DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications

Research on the Method of Linking Scientific Data and Literature Data through Metadata Fusion and Ontology Construction ——from the Perspective of Agricultural Science and Technology Management in China[1]

Chai Miaoling
Chengdu Library and Information Center, Chinese Academy of Sciences, China;
Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, China
chaiml@clas.ac.cn

Zhu Jiang[2]
Chengdu Library and Information Center, Chinese Academy of Sciences, China;
Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, China
zhuj@clas.ac.cn

Zeng Yi
Information Systems, College of Business, City University of Hong Kong, Hong Kong; Information Management, College of Business, Southern University of Science and Technology, China
yzeng36-c@my.cityu.edu.hk

Zhang Di
Chengdu Library and Information Center, Chinese Academy of Sciences, China;
Department of Information Resources Management, School of Economics and Management, University of Chinese Academy of Sciences, China
zhangdi@mail.las.ac.cn

Kangjilamu
The Institution of Science and Technology Information of Tibetan Autonomous Region, China
kjlmgg@126.com

Abstract

[Objective] This study proposes a metadata-ontology fusion method from the perspective of agricultural science and technology management in China, which aims to provide a method and case for cross-departmental and cross-domain scientific data sharing and fusion in China's agricultural science and technology management. [Methods/Approach] Firstly, the study presents the connotation and methods of the correlation between scientific data and Science and Technology literature （S&T literature）, and analyzes unstructured data. Secondly, the data characteristics of agricultural science and technology management are explored. Thirdly, the ontology of agricultural science and technology management is constructed to support the integration of multi-source heterogeneous data. In the empirical part, the data requirements of agricultural science and technology management in Sichuan are targeted, and the industrial chain

[1] Funding Sources: 2017 Sichuan Province International Cooperation Project, No.2017HH0094; 2023 Key Research and Development and Commercialization Project of Science and Technology Plan of Tibet Autonomous Region. Project No. XZ202301ZY0004N; 2019 Open Fund of Key Laboratory of Mountain Hazards and Surface Processes, Chinese Academy of Sciences.
[2] Corresponding author.

https://doi.org/10.23106/dcmi.953357797

and data chain are integrated to propose 20 data requirements. Finally, the ontology is established and revised, and the linkage between scientific data and S&T literature is realized on a demonstration platform. [Conclusion] The ontology of agricultural industry management is realized, and the correlation and fusion of multi-language (Chinese/English) data and unstructured data are achieved. The study builds two demonstration platforms, integrates 24,200 data, including 2,119 expert data, and realizes the correlation of 16 types of agricultural science data and 4 types of S&T literature, verifying the feasibility of the ontology.

Keywords: Metadata; Ontology; Agricultural Scientific Data; S&T Literature; Knowledge Organization; Industrial chain; Agricultural Science and Technology Management; China.

1. Introduction

Hey et al. (2012) proposed data-intensive scientific discovery and constructed scenarios for the linkage between scientific data and S&T literature, believing that the linkage could increase the "information rate" of science and enhance the scientific productivity of researchers. Berners-Lee (2006) proposed linked data, suggest libraries attempt to fuse metadata through linked data, which promoted knowledge discovery and efficient data analysis.

The fusion of metadata and ontology is an important method for reusing and interoperating data by using two different knowledge organization methods to integrate and relate data at different levels. In information systems, metadata is commonly used at the data level to describe and reveal basic information, while at the integration level, data from multiple sources are fused using methods such as linked data and ontologies. Data at these two levels serve the needs of users to access resources.

China is a major agricultural country and attaches great importance to the development of agricultural scientific data and systems. With the establishment and development of big data engineering and the Big Data Bureau in recent years, to some extent, the fragmentation of data has been remedied. However, in terms of specific requirements and goals, effective association and service support are still lacking. Specifically, firstly, there is a lack of communication and collaboration between different departments and institutions, and there is a serious problem of duplicate construction due to the lack of overall planning for digital resources. Secondly, the structure of digital resources still needs to be optimized. Resources such as experimental data, varieties, technologies, patents, and other achievements are relatively rich, while talent and policy are relatively closed and backward, and there is a lack of agricultural S&T literature. Thirdly, databases are incompatible with each other, and data sharing and correlation are weak.

Therefore, to address the above issues, this paper studies the establishment of an agricultural industry ontology with industrial characteristics and reflects management decision-making needs from the perspective of both macro and meso-level, with a view to providing a reference for the establishment of data correlation for Chinese agricultural science and technology development in terms of industrial layout and key technology identification.

2. Research Framework

The research framework includes five steps: question formulation, theoretical research, methodology, empirical analysis, discussion and conclusions. (Fig.1.) The details are as follows:

Step 1, poses a question. Based on the goal of data-driven decision-making in Chinese government agricultural science and technology, the research proposes a scientific problem of integrating and organizing scientific data and S&T literature.

Step 2, discusses theoretical issues. In this step, the research is conducted to address the issues that arise in the process of integrating metadata and constructing an ontology for agricultural Science and Technology management (S&T management). Specifically, the

https://doi.org/10.23106/dcmi.953357797

combination between data chain and industry chain, enhancement of common metadata associations, and how to construct an ontology from the perspective agricultural management.

Step 3, is focus on methodology. It involves constructing a model for integrating scientific data and S&T literature based on the proposed research methods. The agricultural industry management ontology is then constructed.

Step 4, is empirical analysis. The constructed ontology and integrated model are applied to Sichuan Province agricultural management. Based on the results, the researchers modify the ontology, and validate it on the system platform.

Based on these steps, the last section is talks about discussion and conclusions. The paper discusses the ontology and proposes 3 recommendations and the next steps for research.

The fusion of metadata and ontology is an important method for metadata reuse and interoperate. It uses two different information organization methods to integrate and relate data at different levels. Metadata is commonly used at the basic level to describe and reveal data and information, while at the integration level, data and information from multiple sources are fused using ontologies, etc.

