the PDB protein structure database with Uniclust30 protein sequence data to learn protein structure assembly rules. This demonstrates AI model demands for integrated multi-source heterogeneous data under the intelligent research trend. Multi-source heterogeneous data integration not only benefits AI model learning but also facilitates researchers' data collection, improving research efficiency.

(2) Fine-Grained Data Structuring
Under the intelligent research trend, AI models pay greater attention to learning implicit rules within data, i.e., requiring fine-grained mining of data features and associating different data types to construct structured datasets understandable by both researchers and machines, facilitating researcher access and utilization. Typical cases are represented by extraction, organization, and structuring of experimental method information from literature, such as organization of experimental decomposition flowcharts centered on target synthetic substances, which not only meets researchers' fine-grained knowledge learning needs but also satisfies AI model learning needs for implicit knowledge in literature data.

(3) Exploration of Human-Machine Interaction Data Representation
Existing data organization forms are presented in human-understandable forms, meeting researchers' knowledge learning needs. However, intelligent research must not only focus on researchers' knowledge learning needs but also address AI model knowledge learning needs, requiring further transformation of existing knowledge into AI model-understandable forms to build bridges between human language and machine language. This is represented by vectorized database construction, such as Science Navigator, which uses vector computing technology and large language models to achieve vectorized representation, semantic search, and similarity calculation of unstructured literature data as infrastructure support for intelligent research development.

(4) Data Association Mining
Knowledge discovery under the intelligent research trend centers on association identification and mining, constructing association relationships between different hierarchical data to achieve data or feature prediction goals,

such as constructing complex structure-activity relationships like material composition-structure-process-performance in materials science and molecular structure-property-function in chemistry. Additionally, existing AI models are "black box" models lacking interpretability in data association mining, making prediction results difficult to understand. Therefore, facing intelligent research demands, it is necessary to construct interpretable AI models to facilitate understanding of association rule mining and further achieve explicit representation of implicit relationships between domain research data.

(5) Enrichment of Research Data Types

Intelligent research increasingly emphasizes extraction and organization of experimental procedure plans, which are important data resources for intelligent research robot learning and the core of intelligent analysis and discovery of experimental combination patterns. The content involves combination relationships between different experimental elements and their quantities, representing refinement and summarization of scientific literature that enhances machine readability of experimental protocols. Existing experimental protocol combinations use simple text forms, with retrieval centered on single experimental substance names, which cannot meet user retrieval needs centered on experimental objectives, experimental steps, or experimental principles. In the future, domain experimental protocol knowledge graphs combined with precise recommendation technologies can be constructed to provide recommended content based on users' multifaceted needs, assisting researchers in efficiently selecting experimental protocols.

5 Recommendations for Research Data Services in the Era of Intelligent Scientific Research

Regarding how library and information institutions can deeply participate in the new research paradigm and leverage their advantages in data services, this section provides the following recommendations for research data service development based on the above analysis.

(1) Accelerate Construction of High-Quality, Comprehensive Domain Data Networks

Data is one of the important driving forces for intelligent research, and its accessibility, comprehensiveness, and usability affect the efficiency and quality of intelligent research. From the perspective of data sources used, existing intelligent research cases primarily use data from foreign databases, open data platforms, or commercial publishers, with less utilization of domestically constructed databases and open data platforms, indicating low utilization rates. This also suggests that domestically constructed databases need to further improve data quality and structuring degrees and strengthen the establishment and promotion of open, high-quality databases. From the perspective of the number of data sources used, data used in intelligent research is dispersed across multiple data platforms, requiring different data acquisition and analysis methods for different platforms, with data mining and analysis accounting for a large

proportion of research and reducing research efficiency. Therefore, it is necessary to construct unified, high-quality, standardized domain data platforms that integrate open, commercial, and private research data to meet intelligent research demands for multi-source data.

(2) Emphasize Mining Experimental Information from Textual Data
Existing intelligent research emphasizes fine-grained content mining of textual data to construct structured associated knowledge. Intelligent research in basic sciences is represented by association mining and structuring of experimental information, becoming core data for automated experimental workflows. Existing experimental information is represented by Springer Nature' s Protocols laboratory guide database and CAS' s Synthetic Methods synthesis experimental method database, which are suitable for researcher query and learning but unsuitable for use and input in intelligent research. Therefore, future construction of experimental procedure information databases is needed to support intelligent research.

(3) Explore Human-Machine Interaction Languages
AI models are important participants in intelligent research. Existing knowledge serves human learning needs, but its knowledge connotation and semantic relationships need further transformation into machine data representation modes, with feature or representation patterns directly affecting knowledge discovery accuracy. Therefore, it is necessary to construct standardized knowledge representation languages oriented toward different research objectives to build bridges between human knowledge and machine learning.

