Research on the Data Conversion Specification for Chinese Agricultural Product Quantity Safety

Kaimeng Sun

Laboratory of Digital Agricultural Early-warning Technology of Ministry of Agriculture of China; Institute of Agricultural Information, CAAS,
100081 Beijing, China
kmsun@mail.caas.net.cn

Abstract. This paper summarizes the current main methods of heterogeneous data conversion, focuses on the necessity of establishing the data conversion specification for China's agricultural product quantity safety, as well as specifies the main designs and solutions for data conversion structure, data conversion process and data conversion format in research. Establishing a data conversion specification is a pioneering work of applying modern information technology management methods in the research areas of agricultural macroeconomic policy-making system, which plays a positive role in promoting the study of modern agricultural information system. The establishment of data conversion specification can not only improve the level of agricultural information technology research, but can also create a better research and development environment for the study of agricultural information system.

Keywords: specification, agricultural product, data conversion

1.Problem Statement

Chinese agricultural product quantity safety management system is an information management, analysis and decision-making system established for the study of agricultural product quantity safety. It involves large amounts of data related to agricultural product quantity safety, including the agricultural economic base data, field collection dynamic weather data, soil and crop growth data, agricultural product price data from internet, agricultural product quantity safety historical data for statistical analysis, etc., with different data sources and types. For overall system needs, to establish a unified and standardized database that can serve the agricultural system is the base for agricultural product quantity safety research.

The key indicators involved in agricultural product quantity safety management system are divided into two categories. The first category is production indicators: including the arable land area, acreage, production, chemical fertilizer and pesticide use, etc. of crop production. As the agricultural product quantity safety in our study includes various consumer goods needed by human life, here the agricultural product is a broad concept, and therefore the indicators also include the breeding area,

quantity, meat, milk, etc. of livestock production; the forest area of forestry production, all kinds of fruit production, etc.; production environment such as soil conditions, texture, moisture, fertility, weather factors, natural disaster indicators and others; the second category is consumer indicators: including demographic, agricultural economy, all kinds of income, number of expenditures, various related indices and so on. Overall, the indicators of system show a large quantity, a wide variety of sources, complex formats and other features.

Taking agricultural economic base data as an example, data are from the Ministry of Agriculture, national statistical offices and various types of information and internet dynamic snapshots, due to the differences in database platform and type of various departments, there will be heterogeneous database problems. These databases include SQL Server, Oracle, Access, etc. As different databases have different data types, storage and data retrieval methods, how to convert and integrate data so as to access to information from the huge data resources, has become a hot topic in the current database research field, especially the data conversion of heterogeneous database has become a pressing issue now.

The study of agricultural product quantity safety management system is based on the establishment of agricultural product quantity safety-related databases, while the most important issue for establishing agricultural product quantity safety databases is the conversion of various types of data, which means in accordance with the unified format requirements of agricultural product quantity safety database, to convert different types of data into agricultural product quantity safety databases in the format required. Therefore, how to convert various sources of data into a unified data structure to enter the agricultural product quantity safety database, requires creating a data conversion specification to uniform data formats, types, and conversion methods.

Specification is defined as standards of conduct established by groups, the content of specification document is binding and regulating people's behavior in nature. To establish the data conversion specification in agricultural product quantity safety-related research, is to define data formats and standards of conduct for the relevant data conversion. The meaning of establishing specification is to unify the conversion method, format and process of agricultural product quantity safety data, that is, to keep access to the data of agricultural product quantity safety management system under a unified agreement, so as to ensure various types of data convert into system use in accordance with system requirements and unified steps , in order to protect the quality of the data.

2. The Current Main Methods of Data Conversion

Data conversion means the process of converting data structure, form of expression and file format from one into another. Taking agricultural product quantity safety database as an example, data are accessed through a variety of sources, the database structure and storage form of data are not the same, as the amount of data increaseing, the original unreasonable data structure cannot meet various requirements. The replacement of database and data structure requires the conversion of data themselves. Only by data conversion to generate data and document exchange of heterogeneous

system, can data structure of database be unified in accordance with the requirements of agricultural product quantity safety research, and data be used for research to play their real role.