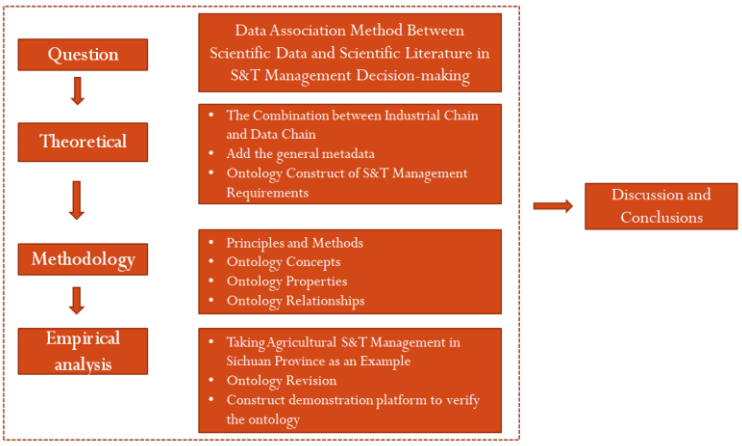

FIG. 1. Research framework.

3. Related works

3.1. Related Works of Scientific Data Research

The research on scientific data can be traced back to the modeling, collection, and preservation of earth science data by the National Center for Atmospheric Research established in the United States in 1960 (Bai et al., 2019). In the field of agriculture, the construction of scientific data resources is an important area of research. In the United States, numerous agricultural companies, professional associations, and farms have adopted computing and network technologies, and the largest agricultural computer network system in the world, AGNET, has been established, covering 46 states in the United States, 6 provinces in Canada, and 7 other countries outside the US and Canada. It is connected with the US Department of Agriculture, a large number of agricultural enterprises, 15 state agricultural departments, and 36 universities (Yan, 2019). In 2017, Germany developed and launched the Spatial Data Infrastructure (SDI) for uploading and providing soil agricultural research data （Specka et al., 2019）. The Rural Development Administration (RDA) of South Korea has jointly developed a Mobile Agriculture Information Service System (MFISS) with relevant departments, providing information support for agricultural science and technology workers and farmers, including a mobile pest and disease warning system and an origin tracking system (Guan et al., 2018). From the perspective of information resource management, existing data correlation research covers data interoperability

https://doi.org/10.23106/dcmi.953357797

(Enayat et al., 2015), interdisciplinary data management (Martin et al., 2017), and security and privacy issues related to agricultural data openness (Ferrag et al., 2020).

The research on scientific data in China can be traced back to the launch of the First Pilot Project of the National Scientific Data Sharing Project, the Meteorological Science Data Sharing Project, at the end of 2001. It has subsequently led to the construction and sharing of scientific data in the fields of resource and environment, agriculture, population and health, and basic and frontier sciences (Science Data Sharing Engineering, 2020). In order to ensure the integration of agricultural scientific data resources, the Agricultural Information Institute of the Chinese Academy of Agricultural Sciences, taking advantage of the opportunity of the "National Science and Technology Infrastructure Platform Construction" project, led the construction of the Agricultural Science Data Sharing Center (Zhu et al., 2017). Chinese researchers cover the entire process of information management, including resource integration (He, 2014), data storage (Wang & Huang, 2015), data supervision (Lu et al., 2017), data submission and management (Zhao, 2009), data openness and publication (Zhao & Wang, 2016), and data reuse (Peng et al., 2019).

3.2. Related Works of Fusion of Metadata and Ontology

Metadata refers to data about data, used to describe the characteristics and attributes of the data itself in scientific data. The idea behind this method is to find physical connections by describing the data and related nodes based on the external and internal feature descriptions of multi-source heterogeneous data.

Ontology is a fundamental method of semantic correlation, which achieves data integration by conceptualizing data content. This method is beneficial for discovering entities and establishing correlations from a semantic perspective, and enhancing data commonality. In the field of agricultural knowledge, ontology has been widely researched and applied as a knowledge organization method based on data semantics. For example, mainstream agricultural ontologies such as AGROVOC Multilingual Thesaurus and the United States Agricultural Thesaurus are based on the modification of agricultural thesauri, transformed into ontologies using Simple Knowledge Organization System (SKOS) or OWL, and are focused on generality in specific applications, which cannot be directly applied to a specific field. Especially for macro and meso-level agricultural technology managers, data fragmentation and different institutional data systems are major challenges.

The ontology construction method for metadata fusion is a logical approach that uses metadata to establish relationships between heterogeneous data sources, enabling the discovery of entities and logical relationships from a resource description and semantic perspective, and enhancing data commonality. This method is beneficial for promoting the integration of natural language and controlled language in retrieval, establishing logical relationships between scientific data, and achieving knowledge reasoning. The integration principle follows the FAIR principles proposed by European Science Cloud (EOSC) in 2014, which refers to making data Findable, Accessible, Interoperable, and Reusable. Cui et al. (2022) proposed a four-level Equipment Integrated Logistics Support (ILS) data modeling framework based on metadata and ontology to solve the problem of unified expression of ILS data. Zhou (2022) used user interviews to construct a top-level ontology by reusing CIDOC Conceptual Reference Model (CRM) and Time Ontology and used a combination of top-down and bottom-up concept extraction and expansion to obtain knowledge concepts and examples for a great website. Sun et al. (2017) proposed the linkage model between scientific data and scientific literature can be divided into two types from the similarity perspective. One is citations, and the other linkage is the similarities of description. Chen et al. (2022) took metadata as the entry point and extracted relevant metadata items of journal articles through the "related articles" field to conduct association and integration research between data papers and journal articles. Huang et al. (2021) based their research on external and semantic features of data and used semantic entities to create relevant relationships between

https://doi.org/10.23106/dcmi.953357797

literature and data, achieving deep association discovery between scientific data and academic literature.