(4) Deepen Embedded Research Data Service Models
In the data-intensive era, China' s data service model has transformed into a data-centric service model, focusing on data collection, acquisition, and mining services in front-end data services, with lower attention to research preparation services, data processing and analysis technology selection, and data publication services in the research lifecycle, resulting in lower support and impact of embedded data service models on research innovation. Therefore, providing embedded data services from the perspective of the research lifecycle and improving researchers' information literacy and data literacy from a practical standpoint to stimulate researchers' creativity, innovation capability, and research capacity can enhance the participation, influence, and competitiveness of library and information institutions in the intelligent research trend.

(5) Enhance Librarians' Domain Knowledge and AI Literacy
Intelligent research is the result of cross-disciplinary integration between AI and other fields, with its core being the fusion of domain knowledge and AI technology. This also presents new requirements for librarians' data service capabilities. On the basis of library and information science knowledge, they must possess data analysis and mining knowledge to become data librarians; on the basis of data analysis and mining knowledge, they must possess domain knowledge to become subject librarians; and on the basis of domain knowledge, they must possess AI knowledge to become intelligent librarians, thereby improving data

service quality under the intelligent research trend.

References

[1] Pingping Du, Yuke Li, Yue Chen. Research on the Assetization Storage and Data Reuse Rights Licensing of University Scientific Research Data [J]. Library and Information Service, 2022, 66(03): 45-53.
[2] Jinghua Xue, Huiting Xu, Guangyu Chen et al. Research on Global Trends in Digital Transformation of Scientific Research Paradigms [J]. Competitive Intelligence, 2022, 18(06): 54-63.
[3] Yunqiang Zhu, Peng Pan, Lei Shi et al. Progress and Challenges in Scientific Big Data Integration and Sharing [J]. China Science and Technology Resources Review, 2017, 49(05): 2-11.
[4] INITIATIVE M G. Materials Genome Initiative Strategic Plan [Z].
[5] MIKSA T, CARDOSO J, BORBINHA J. Framing the scope of the common data model for machine-actionable data management plans; proceedings of the 2018 IEEE International Conference on Big Data (Big Data), F, 2018 [C]. IEEE.
[6] Daqing Chen. Investigation and Enlightenment of Data Management and Sharing Policies of UK Research Funding Agencies [J]. Library and Information Service, 2013, 57(08).
[7] JIANG G, SANTIAGO I A, HANYU G, et al. Automated Chemical Reaction Extraction from Scientific Literature [J]. Journal of chemical information and modeling, 2021.
[8] MEHR S H M, CRAVEN M, LEONOV A I, et al. A universal system for digitization and automatic execution of the chemical synthesis literature [J]. Science, 2020, 370(6512): 101-8.
[9] ADITYA N, CHENRU D, J K H. Using Machine Learning and Data Mining to Leverage Community Knowledge for the Engineering of Stable Metal-Organic Frameworks [J]. Journal of the American Chemical Society, 2021.
[10] MYSORE S, KIM E, STRUBELL E, et al. Automatically extracting action graphs from materials science synthesis procedures [J]. arXiv preprint arXiv:171106872, 2017.
[11] OLGA K, HAOYAN H, TANJIN H, et al. Text-mined dataset of inorganic materials synthesis recipes [J]. Scientific data, 2019, 6(1).
[12] COURT C J, COLE J M. Magnetic and superconducting phase diagrams and transition temperatures predicted using text mining and machine learning [J]. npj Computational Materials, 2020, 6(1).
[13] EDWARD K, ZACH J, ALEXANDER V G, et al. Inorganic Materials Synthesis Planning with Literature-Trained Neural Networks [J]. Journal of chemical information and modeling, 2020, 60(3).
[14] WANG Z, KONONOVA O, CRUSE K, et al. Dataset of solution-based inorganic materials synthesis procedures extracted from the scientific literature [J]. Scientific Data, 2022, 9(1): 231.
[15] M J D, E A S, L W E, et al. OSCAR4: a flexible architecture for chemical

