STUDY ON THE DATA QUALITY MANAGEMENT AND THE DATA QUALITY CONTROL—A CASE STUDY OF THE EARTH SYSTEM SCIENCE DATA SHARING PROJECT

Chongliang Sun*, Juanle Wang

Institute of the Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101
suncl@lreis.ac.cn

Commission II, WG VI/4

KEY WORDS: Scientific data, Sharing, Data Quality, Management, Control

ABSTRACT:

With the rapid accumulation of scientific data, there has formed mass data storage, which is ascending in every minute. So it becomes a very desiderate problem to be solved that how to manage the mass data effectively and to control the data quality scientifically. This paper analyzes several current problems in the field of scientific data management based on the author's experience on the data quality management of the Earth System Science Data Sharing Project, such as (1) the data quality problem is still not solved, (2) the absence of data living period design in data quality management, (3) the data quality management strategy always not that clear, etc. And according to the data management regulation home &abroad, the paper talks about an elementary method combining with the data classification and the quality management and some solving measures with a case study on the Earth System Science Data Sharing Project and it's soil vector data with name of *the distribution map of the national soil in a scale of 1:4,000,000(1980s)*. At last, the paper takes a prospect of the data management & quality control problem.

1. INTRODUCTION

With the globalization of information and economy, scientific data has become into the vital strategic resource for the national economy and society development. And it plays a key role in the national economy, social development, national security, and the public services, etc.

Under the condition of network's rapid development, especially the embedded work on the scientific data sharing network, the scientific data has been into massive data. And how to manage these data effectively, and control the data quality efficiently to improve the level on the scientific data production, processing, storage, sharing, and use, has been the hotspot. The paper analyzes the problem based on the work experience of data management, and put up with a resolved method to provide a way for the improvement of the scientific data sharing efficiency.

2. THE CURRENT PROBLEMS

For many years, due to the uncertainty of the scientific data[1], and the traditional data management system, it result in the data resources separated management, isolated using. So the data quality evaluation system has not formed yet. Thereby, it is difficult to exchange and share data, which limit the data use abroad in the field of inter-discipline, inter-branch, inter-territory, and the inter-industry.

Scientific data quality management refers to much content, large scope, and many domains, which include data quality itself, as well as the quality problem during the data producing process.

The paper analyzes several main problems of the data quality management, based on the long term work on the data quality control. There are, (1) the data quality problem is still need to be solved, (2) the absence of data living period design in data quality management, (3) the data quality management strategy always not that clear, etc.

3. ANALYSIS OF DATA QUALITY MANAGEMENT

Data quality refers to the accordance between the data and the impersonality state which described by the data during the process of data production, processing, storage, sharing, and data use. And to the scientific data, it can be described as the matchup degree during the above processes. The scientific data quality indexes include the data authenticity, completeness, and self-consistency. The data process quality include the transfer quality, storage quality, use quality, etc[2]. For the spatial data, the quality indexes mainly consist of data completeness, data logical consistency, data position accuracy, property accuracy, temporal accuracy and some narration of the data.[3]

Data quality management materializes in the stages of data production, storage, transfer, sharing, and the use, during which the management problem could be paid more attention in practice. And the data quality management should meet the need in the following aspects, such as the data format, data completeness, and data readability, etc.

Data format: Scientific data has the special storage format, such as the vector format, grid format, and property format, etc, which can be selected by the data characteristic.

Data completeness: scientific data's completeness mainly refers to whether the data cover the range of concrete entity

[*] Corresponding author. E-mail: suncl@lreis.ac.cn.

entirely or not[3]. The scientific data stored by a special format should be a complete entity followed by special request.

Data veracity: the scientific data, as an integral individual, should be read by the given software, and recognize the given parameters exactly, which include the accurate position information, temporal information, and the theme information, etc. For example, the soil vector data in geo-science, should be opened by the geo-software, such as the ArcGIS software or the SuperMap software, and read the several parameters from the property sheet.

In general, according to the different data type, we could divide the quality evaluation elements into two kinds, that is the Quantitative elements, and Non-quantitative elements. And the Quantitative elements include,

Completeness: superabundance, and scarcity.