In addition to metadata and ontology, methods such as linked data, standardized data, and data modeling are also commonly used for existing scientific data integration and compilation （Zhang et al., 2022）.

4. Issues

4.1. Insufficient data to meet decision-making needs

As a major agricultural country, China has established a large number of data and resource platforms, providing a good foundation for agricultural technology management. However, there are still phenomena such as insufficient data and data silos. Specifically, this manifests in three ways:

First, data usage is guided by national policies. In terms of data usage, management departments focus on the use of statistical data, result transformation and talent data, with characteristics of industrialized management. This characteristic is determined by the nature of the management department, and it specifically manifests in the use of macro and meso-level data with statistical significance, emphasizing result transformation and talent, and not pursuing scientific research content and details. Relevant data types include policy and regulation data, scientific and technological project data, award-winning achievement data, and expert data. In recent years, data usage has supported industrial management and also reflected the characteristics of science and technology poverty alleviation work, specifically in the collection of data on science and technology experts, talent from impoverished areas, and science and technology envoys. See Table 1 for the data usage situation of the Sichuan Rural Technology Development Center.

Second, the decision-making bases on the industrial chain. In agricultural management, to facilitate decision-making and guide industrial development, the data needs of China's agricultural technology management departments are based on the industrial chain. Taking the Sichuan Rural Technology Development Center as an example, it has designed five industrial chain links that are adapted to science and technology management, covering breeding, planting and animal husbandry, processing, logistics, and by-product utilization. This design is closely related to the functions of institutional management, with clear target direction. However, with the increasing attention paid to high-quality breeding and deep processing, the original links need to be further subdivided.

Third, participate in technology management data focuses on macro level. Another characteristic of S&T management needs is that data is classified according to industry characteristics, with more macro-level categories. Taking the data of the Sichuan Rural Technology Development Center as an example, in order to be closer to the needs of industrial development, the standardization of data processing tends to be simple and practical in terms of industrial classification. The classification methods commonly used in S&T literature, such as the Chinese Library Classification and the Chinese Academy of Sciences Library Classification, have significant differences in refinement. In terms of classification levels, science and technology management data is similar to the second or third-level categories under the agriculture classification.

4.2. Insufficient association points for heterogeneous data

Research selects data from institutions and platforms, compares and analyzes metadata content and features, assesses the feasibility of integration, and conducts a feasibility analysis of the integration of agricultural scientific data and S&T literature from the perspectives of management requirements, data characteristics, and industrial goals. Based on the characterization of

https://doi.org/10.23106/dcmi.953357797

agricultural scientific data and literature, they are divided static and dynamic data. Static data mainly including the content and external structure of resources, such as titles, authors, subject classification codes, keywords, etc. Dynamic data mainly reflect the spatiotemporal characteristics of the data, such as grain yield, yield growth rate, etc., described by a series of dynamic time sequences. As shown in Tables 1 and 2, the metadata can be summarized as external features that describe the objective existence of the resource and internal features that express the content. Due to the diversity of data types, different data sources have different structures. There are relatively few common external and internal features as well as similarities between the data. If data linkage is required, it is necessary to supplement other metadata for scientific data and S&T literature, such as classification, industrial distribution, geographical location, language, etc., to enhance the linkage. Therefore, the ontology construction based on metadata is proposed in the data organization scheme.

4.3. Current ontologies cannot support S&T management needs

Ontologies can be classified as domain ontology, scientific ontology, application ontology, etc. by its function and characteristics. Due to the specific ontology under S&T management needs, which spans both management and research, it requires a specialized one. The current ontologies cannot meet the needs, so it is significance to rebuild one to support the Science & Technology decision-making.

5. Methodology

5.1. Integration of Agricultural Science and Technology Management Data Chain and Industry Chain

The research proposes that there are three parts that need to be integrated.

Firstly, data needs to support decision-making, prediction, indicator analysis and policy-making, target and strategic planning. Since science and technology management is guided by national policies in data usage, it focuses on supporting macro and meso-level management.

Secondly, reconstruct the industry chain on the basis of the original science and technology management industry chain structure. Design six links including breeding, cultivation, primary processing, deep processing, comprehensive utilization of by-products, and storage and transportation logistics. The processing links are divided into primary processing and deep processing, and the logistics and transportation links are reordered and defined.

Thirdly, industry management focuses on the related elements of industrial development, and data processing in management emphasizes the identification of industry chain links. Therefore, the data reflecting the relationship between industrial input and output is the core dataset, which is considered to include production data, policy regulations, technical requirements, scientific and technological achievements and awards, project investment, expert data, scientific and technological literature, patents, etc. From the perspective of data scope, it covers scientific data and literature data, numerical data and text data, Chinese data and English data, and other data types.

5.2. Analyze Metadata

The first step in data fusion and integration is the selection of data sources. It will help to find relevant points of scientific data and S&T literature connection.

The study is based on the agricultural science data in Sichuan. The selected data sets come from agricultural research management institutions, library and information institutions, and scientific research institutions/scientific data platforms. 16 data sets maybe commonly used in science and technology management (Table 1), while 4 types are used in S&T literature (Table2). Then, through the analysis of these data sets, it is found that: (1) the metadata of scientific data

https://doi.org/10.23106/dcmi.953357797

can be divided into two categories: external feature description and internal feature description; and (2) it is possible to enhance the metadata of scientific data and literature by supplementing metadata, such as adding classification, industry chain, geographic location, language, etc.

TABLE 1: Metadata schema of agricultural science data.