Currently there are a variety of ways of data conversion for different information systems.

2.1 Conversion based on the timeliness of data

Based on timeliness, data conversion consists of real-time and non-real-time (delayed treatment). Real-time data conversion is a system as a data source providing data call service for another system. Receiving service conducts real-time conversion and processing of the data accessed to provide data services for the system. This real-time data conversion is usually found in the network. For example, the agricultural product quantity safety management system calling the content of a particular website that provides data services, or the system calling some of the external programs, are all real-time conversion, which requires filtering, extracting, computing, converting and sorting the data or program called from the website by definition, then adding the converted data in accordance with certain requirements of form to the database of agricultural product quantity safety management system. This processing is required to be completed within a very short time, that is, the timeliness of data access.

Non-real-time data conversion refers to data that are difficult to convert and do not need to immediately convert after data extraction between two information systems, such as databases, office software, etc. Through some technical means, such as the application of some intermediary format, the data has been converted into a common data format, and after data conversion, it can be directly applied to different information systems. Or by programming, specially write a program for the two data systems, to convert data extracted from the two systems so that it can be applied to another system.

According to the conversion process, it can be divided into intermediate format conversion and direct conversion between source data and target data.

Comparing the two methods, we can find direct conversion between two different data formats is more efficient and of less negative effect (noise brought about, a small amount of data forced to discard, the degree of structure forced to reduce), which means the mapping from original format of data to target format of data will be maintained better. However, the possibility of direct conversion with ready-made means is not big, very often requiring special programming for converting from the source format to the target format.

2.2 Conversion based on different data structure

Through studying heterogeneous databases and making efforts to develop solutions to convert heterogeneous database data, we can achieve centralized management and uniform use of data in a unified environment. Heterogeneous database conversion is relatively complex in techniques, and currently it is still difficult to use a generic DBMS (Database Management System) to solve this problem. The common language

specification XML (Extensible Markup Language) defines data structures in an open self-described manner, which can highlight the description of the structure while describing the data content, so as to reflect the relationship between the data, therefore is the most widely used normative language for data conversion. XML can achieve the separation of data representation with data content, allowing seamless integration of data from different sources and different processing of the same data. XML language uses XML Schema to describe the structure of the document, meanwhile XML Schema has rich data types and powerful ability to express constraints, making it feasible and convenient to take XML schema as the common data model to establish a data conversion tool between unified pattern of relationships. With the development of XML technology and XML'sadvantages in data applications such as easy to express, cross-platform, semi-structured data description, good scalability, separation between content and form, etc., favorable conditions have been provided for heterogeneous database data conversion. The heterogeneous database data conversion based on XML is not only beneficial for data sharing, making savings in software development investment, but also conducive to better management and use of data, laying the foundation for government and enterprise applications of information.

The importance of information system lies in the information exchange between different information systems. The information stored in separate information systems is not complete and single, only by exchange of information can information be enriched and completed. However, with the ubiquitous presence of information systems, their development is towards diversification and specialization. Therefore, information sharing and data and document conversion between information systems has become a subject to be solved now. Only with exchange of information, as well as timely and accurate information conversion, information is useful. A data conversion specification for agricultural product quantity safety should be established, so that various types of data conversion is in accordance with the requirements of system design, exchange of different types of information is smooth, information resources are fully utilized, in order to ensure the research work of agricultural product quantity safety be carried out smoothly.

