The Standard of Data Quality Control Technology Based on the Share of Rural Science and Technology Data

Dan Wang, Xiaorong Yang, Jian Ma, Yang Sun

The standard of data quality control technology based on the share of rural science and technology data

Dan Wang[1,a] Xiaorong Yang[1,b] Jian Ma[1,c] Yang Sun[1,d]

[1] Institute of Agricultural Information，Chinese Academy of Agricultural Sciences，Beijing 100081，China;

[2] Key Laboratory of Agricultural Information Service Technology（2006-2010）, Ministry of Agriculture, The People‘s Republic of China

[a]wangdan01@caas.cn, [b]yangxiaorong@caas.cn, [c]majian@caas.cn, [d]sunyang@caas.cn

Abstract. The standard of data quality control technology based on the share of rural science and technology data is one of the important standards to share network information source. In the paper, data quality control system is presented. The technical specification of data quality control in the data collection, data input, subject indexing, data storage construction, data description and data unit was prescribed, too. At last, it produces workflow and estimate index of data quality control.

Keywords: rural science and technology data, data quality control, technical specification, data sharing

1 Introduction

The construction of rural science and technology data sharing platform (referred to as a Shared Platform) is an important part of agricultural informatization, which is an effective way to solve rural technology "islands of information" and "last mile" problem and the service to “agriculture, rural areas and farmers”. “Shared Platform” has a variety of data types and a huge amount of data, creating a standard of data quality control technology based on the share of rural science and technology data to effectively control the quality of the data is very necessary in the process of data collection, processing and retrieval using.

2 “Shared Platform” data quality control system

“Shared platform” data quality control system includes two parts of data quality control management specifications and technical specifications.

Data quality control management department is composed by the quality control system administration and quality control technology group. The former is mainly responsible for organization and management of quality control standards. The latter is mainly responsible for the technical specifications of quality control of data.

"Quality control technical specifications" is the main indicator and basis for controlling rural science data quality, its main contents are as follows:

2.1 Quality control of data collection

In the data collection process data quality control should be noticed from a few aspects about the scientificity and practicability and use effect of the data system structure and data content, and should follow the following technical specifications:

1）Numbers and text collected must be accurate, the accuracy of data must reach 99.5%；

2）Document data must be complete, detailed and accurate；

3）Numerical data are accurate, standardized, uniform units of measurement, and facilitate statistical analysis；

4）Data quality of graphics and image includes a measurement error (error of the system, operator error, accidental error) Mapping errors, digital error, editing error of correction and analysis, data conversion error is less than 0.5%.

2.2 Quality control of data entry

Comply with national standards of existing data entry. For non-standard input， professional department is responsible for data entry rules. For data sources involved by data of multi-sector should be unified using a sector’s data. In the data entry process we should timely inspect the quality of data entry, at the same time, also should take the following quality assurance measures:

1）Entry personnel qualification: Input rate of 100 words per minute or so, bit error rate is less than five out of 10,000; Strictly obey the input rules and operational procedures;

2）In accordance with the requirements of entry indicators, if the program can be used to control data entry, entry control procedures must be used;

3）Strengthen proofreading, improve the quality of data entry (manual re-recorded proof method, machine proofreading method);

4）Select to ensure the accuracy of the input device and related information carrier;

5）There is a perfect data management and quality inspection system;

6）Data on the scan input should check and conform to the degree of the real data, requiring graphics and images are clear, identification accuracy is above 99.5%.

2.3 Data Classification and Indexing Quality Control

Classification indexing is a method of classification knowledge base on level enumerating method. It can more fully reflect the whole knowledge and its inner logical relationship. It is systematic, family of knowledge retrieval ability and enlarge/shrinkage function. "Shared platform" data classification indexing quality control to follow these guidelines:

1）Since rural science and technology data resource type is various, content is wide, there is a variety of classification indexing scheme, but must provide a classification scheme choice, that is as "rural science and technology data classification and code" technical specification as the basis of classification indexing;
2）Categories and classification codes of indexing object are mandatory fields, must be marked with the contents (values), allows duplicate content to appears (multi-value) with duplicate content between the “:” character.