Resource Type	Metadata	Data Source	Data Type
S&T Plan Projects	Name, Project leader, Contact institution, Project type, Approval Year, Location	Sichuan Rural Technology Development Center	Text data
New Variety & New Product	Name, Certificate number, Owner, Source, Characteristics, Cultivation points, Feature characteristics, Quality, Yield performance, Approval year, Range, Institution location		Text data
Experts	Name, Title, Location, Institution, Research field		Text data
S&T Commissioner	Name, Position title, Location, Institution, Research field		Text data
Talent in Impoverished Areas	Name, Gender, Professional title, Industry, Professional field, Work unit, Assigned to city/state		Text data
Prize-winning Achievements	Project name, Registration number, Main completion person, Main completion unit, institution location, Achievement level, Award name, Award year, Award level, Award category, Recommended department, Achievement introduction		Text data
Technology Demands	Demand, Classification, Demand period, Demand institution, Demand location		Text data
Service Organization	Name, Description, URL, Type, Institution level, Establishment time		Text data
Enterprise	Name, Description, URL, Type, Institution level, Establishment time		Text data
Grain Production Data	ID, Crop type, Unit yield/mu/kilograms, Planting area/ 10,000 mu, Total output/10,000 tons, Yield growth rate /mu (%), Area growth rate (%), Output growth rate, Proportion of area to grain production (%), Proportion of output to grain production (%), Yield/mu equivalent to the national level, Area proportion to national average level, Output proportion to national level, Year, Province	National Science and Technology Infrastructure Platform-National Agricultural Science Data Center	Numeric data
Livestock & Poultry Categories Production Data	ID, Output in 10,000 tons, Output proportion to national level, Year, Location		Numeric data
Sheep Production Data	ID, Output in 10,000 tons, Output proportion to national level, Year, Location		Numeric data
Cattle Production Data	ID, Output in 10,000 tons, Output proportion to national level, Year, Location		Numeric data
Rabbits Production Data	ID, Output in 10,000 tons, Output proportion to national level, Year, Location		Numeric data
Poultry Production Data	ID, Output in 10,000 tons, Output proportion to national level, Year, Location		Numeric data
Bees Production Data	ID, Output in 10,000 tons, Output proportion to national level, Year, Location		Numeric data

TABLE 2: Metadata schema of scientific and technological literature.

Resource Type	Metadata	Data Source	Data Type
Patents	Grant year, application number, title, main classification code, classification code, applicant (patentee), inventor (designer), publication date, publication number, patent agency, agent, filing date, address, abstract, province code, patentee's location	China National Intellectual Property Administration	Text data
Chinese Journal Articles	Title, author, organization name, keywords, abstract, publication year, journal name, volume and issue number, publication time, source, location	CNKI	Text data
Foreign Language Literature	Work ID, title, other language titles, collection ID, collection title, organization name, contributor organization name, contributor name, contributor type, source ID, source name, publication type, OA in One classification code, keywords, alternative keywords, language, abstract volume, issue, publication year, publication date, self uri	OAinOne	Text data
Policies & Regulations	Title, organization name, OA in One classification code, industry field, industrial chain, type of policy or regulation, release time, effective time, etc.		Text data

https://doi.org/10.23106/dcmi.953357797

5.3. The Fusion between Metadata and Ontology

(1) Principles and Methods

The study aims to establish an ontology for agricultural industry management based on the agricultural management characteristics. It provides industry knowledge retrieval services to managers, industry entities, and industry service providers from the perspective of macro and meso-level decision-making in agricultural industry technology.

The study uses the seven-step method to build the ontology. At the beginning, based on existing metadata and experts' opinions, we determine the professional fields and scope of the ontology, then examine the possibility of using existing ontologies, list important terms in the ontology, and define the concepts/classes and their hierarchical structure, the properties of the concepts, the facets of the properties, and create instances. To ensure the quality, the study checks the concepts, properties, and relationships manually.

(2) Structure and Content

The ontology structure includes concept/class, property and instance.. In this study, concepts are collections of individuals usually organized by classification, and relationships can be inherited through classification. Properties are used to describe the properties of concepts or instances. Instances are entities linked with a specific concept.

The ontology following Noy and MCGuinness's（2001）definition of a simple ontology, and have three parts: the first part is establishing a finite (expandable) vocabulary, the second one is interpreting the relationships between concepts and terms, and the third one is a strict hierarchical subclass relationship between classes. The ontology design fully considers the core concept set of the ontology, conventional requirements, existing conventions, regional characteristics, and values the ambiguity of domain boundaries. Furthermore, it should fully consider the characteristics of industry classification and industry chain and highlight the division of industrial fields and agricultural regional characteristics. From the content perspective, the ontology is the construction of the concept layer and instance layer. The concept layer includes the definition of classes, the hierarchical structure of classes, class relationships, and properties, forming a semantic network at the concept level. The instance layer mainly describes specific instances on the basis of the concept layer, filling instances into the semantic network of concepts, and realizing the organization and application of specific knowledge.

(3) Concepts

According to the method, the study investigates ontologies, thesaurus, and classification related to industrial construction in the agricultural domain and general domain. The AGROVOC multilingual thesaurus, the Science and Technology Knowledge Organization System (STKOS), the Agricultural Technology Extension Law of the People's Republic of China (20120831), the National Economic Classification, the Chinese Library Classification, and the Chinese Academy of Sciences Library Classification are referenced. Since the existing ontologies are not entirely suitable for industrial ontology construction, the study comprehensively attempts to transform and absorb concepts with industrial characteristics. The industry ontology of the study covers a variety of scientific data, literature and industry data. In order to describe and organize these data, together with the agricultural industry experts, we summarized 13 first-level concepts by top-down concept refinement and bottom-up clustering. Among them, considering the characteristics of multi-source heterogeneous data, the study adds concepts that can express common features, such as geographical location, domain, industrial chain links, and language. The following table (Table 3) shows the data served by each concept.

TABLE 3: Concepts and corresponding data of industrial ontology.