3. The Functional Structure of Agricultural Product Quantity Safety Data Conversion

3.1 The functional structure of data conversion of agricultural product quantity safety management system

Based on the data sources and structure of agricultural product safety management system, through analyzing and researching, the functional structure of data conversion is established, consisting of database operation module, database source extraction module, data structure conversion module and data key-in module according to functions. The functional structure chart of data conversion is as shown in Figure 1:
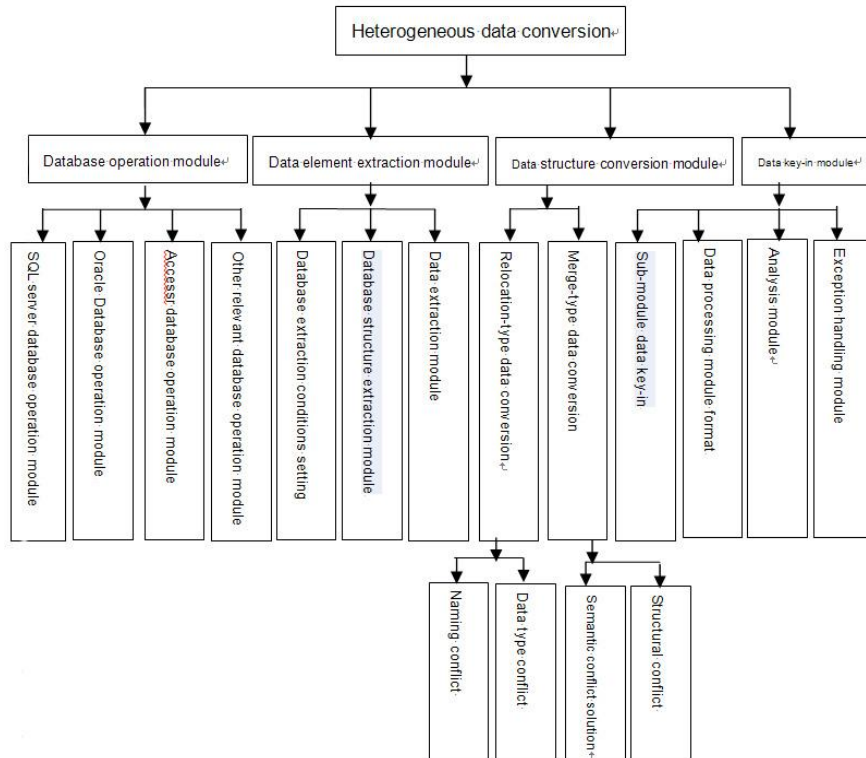
Figure 1 The functional structure chart of data conversion of agricultural product safety management system

3.2 The main functional module features of data conversion:

3.2.1 Database operation module

This module creates different classes respectively for different databases to achieve connection between a variety of databases, data query, insert, delete, modify and other operations. When you need to add a database type, you can just add the corresponding class to improve the scalability of the tool.

3.2.2 Data extraction module

This module sets data extraction conditions according to the needs of data conversion to extractthe source database data as XML data files and XML structure files.

3.2.3 Data structure conversion module

This module implements the data structure conversion between source database and target database, including two data conversion solutions, the relocation type and the merge type, to solve the existing naming conflict, data type conflict, structural conflict and semantic conflict in two data conversion solutions, achieve multi-table conversion under complex conditions, and finally generate data structure conversion files for controlling the key-in of XML data from the source database to the target database.

3.2.4 Data key-in module

According to the corresponding data insert statement generated by the data structure conversion files generated by data structure conversion module, this module inserts and updates the data of XML data files generated by data extraction module from the source database to the target database. In the data key-in process, sub-data key-in can be selected according to the data file size; when sub-data key-in is selected, XML data file will first be cut into several smaller XML data files, and then each XML data file will be circularly processed to complete the key-in of XML data to the target database. Meanwhile, the data key-in module can also complete data format adjustment based on the actual situation.