text-mining [J]. Journal of cheminformatics, 2011, 3(1).
[16] LEZAN H, M J D, NICO A, et al. ChemicalTagger: A tool for semantic text-mining in chemistry [J]. Journal of Cheminformatics, 2011, 3(1).
[17] C S M, M C J. ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature [J]. Journal of chemical information and modeling, 2016, 56(10).
[18] XINTONG Z, STEVEN L, SEMION S, et al. Text to Insight: Accelerating Organic Materials Knowledge Extraction via Deep Learning [J]. Proceedings of the Association for Information Science and Technology, 2021, 58(1).
[19] PANG N, QIAN L, LYU W, et al. Using pretraining and text mining methods to automatically extract the chemical scientific data [J]. Data Technologies and Applications, 2021, 56(2).
[20] PARK H, KANG Y, CHOE W, et al. Mining Insights on Metal–Organic Framework Synthesis from Scientific Literature Texts [J]. Journal of Chemical Information and Modeling, 2022, 62(5): 1190-8.
[21] ZACH J, EDWARD K, SOONHYOUNG K, et al. A Machine Learning Approach to Zeolite Synthesis Enabled by Automatic Literature Data Extraction [J]. ACS central science, 2019, 5(5).
[22] BURGER B, MAFFETTONE P M, GUSEV V V, et al. A mobile robotic chemist [J]. Nature, 2020, 583(7815).
[23] W C C, A T D, M L J A, et al. A robotic platform for flow synthesis of organic compounds informed by AI planning [J]. Science (New York, NY), 2019, 365(6453).
[24] M G J, LIVA D, VINCENZA D, et al. Controlling an organic synthesis robot with machine learning to search for new reactivity [J]. Nature, 2018, 559(7714).
[25] ALESSANDRA T, PHILIPPE S, ANTONIO C, et al. Unassisted noise reduction of chemical reaction datasets [J]. Nature Machine Intelligence, 2021, 3(6).
[26] CHRISTOPH S, CHRISTIAN H, STEFAN K, et al. Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics [J]. Current pharmaceutical design, 2006, 12(17).
[27] SCHWALLER P, HOOVER B, REYMOND J-L, et al. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions [J]. Science Advances, 2021, 7(15): eabe4166.
[28] Siqi Shi, Zhangwei Tu, Xinxin Zou et al. Application of Data-Driven Machine Learning in Electrochemical Energy Storage Materials Research [J]. Energy Storage Science and Technology, 2022, 11(03): 739-59.
[29] FALAK N, ANIRUDDH H, JANAMEJAYA C, et al. A generative adversarial network-based synthetic data augmentation technique for battery condition evaluation [J]. International Journal of Energy Research, 2021, 45(13).
[30] KIM E, HUANG K, JEGELKA S, et al. Virtual screening of inorganic materials synthesis parameters with deep learning [J]. npj Computational Materials, 2017, 3(1).
[31] LOOKMAN T, BALACHANDRAN P V, XUE D, et al. Active learning in materials science with emphasis on adaptive sampling using uncertainties for

targeted design [J]. npj Computational Materials, 2019, 5(1).
[32] VISHU G, KAMAL C, FRANCESCA T, et al. Cross-property deep transfer learning framework for enhanced predictive analytics on small materials data [J]. Nature Communications, 2021, 12(1).
[33] CHERRINGTON M, THABTAH F, LU J, et al. Feature selection: filter methods performance challenges; proceedings of the 2019 International Conference on Computer and Information Sciences (ICCIS), F, 2019 [C]. IEEE.
[34] BENNASAR M, HICKS Y, SETCHI R. Feature selection using Joint Mutual Information Maximisation [J]. Expert Systems With Applications, 2015, 42(22).
[35] KHAIRE U M, DHANALAKSHMI R. Stability of feature selection algorithm: A review [J]. Journal of King Saud University - Computer and Information Sciences, 2019, (prepublish).
[36] WEIHAO Z, TEHILA E, TINGTING W, et al. Multi-feature based network revealing the structural abnormalities in autism spectrum disorder [J]. IEEE Transactions on Affective Computing, 2019.
[37] JIHENG F, MING X, XINGQUN H, et al. Machine learning accelerates the materials discovery [J]. Materials Today Communications, 2022, 33.
[38] Yue Liu, Shuchang Ma, Zhengwei Yang et al. Data Quality Governance for Machine Learning in Materials Science [J]. Journal of the Chinese Ceramic Society, 2023, 51(02): 427-37.
[39] COLEY C W, JIN W, ROGERS L, et al. A graph-convolutional neural network model for the prediction of chemical reactivity [J]. Chemical science, 2019, 10(2): 370-7.
[40] KANGJIE L, YOUJUN X, JIANFENG P, et al. Automatic retrosynthetic route planning using template-free models [J]. Chemical Science, 2020, 11(12).
[41] OPENAI. GPT-4 Technical Report [J]. ArXiv, 2023, abs/2303.08774.
[42] CHARLIN L, ZEMEL R. The Toronto paper matching system: an automated paper-reviewer assignment system [J]. 2013.
[43] YUAN W, LIU P, NEUBIG G. Can we automate scientific reviewing? [J]. Journal of Artificial Intelligence Research, 2022, 75: 171-212.
[44] Zhihong Shen, Xiaolin Zhang, Xiaohuan Zheng et al. PARIS Principles: Scientific Data Availability in Open Collaboration Environments [J]. Big Data Research: 1-16.
[45] ZHU Q, ZHANG F, HUANG Y, et al. An all-round AI-Chemist with a scientific mind [J]. National science review, 2022, 9(10).
[46] Siyuan Wu, Yuqi Wang, Ruijuan Xiao et al. Development and Application of Battery Materials Databases [J]. Acta Physica Sinica, 2020, 69(22).
[47] JOHN J, RICHARD E, ALEXANDER P, et al. Highly accurate protein structure prediction with AlphaFold [J]. Nature, 2021, 596(7873).

(Corresponding Author: Tao Han E-mail: hant@mail.las.ac.cn)

[Author Contribution Statement] Jingrui Zhang: Literature research and organization; paper writing. Tao Han, Mengge Sun: Paper revision and review.

Source: ChinaXiv —Machine translation. Verify with original.