Logical consistency: concept, codomain, format, topology.

Position accuracy: absolute, relative, grid.

Temporal accuracy: temporal survey accuracy, temporal consistency, temporal correctness.

Theme accuracy: classification correctness, non-quantitative property accuracy, quantitative property accuracy.

The non-quantitative elements include the data-use object, purpose, data log, etc.

The highly efficient management of scientific data depends on the high-quality management method. In this paper, to manage the data comprehensively, we introduce the concepts of completeness, superabundance and scarcity, which differ from ISO 19113 and ISO 19114, but only used in this paper and no need to elaborate them. Based on the above cognition, we analyze the management request deeply according to the management demand, and put up with a resolving method for the quality management. That is to analyse the living period of data management, and implement them into the management the data quality.

4. THE LIVING PERIOD OF DATA FOR THE DATA QUALITY MANAGEMENT

As well as any other process, the scientific data has its own living periods, which include the production, storage, use, even to the finish. For the different period, there exists different data quality discipline, so the data quality control flow would be separated into the following parts, there are data production period, data process period, data storage period, data sharing period, and the data use period. The quality control measures should be in all of the above period. The logic frame of the data quality management could be described by the fig1.

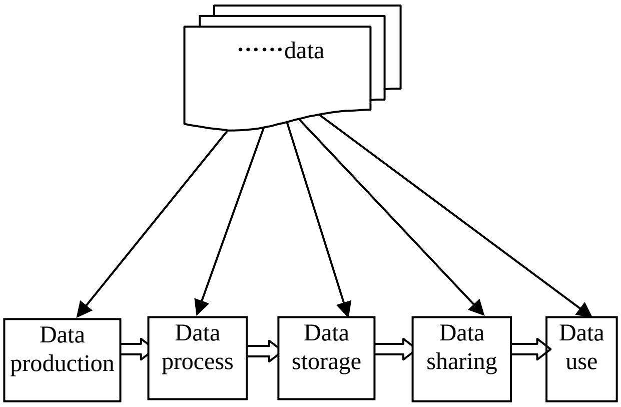

Figure 1 the logical stream of data quality management

5. THE IMPLEMENTATION OF THE DATA QUALITY MANAGEMENT STRATEGY

Given to the above analysis, the physical operation of the data quality management should include the two parts, there are the data quality evaluation method and the evaluation indexes control.

5.1 The evaluation method and flow

Due to the several types of the scientific data, and every kind has its own characteristic. So they need many evaluation methods. For the vector data, the proper method to data quality evaluation is the method based on classic mathematics. While for the property data, the right method may be the one based on the statistic mathematics, the chaos maths[4], and rough set theory[5]. But in general, the evaluation flow would be in the same control chart, and the main evaluation content consists of the followings, which are the data quality evaluation method, the evaluation operator, the evaluation date, and the evaluation result, etc.

5.2 Data quality evaluation indexes control

According to the scientific data characteristic, the data type could be divided into the vector data, grid data, property data, and the metadata during the sharing period, see fig2. So the data quality evaluation indexes should include the index of the different data types. There are the vector data evaluation index, grid data evaluation index, property data evaluation, and the metadata evaluation index, etc.

Vector/grid data indexes include the following factors, such as the storage format, project parameter, data coding, data scale, polygon closing, data document, data name criteria, data version, the last evaluation rank.

Property data evaluation indexes include the following factors, such ad the storage format, field inspection, record inspection, data document inspection, data name criteria, data version, the last evaluation rank.

Metadata evaluation indexes include the following factors, such as the metadata ID, metadata name criteria, the abstract standard, the responsibility information, the key words evaluation, and the data services mode(off-line/on-line).

The data types should be as the fig 2.