2.4 Subject indexing of data quality control

To ensure the effect of "Shared platform" data retrieval, data subject indexing is indispensable, the following subject indexing specification is adopted:

2.4.1 Subject indexing method

1）Subject indexing tool is selected. For example, the “Chinese Library Classification”, “Agricultural Professional Classification”, “Thesaurus of Agricultural Sciences” are as the word choice basis of subject indexing;
2）Strict enforcement word choice process of indexing work, word choice must be standard and accurate；
3）Fully considering overall and specificity of theme analysis, the maximum to meet the requirements of precision and recall；
4）Subject factors (the research object, material, method, process and conditions)，common factor, location factor, time factor and data types constitute subject factors，we carry out the subject indexing from the above 5 factors;
5）Data of graphics, images and charts are indexed keywords which are connected in order to retrieving. Keyword choice must be standardized.

2.4.2 Subject indexing rules

1）Objectively reflect the information content, avoid introducing indexing personnel's personal views；
2）Keywords selected come from the thesaurus. Synonym keywords need to be converted into formal synonyms. Informal word cannot be used as indexing words；
3）When there is no specific corresponding keywords, you should choose the most directly related to a few keywords equipping indexing;
4）If the combination of indexing are still unable to meet the requirements, you should select the most direct hypernym Indexing;
5）If hypernym still not appropriate, you can use keywords indexing (free words);
6）Subject indexing depth is generally seven keywords (mean value), general indexing word are 5-10 keywords.

2.5 Subject indexing of data quality control

2.5.1 The data dictionary

To facilitate data sharing, improve the efficiency of data use and development, reduce development costs, "Sharing Platform" need to build the data dictionary of rural science and technology to regulate data storage structure. Data dictionary should include the following:

Data Item Name	Explanation
Data System Name	Name of data systems, such as rural science and technology data sharing system
The database name	The name of the database file
Data Name	The name of the data item
Data store name	The name of the data field
The data type	The type of data, such as digital type, character type, date type, etc.
The length of the data storage	Data stored in the computer space represented by byte
Unit	Measurement Units of data (unit of measurement)
Code description	the use of code system and coding rule
Precision	Effective number of minimum digit position
The lower limit of data	The reasonable lower limit of data
The upper limit of data	The reasonable upper limit of data
Access to the means of data	Data measurement methods or reference sources
Time and/or the environment	The time of get data and / or the environment
Remark	To add other instructions

2.5.2 Data structure specification

“Sharing platform” of agricultural science and technology data resource involves science and technology data of agriculture, forestry, water conservancy, meteorology and other fields. Data types have papers, books, journals, meetings, news, newspapers, patent information, policies and regulations, standards, non-book materials, etc. Data type is very complex, you should follow the following technical specifications in the design of the database’s data structure:

1）Follow the international and domestic existing metadata standards；Or follow the Dublin core element set；Or follow the MARC standard (USMARC, CNMARC)；Or follow the "Geographic Information - Metadata" standard, (ISO 19115:2003,

MOD, draft)；Follow the “Metadata Specification of Rural Science Data Sharing Platform” based on the above criteria.
2）In the database of "Shared platform" the required data item is defined as: title, author, subject category, information source code, keywords, description, origin, date and ID number.
3）Data structure is designed to meet the requirements of rural science and technology data sharing system. Date structure has been recognized by experts in the relevant databases. Avoid arbitrariness of database structure definition.
4）Name of the data item must be standardized, meaning must be accurate.

2.6 The technical specification of data description

GB3793-83 in "Rural science and technology data resource sharing platform" is one of the cataloging rules must be obeyed.

2.6.1 Data description of graphics and tables

There are two rules of the cataloging rules of graphics and bibliographic data:
1）For powerful database systems such as Oracle, you can put graphical information, form data directly into the field.
2）For less powerful database systems such as SQL Server, you can use the link technology, through the fields of subject (key words) associated with graphics and tables.