Concept	The Data to Describe
Policy and Regulation	Full-text policy and regulation documents
Industrial Scale	Livestock and poultry quantity, production data for pigs, cows, sheep, poultry, and rabbits; production data for bees and beneficial insects; grain and oil production data

https://doi.org/10.23106/dcmi.953357797

Industrial Input	Project input, funding input (no data available)
Industrial Achievement	Scientific and technological achievement awards, full-text foreign literature, Chinese abstracts, patents, new varieties and new technologies
Industrial Talents	Talents from the Three Zones, scientific and technological envoys, technical experts
Industrial Enterprise	Enterprise data
Industrial Service	Research services, public utility services (no data available), collaborative services, achievement transformation service data
Industrial Technology	Technology requirements, scientific and technological achievement awards data, new varieties and new technologies
Industrial Requirement	Technology requirements, talent requirements (no data available)
Industry Classification	All of the above data
Industry Chain	All of the above data
Industrial Geography	All of the above data
Language	All of the above data

65 secondary subclasses, and several tertiary subclasses are defined and created. The depth of the hierarchy is no more than three levels. The content of the concepts covers the classification of industrial chain links, agricultural field classification, resource type classification, and management element classification, and can describe and link multi-source heterogeneous data, enabling the ontology to be better applied to industrial management application scenarios and improving knowledge utilization efficiency. Table 4 shows the top concepts and corresponding subclasses.

TABLE 4: Top concepts and corresponding subclasses.

Top Concept	**Secondary Subclass**
Policy and Regulation	Agriculture and Rural Land National Policy and Regulation, National Food Production National Policy and Regulation, Farmer Income Growth National Policy and Regulation, Agricultural Subsidy National Policy and Regulation, Price National Policy and Regulation, Market National Policy and Regulation, Open National Policy and Regulation, Sci and Tech. Progress National Policy and Regulation, Quality and Safety National Policy and Regulation, Social Service National Policy and Regulation, Specialized Farmer Cooperatives National Policy and Regulation, Rural Labor Employment Transform National Policy and Regulation, Rural Infrastructure Construction National Policy and Regulation, Rural Finance National Policy and Regulation, Agricultural and Rural Environment National Policy and Regulation, Poverty Alleviation and Development National Policy and Regulation
Industrial Scale	/
Industrial Input	/
Industrial Achievement	Achievements and Awards, Literature, Patent, New Variety and Product
Industrial Talents	Sci Tech Expert, Poverty Stricken Aera Talent, Technical Envoy
Industrial Enterprise	Leading Enterprise, High-tech Enterprise, Minor Enterprise
Industrial Service	Academic Service, Public Service, Farm Helper Service, Achievement Transformation Service
Industrial Technology	Elite Seed Production, Cultivation, Fertilization and Breeding Technology; Plant Disease, Animal Epidemic Diseases and Other Pests Prevention and Control Technology; Harvest, Process, Package, Storage, and Transportation Technology; Safe Use, Quality and Safety Technology; Water Conservancy, Water Supply and Sewage, Soil Improvement and Conservation Technology; Mechanization, Aviation, Agrometeorology and Information Technology; Disaster, Resources and Ecological, Energy Technology; Other Technology;
Industrial Requirement	/
Industry Classification	Grain and Oil, Vegetable, Fruits, Forestry, Horticultural Plants, Plant Fibers, Traditional Chinese Medical Plants, Cash Crops, Livestock and Poultries, Feed Crops, Forage and Green Manure Crops, Sericulture, Bees and Beneficial Insects, Aquaculture, Comprehensive
Industry Chain	Seed-Breeding; Planting and Breeding; Preliminary Processing In Producing Area; Fine and Intensive Processing; Comprehensive Utilization of Byproducts; Storage, Transportation and Logistics
Industrial Geography	Administrative divisions, Planting areas

https://doi.org/10.23106/dcmi.953357797

Language	Chinese, English

(4) Properties

Properties can be categorized as concept and instance properties. Concept properties are mainly used to describe the characteristics or structure of concepts, and they exist in the form of attribute values, which are expressed as metadata in specific datasets. In the process of defining ontology properties, concept properties and instance properties are established according to actual needs, and there are intersections between properties. For example, the concept "industrial technology" is both a concept of the ontology and a property of the concept "industrial demand". By improving and enriching the relationships between classes, the degree of association between data is strengthened.

(5) Relationships

In terms of relationship design, it includes basic and custom relationships, which mainly describe the relevance of concepts and the whole, between the concepts, concepts and properties, concepts and instances. The basic relationships include "kind-of" inheritance, "part-of" between concepts and the whole, and "instance-of" between concept instances and concepts, etc. In addition, because relationships cannot be expressed by the ontology construction tool and DC definitions, a free description method is adopted (Table 5). Fig.2 is the example of the ontology structure. The study uses the Vocbench 3.0 to construct the COFCO (Chengdu) Grain and Oil Industry Co., Ltd (COFCO, Chengdu) as an example. COFCO, Chengdu has url， and wheat, rice, grain etc. related to it. COFCO, Chengdu is a province level's institution, and the institution's type is agricultural Ins.

TABLE 5: Example of agriculture industrial ontology relationship.