4. The Basic Data Conversion Process of Agricultural Product Quantity Safety

The data conversion process divides data conversion into two cases: relocation-type data conversion and merge-type data conversion. The relocation-type data conversion refers that the target database is empty, just needing to completely relocate the data to be converted from the source database to the target database, which approach requires to establish the target database in full accordance with the source database structure; the merge-type data conversion means the target database already exists and there is data, needing to establish the corresponding relationship between source database structure and target database structure, inserting and updating the source database data into the target database according to the structural correspondence setting.
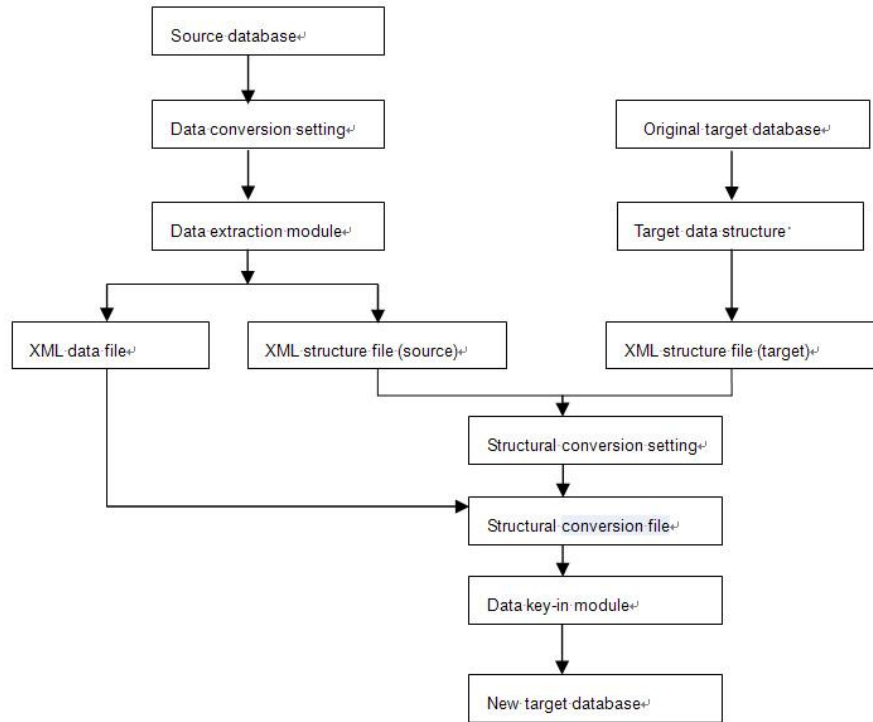
The conversion process chart from the source database to the target database is as shown in Figure 2.

First, data extraction, which is the starting point of data conversion, is responsible for accessing information from the source database to match the target database, specifically including:

- calling the database operation module
- setting the conversion data according to needs
- calling the data extraction module, to extract the setting conversion data as XML data files, and extract the structural information of conversion data as XML structure files.
- extracting the target database structure as XML structure files to prepare for the conversion between the source database and target database data structure.

Then comes data structure conversion, the data structure conversion between source database and target database is the core of data conversion process, which sets the corresponding relationship between source database and target database structure based on the XML structure files of source database and target database, to resolve the semantic conflict, naming conflict and type conflict in heterogeneous database data conversion, and generate data structure files to control the updating and inserting of source database data into the target database.

Finally, data key-in is achieved, according to the corresponding data insert statement generated by data

Figure 2 Data conversion process chart

structure conversion files, updating and inserting the source database into the target database to complete the data conversion.

5. The Exchange Format Design for Basic Data and Shared Data of Agricultural Product Quantity Safety

5.1 Extraction and standardization of data elements

5.1.1 Data source extraction
Using the logical analysis of natural language, the basic data objects of agricultural product quantity safety that need exchange format design are analyzed in terms of role, location, thing, event and time, of which the thing and event, each also contains two aspects: identification and characterization. Detailed steps are as follows:
- Make sure all relevant details of participants;
- Determine the identity elements, such as: ID numbers and names fall into this

category;
- Make sure all the details related to the event. Event must have a date / time and a type. In general, there should be a location / position where event takes place;
- Define the relevant participants. Party A and Party B are participants related to the registered event;
- Define the location / position. There are three basic types of location: geographical address, physical address (latitude and longitude) and mailing address.