2.6.2 Data record consistency

1）Information source name, organization name should be consistent；
2）Units of measurement should be consistent;
3）Description of special characters should be consistent;
4）Description of numbers and dates should be consistent

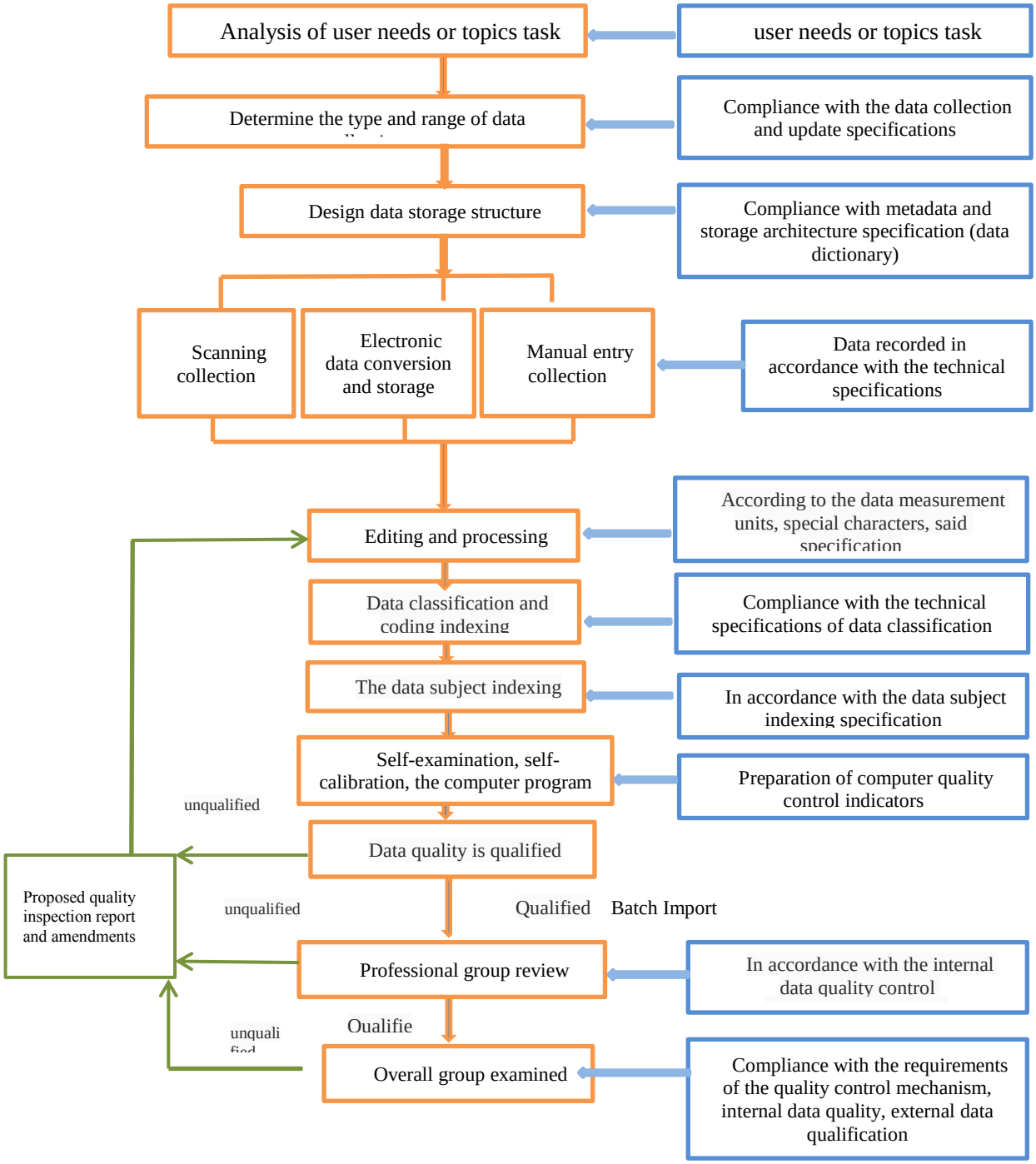
2.6.3 Data record consistency

"Sharing Platform" obey GB / T 17295-1998 international trade measurement unit code standards. Units of measurement appearing in Rural Science database must be strictly enforced this standard. Numeric database must have units of measurement instructions.