Relationship Type	**Relationship Description**
Class Inheritance Relationship	kind-of
Class and Whole Relationship	part-of
Class and Instance Relationship	instance-of
Inter-Class Relationship	located_in
	leaded_by
	relate
	influence
	language

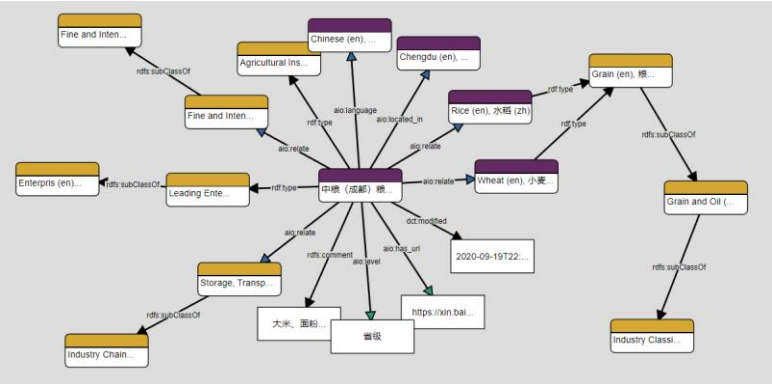

FIG.2: COFCO's agriculture industrial ontology structure.

https://doi.org/10.23106/dcmi.953357797

6. Empirical Research

6.1. Empirical Research in Sichuan Province of China

The study chose the Sichuan Province agricultural management as an example. Sichuan Province is an important agricultural province in China but lack of multi data support. Currently, there are several information platforms in Sichuan Province, such as Sichuan Province Science and Technology Achievement Transformation Information Service Platform, the Sichuan Breeding Research Data Sharing Platform, and the Multimedia Database of Agricultural Pests and Diseases. However, these platforms still fail to meet the needs of decision-making and management. Specifically, there are three main issues: first, there is a lack of overall planning for scientific data, leading to severe duplication of construction and unclear knowledge service orientation; second, databases are incompatible with each other, with uneven distribution of data resources and poor data sharing; third, there is a lack of multi-language agricultural S&T literature.

6.2. Methodology and Argumentation

In order to match data with agricultural management decision-making needs, the study achieves this goal from three aspects.

First, matches the data chain with industrial chain. Based on the work of Sichuan Rural Technology Development Center, the study rebuilds the five industrial chain links to six. The new industrial chain is covering breeding, planting and animal husbandry, primary processing, intensive processing, logistics and transport (Chai Miaoling, et al. 2020.). The data type corresponds to the industry chain, and chooses the S&T datasets and literature datasets which covered 16 types of agricultural scientific data and 4 types of literature listed in 4.2.

Second, supplements core metadata to assist in data association. Due to the heterogeneity of multiple sources of data, it is necessary to find common features in the metadata to establish the association. The study is based on the external and internal common characteristics of scientific data and S&T literature, chooses key concepts as industrial processes, industrial classification, industrial technology, and industrial geography. Whether these four concepts can be incorporated into all types of data will affect the accuracy of data association between unstructured data. The incorporation of ontology into data argumentation is of great significance for improving ontology classes, attributes, and relationships. After analyzing the metadata of 20 types of datasets, the above four types of information are extracted 100% from the dataset.

Third, the study conducts semantic analysis of data from description accuracy, areal feature and operability by hand. Yet, semantic analysis has performed about 26,190 pieces of data based on accuracy of description, regional characteristics, and operability. 23,100 pieces of vailed data have been generated, with semantic extraction rate of 88.20%. Emphasis has placed on analyzing the semantic relationship among industrial processes, industrial classification, policies and regulations, industrial technology, and industrial geography in scientific data and S&T literature.

In order to discover the effectiveness of ontology, the study establishes a demonstration platform to verify data could integrate, organized and visualized. Fig.4 takes the project of "Breakthrough Rice Mianzhuan 725 and its Variety Promotion and Application" as an example to realize the match of policies, papers, patents and needs. This step will show the possibility in the agriculture management in the future.

https://doi.org/10.23106/dcmi.953357797

FIG. 3. AGROIN demonstration platform for agricultural industry knowledge services.

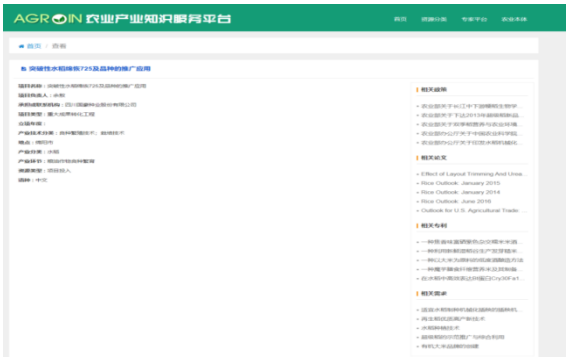

FIG. 4. The project of "Breakthrough Rice Mianzhuan 725 and its Variety Promotion and Application" and the demonstration of policies, papers, patents and needs.

6.3. Ontology Revision

The verification process mainly consists of expert discussions and data practices, and is divided into four steps: (1) 6 experts in the field of agriculture or agricultural management, and 2 experts in agricultural library and information science are invited to discuss ontology and modify it; (2) 2 PhD. Students and 3 master students in the fields of agriculture and 3 master students in library and information science are invited to validate the content; (3) some test data are imported to verify data associations; and (4) completes ontology revision.

This section mainly describes the application and revision of the core concept set in ontology: industrial processes, industrial classification, policies and regulations, industrial technology, and industrial geography in scientific data and S&T literature. Table 6 is the classification of the concept set. The testers follow three principles: accuracy, regional, and interoperable.

TABLE 6: Explanation of data testing scope.

Concept	Testing Data Scope
Industry Link	All data
Industry Classification	All data
Industry Geography	All data
Industry Technology	Technology award achievements, projects, and technology demands
Policy and Regulations	Policy data

The researchers have tested the concepts and find that concepts are effective in meet the classification needs of both scientific data and S&T literature, take into account the macroscopic nature of extracted scientific data and the microscopic characteristics of literatures. The classification process is simple, easy to use, and easy to index, which is in line with the characteristics of Sichuan's agricultural industry. The concepts are accurate and unambiguous. As an example, the industry technology concept is tested and yielded good classification results.