5.1.2 Standardization of data elements

As for the data elements extracted, experts in the related field of agricultural product quantity safety basic data sharing will standardize the data elements in accordance with the basic principles and methods and other relevant standards of data element standardization, so as to form reusable data elements that can directly compose of syntax-independent data exchange format.

5.2 Data classification and coding

Based on the data elements formed, code processing is needed for the data elements which require coding in accordance with the methods described in the basic principles and methods of data classification and coding and other relevant standards, so as to form the basic data classification and coding for agricultural product quantity safety, the code mapping of code-type data elements is the attribute in XML Schema.

After combing the business process, relevant information model is formed, how to convert these information models into syntax-independent, data element-composed data exchange format, can be illustrated in the following two cases:

5.2.1 Relevant data elements are readily available

If a data element directory has been established in specific areas of agricultural product quantity safety basic data sharing, related data elements in the directory should be used with a priority, for the standardized processing of data in the information model, so as to form data element-based, syntax-independent data exchange format with mapping to any syntax corresponding to the information model.

5.2.2 Relevant data elements are not readily available

If there is no established data element directory, or there are no reusable data elements in the established data element directory to meet the needs of data exchange, you need to extract data elements based on the information model, in order to develop a data element directory in the field. On the basis of the directory, as described in 5.2.1, further form data element-based, syntax-independent data exchange format with mapping to any syntax corresponding to the information model.

In accordance with the mapping rules provided by this specification, the data element-based, syntax-independent data exchange format is mapped to XML Schema, and finally will form XML Schema that can achieve unambiguous exchange between homogeneous or heterogeneous systems.

5.3 The composition of data exchange format

Data exchange format consists of three parts, including head of document, body of document and end of document.

- Head of document includes document identification information, such as: document number, document date, role information and so on.
- Body of document includes the specific business information of data exchange format, such as: the registered data content.
- End of document is some descriptive information, which can be omitted sometimes.
  Due to space limitations, the XML mapping rules of data exchange format are described in detail in the article.

6. Summary

To ensure the research of agricultural product quantity safety management system, the data accuracy, timeliness and effectiveness of database system must be ensured, for data quality determines the success or system. As already mentioned, due to large volumes of data and a wide range of data sources, data conversion is the basic work. In order to ensure the quality of the data, it is necessary to establish a data conversion specification in line with the requirements of agricultural product quantity safety management system. This specification includes the definition of data formats, data conversion functions and data conversion processes, is one of the most important work of the system database. The same can be said that the data specification will directly affect the statistical analysis, model operations and decision-making results of the system. The data conversion specification research is a useful attempt to use modern information technology management methods in the research field of agricultural macroeconomic policy-making system, and is also an inevitable trend of the continuous development of modern agricultural information system R&D work.

The development of data conversion specification itself is an important part of the research of agricultural product quantity safety management system, which is basic research. In the previous study of agricultural information technology, very few norms and standards were involved; therefore, the study of norms and standards is a new attempt. In addition to data conversion specifications, we have also studied and established data access and data processing specifications, the establishment of which not only improved the level of agricultural information technology research, but also created a better development environment for the study of agricultural information system, so that agriculture information technology research was carried out in strict norms, promoting the standardization, industrialization and modernization of agricultural information system research. It can be said that the establishment of various norms have injected new content into the development of agricultural information technology research, with such norms and standards, agricultural information technology will be able to better serve macroeconomic policy analysis, serve agricultural production and life of the peasants, as well as serve agricultural modernization in China.

Acknowledgement

References

1. Zhang Chi, "Study of the XML-based Data Exchange Technology of Heterogeneous Databases". Computer Knowledge and Technology, 2008.3 (8).
2. Geng Xiangyi, "XML-based Tutorial", Beijing: Tsinghua University Press, 2006.
3. Du Shaomin, "The Design and Implementation of XML-based Data Conversion Tool for Heterogeneous Databases". Master's degree thesis of Shenyang Polytechnic University, 2009.