3 Data quality control as a whole

3.1 Data quality control process

Data quality control process is divided into self-examination, the quality control of a professional group and the overall group acceptance three stages. Each stage has its corresponding inspection standards, quality control flow chart below:



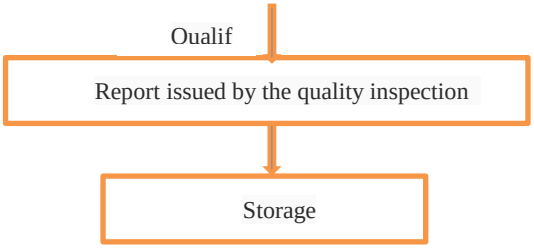


Fig. 1.

3.2 Rural Science and Technology Data quality evaluation

Order number	Evaluation of the project	Evaluation of Content	Evaluation basis	Evaluation of quantized values
1	Collection scope	1. Defining the rationality of the collection scope 2. Defining the comprehensive of the collection scope 3. Defining the maneuverability of the collection scope	➢ user requirement ➢ Project specification	5
2	Acquisition methods	1. Scan input and recognition 2. Electronic data conversion and storage 3. Manual entry collection 4. The instrument automatically collect	➢ Clarity of graphic and image ➢ The recognition accuracy of 99% ➢ Manual entry error rate is less than 0.05% ➢ To ensure that the data statistical error	15
3	The data structure	1. DC, MARC or professional metadata specifications conform to the degree 2. Required fields are complete 3. Metadata attributes are standardized	➢ Follow metadata standards ➢ Prescribed 9 required fields ➢ Eight properties of metadata	15
4	Data description	1. The degree of compliance with the data type record specification 2. Data record consistency degree 3. Science, consistency of data measurement units 4. The consistency degree of special characters	➢ Reference description standard or description in the database itself ➢ Check the fields which has requirement for statistics and retrieve ➢ In accordance with international standards of measurement ➢ Accordance with the requirements of the special characters said ➢ Description of the database	20

			itself	
5	Data classification	1. Classification system prescribed by the State 2. "Shared platform" classification specification	➢ Scientific of Category name ➢ The accuracy of classification and coding ➢ The consistency of record	10
6	The data subject indexing	1. The thesauri stipulated by the state 2. "Shared platform" to provide a scientific term 3. Comprehensiveness and technicality and subject indexing	➢ Reasonableness of the choice of words ➢ Indexing depth control for seven keywords ➢ The consistency of word choice ➢ The degree of leakage	10
7	The data content	1. Topics of data is reasonable (over-range or missing included) 2. Scientific of data content 3. Stability of data sources	➢ User needs ➢ Project mission statement Meet the qualification requirements of the data supplier	20
8	Data quality control mechanism	1. Data quality control organization 2. The organization staffing 3. Stability of data quality control personnel	➢ Whether the establishment of appropriate quality control mechanism ➢ Quality control personnel's knowledge structure is reasonable ➢ The stability of the data quality control personnel	5

4 Data quality assessment report

After the overall quality control of the data, you shall issue the corresponding data quality assessment report. The content includes: the report name, data content review, the unit responsible for the data, data quality assessment, the person completing the report, the evaluation unit, reporting to fill time.

5 Application

This paper discusses the standard of data quality control technology has been used in China Agricultural Science and Technology Information Website and China Rural Science Information Website and other large national websites. These websites' data collection, data input, subject indexing, data storage construction, data description and data unit was prescribed. After these years of operation, these websites data quality control very well, these websites ranked among the best in the field of domestic agricultural websites.

Acknowledgment

The work is supported by the special fund project for Basic Science Research Business Fee “Website of CAAS content resources organizations and service models research”, AII (No. 2014-J-008).

References

1. Zhangxiaolin. Metadata Research and Application. Beijing. Beijing Library Press,2002
2. Shaoquanqin. Several Key Issues in GIS Database Development. Acta Geographica Sinica.1995. Supplement: 34-42
3. Huajipeng. The Quality Control Method of Collecting Product Data. Applications of The Computer Systems. 2003(01): 77-79
4. Liangliping,Wuyang. Verification and Quality monitoring of Information Network Data.Chinese Journal of Hospital Statistics, 2003(03):192
5. Fang Youlin, Yang Dongqing, Tang Shiwei, Zhang Weihua, Yu Libo, Fu Qiang.Data Quality Managements in Data Warehouse. ,Computer Engineering and Applications,2003(13):1-4
6. Luo Man. Quality Control of Open Databases. China Intelligence Information, 1994(02):31