After testing, three issues are identified and corrected. First, semantic granularity causes the classification problems. For example, in the data testing of industry links, the classification of agricultural policies, projects, and organizations is more macroscopic, and a single data point may cover multiple third-level categories or even second-level categories. After correction, macroscopic and intermediate data are retested and classified primarily by first-level or second-level categories.

Second, concept supplementation is necessary. Through expert discussions and testing, concept data was supplemented. For example, in industry classification, special or third-level categories such as Sichuan agricultural products, forest products, and medicinal materials were added.

Third, classification perspectives are varied. The test is conducted mainly by Ph.D. and master's students in agricultural economics and management, and library and information science.

https://doi.org/10.23106/dcmi.953357797

The testing process is standardized and the semantic recognition proficiency is high. However, there are differences in classification perspectives, and the cross-disciplinary test adds difficulty to the work. Later, through third-party and cross-check audits, issues are discovered and corrected in a timely manner. Further research can establish classification indicators and use algorithmic methods to judge classification.

7. Discussion and Conclusions

7.1. Issues When Developing the Ontology

(1) Integration of Industry Chain and Data Chain

The industrial ontology is developed based on the industrial chain data and is intended to meet the requirements of scientific and technological management. Therefore, the data chain should be established based on the government's decision-making intentions and levels of government agencies to better serve the goal, but this will inevitably require alignment of different granularity data. Therefore, it is necessary to build a rich category hierarchy.

(2) Synonymy and Polysemy Processing

The ontology requires that the semantics of each concept or instance be clear and unique, which means that a semantic expression can only represent one semantic meaning; conversely, one semantic meaning can only be expressed by one concept or instance. However, because the concepts and instances involved in the industrial ontology are numerous, there will inevitably be situations where multiple words have the same meaning, or one word has multiple meanings. Therefore, it is necessary to discern synonyms and polysemy in the naming of concepts and instances.

For synonyms, the "alterLabel" property can be set. For example, "barley (大麦)" and "hulled barley (元麦)" both refer to the same crop, so "barley" has an alternate label of "hulled barley". For polysemy, whether a concept belongs to the situation of polysemy needs to be further identified by experts. For example, "potato" exists in both the "grain and oil" and "vegetable" classes. After judging, it is determined that the "potato" in both categories refers to the same crop, so there is no ambiguity. However, "ginkgo (银杏)" exists in the "forestry", "fruit", and "Chinese herbal medicine" classes at the same time. Different from potato, the "ginkgo" in the three categories refers to three different entities, indicating a situation of polysemy. In this situation, it is necessary to further refine the semantics, and modify them to "ginkgo tree (银杏树)", "ginkgo fruit (银杏果)", and "ginkgo (银杏)" respectively to clarify the semantic.

(3) Transformation from Metadata Description to Properties

When building an ontology based on metadata, since the expression of metadata is slightly different from that of the ontology, the description needs to be converted before building ontology properties. For example, when describing the location property, the metadata field is "Places", but when converted to a property, it should be renamed as "located-in" to express the semantic of "entity-located-in->place", which is closer to the expression of the ontology.

7.2. Conclusions

Based on the analysis of agricultural science data and S&T literature, this study proposes a non-structural data ontology correlation method according to data characteristics and management decision needs. It integrates ontology concepts and industrial elements, establishes an industrial management ontology in agriculture, and verifies and implements it systematically.

Based on the management decision research, the study analyzes the demand and characteristics of agricultural S&T management and proposes a data correlation scheme for constructing an industry ontology model based on semantics.

https://doi.org/10.23106/dcmi.953357797

Based on the ontology research, the study proposes the construction principles and process of ontology, constructs concepts, attributes, and relationships of the ontology, and introduces expert argumentation and expert data measurement in the construction process for ontology correction. The revised ontology has been tested on more than 23,000 pieces data and proves to be feasible. Although there are difficulties in semantic alignment of data in the industrial link and policy classification, reasonable correlations among data are achieved by adjusting the concept classification level of macro and meso data. Finally, the ontology is implemented in Vocbench3.

Eventually, this study realizes the industry ontology concept based on the platform, and correlates research data and S&T literature. The study has built two demonstration platforms, integrating 24,200 pieces of data, which including 2,119 experts, and achieved the correlation of 16 types scientific data and 4 types of literatures. Currently, the main platform integrates more than 23,100 pieces of data, achieves data correlation, language correlation, and information visualization, providing users with a graphical data experience. The agricultural expert sub-platform has built expert, consulting, and expert output modules, integrating 1,184 pieces data, and achieving the correlation of talent data in scientific data with S&T literature.

7.3. Future work

Under the demand for government digital transformation strategy, agriculture, as an important field of national economy and people's livelihood, plays a crucial role in the use of scientific data. It is recommended that scientific research institutions, universities, and library and information institutions, based on their own resource and disciplinary advantages, actively promote the development and utilization of agricultural scientific data. The following recommendations are proposed:

In the first instance, expand the types and scope of scientific data collection. With a focus on scientific and technological management, it is recommended that scientific data collection types and scope should be expanded, from the perspective of decision-making needs, to enhance the collection and integration of multi-source heterogeneous resources and explore new types of scientific data that can be used to support agricultural management. For example, data reflecting changes in crop yields and changes in the scale of animal husbandry can be used to help relevant management departments to grasp macro-level trends.

Second, strengthen research on data fusion and topic recognition methods. Based on the semantic integration of scientific data and S&T literature, it is recommended to comprehensively grasp the metadata, enhance the research on automatic topic recognition methods, and effectively reuse scientific data, while strengthening data circulation.

Finally, serve the agricultural virtual community. With the gradual improvement of digital infrastructure, artificial intelligence can already achieve basic tasks such as data integration and knowledge-based question answering. Therefore, information researchers need to change their information service strategies and focus on promoting and enhancing human creativity. For example, in a digital environment, it is recommended to establish a global agricultural virtual community, mining knowledge and assisting scientists in academic innovation.

Acknowledgements

We would like to express our sincere gratitude to Dr. Zhang Xuefu and Dr. Pan Shuchun from the Agricultural Information Institute of the Chinese Academy of Agricultural Sciences for their guidance and support in the ontology design process. We also thank Dr. Wang Fang from Sichuan Agricultural University for her support in the construction of concepts, properties, relationships, and instances in Vocbench3. And appreciate the support from Research fellow Wang Jingdong and Associate Research fellow Zou Yixing from the Sichuan Rural Science and Technology Development Center for their help in the analysis of agricultural technology management requirements.

https://doi.org/10.23106/dcmi.953357797

We are grateful to the National Agricultural Science Data Center (http://www.agridata.cn) for providing scientific data, Chinese Academy of Sciences Literature and Information Building New Capacity Project - "Open Knowledge Resource System Construction" (http://oa.las.ac.cn) for providing papers, policies, and regulations. We also thank the Sichuan Rural Science and Technology Development Center (http://www.scnckj.org.cn/) for providing scientific data for study.

References

Bai, Y., Yang, Y., & Sun, J. (2019). Advances in the study of domestic and foreign scientific data management methods. Journal of Agricultural Big Data, 1(3), 5-20+4.

Chai Miaoling，Zou Yixing，Tan Rongzhi et al. (2022). Research and Practice on Association of Scientific Data and Scientific Literature Oriented to Knowledge Service of Agricultural Industry. Journal of Library and Information Science in Agriculture. 34(3): 37-50.

Chen, S., Liu, G., & Liu, Q. (2022).A study of the association of metadata-based data papers with journal papers:take the global change science research data publishing system as an example: A case study of the Global Change Research Data Publishing System. Digital Library Forum, (08), 11-18.

Cui, X., Pan, C., Lu, J., & Han, Y. (2022). A Metadata Based Equipment Integrated Logistics Support Data Ontology Modeling Method. In 2022 3rd International Conference on Computer Vision, Image and Deep Learning and International Conference on Computer Engineering and Applications, CVIDL and ICCEA 2022 (pp. 144-148).

Enayat, R., Rajabi, S., Salvador, et al. (2015). A linked and open dataset from a network of learning repositories on organic agriculture. British Journal of Educational Technology.

Ferrag, M. A., Shu, L., Yang, X., et al. (2020). Security and Privacy for Green IoT-Based Agriculture: Review, Blockchain Solutions, and Challenges. IEEE Access, PP(99).

Guan, B., Chen, P., Luo, Z., & Shen, X. (2018). Experience and enlightenment in agricultural information service system construction in developed countries. World Agriculture, (10), 26-31. DOI: 10.13856/j.cn11-1097/s.2018.10.005.

Hey, T., Tansley, S., & Tolle, K. (2012). The Fourth Paradigm: Data-intensive Scientific Discovery. Science Press.

He, L. (2014). Research on the integration of agricultural information resources based on linked data (Master's thesis). Huazhong Normal University.

Huang, Y., Sun, T., Zhao, R., Xian, G., Li, J., & Luo, T. (2021). Research and implementation of linking services between scientific data and academic literature. Library and Information Service, 65(23), 116-125.

Lu, L., Wang, P., & Yu, X. (2017). Research on the construction of agriculture data curation platform. Library and Information Service, 61(10), 68-73.

Martin, C., Cadiou, C., & Jannès-Ober, E. (2017). Data Management: New Tools, New Organization, and New Skills in a French Research Institute.

Noy, Natalya F., McGuinness, Deborah L. (2001). Ontology development 101: a guide to creating your first ontology. Knowledge Systems Laboratory, March, 2001.

Peng, X., Wang, F., & Zhou, G. (2019). Research on the reuse-oriented agricultural scientific data sharing model. Agricultural Economics, (1), 87-89.

Science Data Sharing Engineering. (2020, September 8). Retrieved from http://www.most.gov.cn/ztzl/kjzg60/kjzg60hhcj/kjzg60jcyj/200909/t20090911_72832.htm

Specka, X., Gärtner, P., Hoffmann, C., et al. (2019). The BonaRes metadata schema for geospatial soil-agricultural research data–Merging INSPIRE and DataCite metadata schemes. Computers & Geosciences, 132, 33-41.

Sun, W., & Chang, E. (2017).Correlation analysis of scientific data and scientific literature. Library Theory and Practice, (03), 49-53.

Wang, J., & Huang, C. (2015). Research on the storage architecture and method for big agricultural scientific data. Guangdong Agricultural Sciences, 42(2), 152-156.

https://doi.org/10.23106/dcmi.953357797

Yan, D. (2019). The development and experience of digital agriculture at home and abroad. Yunnan Agriculture, (5), 48-50.

Zhang, X., Yang, Z., Pang, H., et al. (2022). Research on science data integration system and the latest progress. Information Studies: Theory & Application, 45(06), 199-206.

Zhao, H., & Wang, J. (2016). Circumstance on scientific data publication and its inspiration to China's agricultural scientific data publication. Agricultural Outlook, 12(8), 53-57.

Zhao, R. (2009). Research on data remittance and management in agricultural science data sharing. Science and Technology Management Research, 29(8), 284-286.

Zhou, J., & Si, L. (DCMI 2022.). Metadata and Ontology Design for Protection and Utilization of Great Sites. Retrieved from https://www.dublincore.org/conferences/2022/sessions/papers-metadata-as-linked-data-and-kg/#reckRWmMSpzKuC0ID

Zhu, L., Meng, X., Zhao, R., & Zhao, H. (2017). Research on resource construction for National Agricultural Science Data Sharing Center. Digital Library Forum, (11), 15-20.

https://doi.org/10.23106/dcmi.953357797