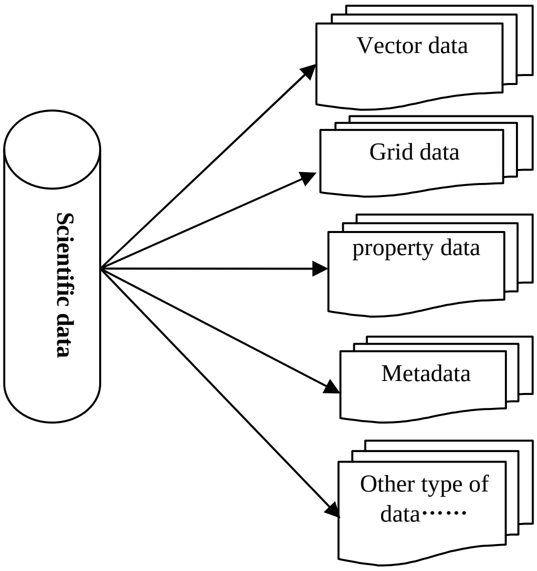

Fig 2 the classification of the scientific data

The relationship among the data evaluation indexes of different data type can be described by the fig3.

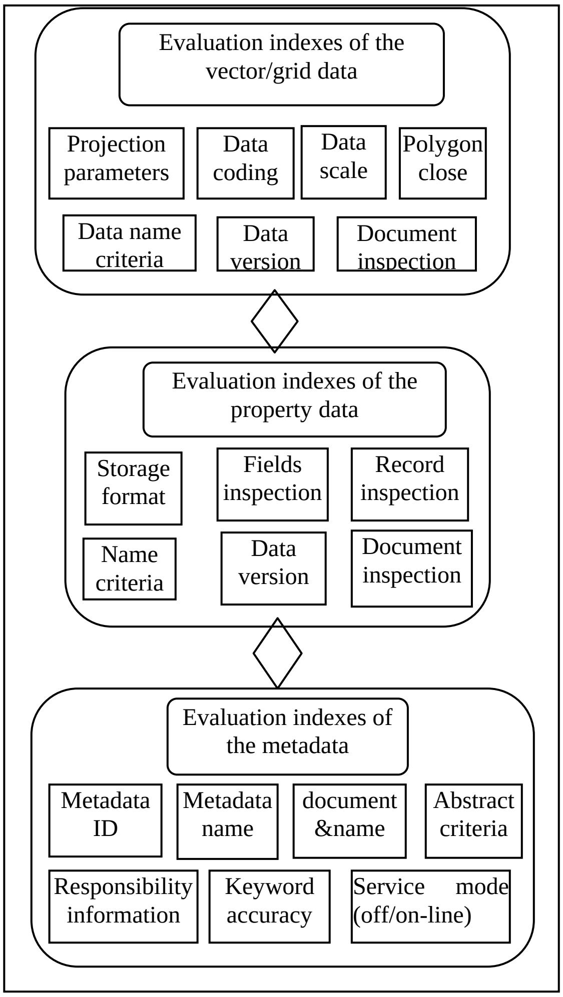

Fig 3 control index chart of the data quality evaluation

The evaluation result could be calculated by the following formula,

$$R= Value_i*Weight_i \qquad (1)$$

where $Value_i$ = evaluation value, ranging from 0-100;
$Weight_i$= evaluation weight, ranging from 0-1.

The value and weight is valued with the criterion of Data quality management guidelines from the institute of Geographic sciences and natural resources research, CAS. The result was divided into the following four classifications, as in the table 1.

Table 1 the table of evaluated value result and their ranking classification for data evaluation

value	90~100	75~89	60~74	<60
ranking	perfect	good	qualification	disqualification

6. CASE STUDY

The scientific data sharing project was put forward by the Ministry of Science and Technology(MOST), China, in 2003 to embark the data sharing work, and this project was transferred into function period in 2006. As the key one in the project, the Earth System Science Data Sharing Platform(ESSDSP) has owned more than 46000 registered users, and more than 24TB scientific data. For so large amount of data, how to manage them efficiently is a crucial problem in the construction of the platform, especially for the data management and etc.

On the above method this paper talk about, the ESSDSP data also divides into the following types, such as the vector data, grid data, property data, and the metadata, etc. According to the respective evaluation indexes, we select a soil vector data for a case study, and the data name is *the distribution map of the national soil in a scale of 1:4,000,000(1980s).* And table 2 is the soil data evaluation index value and the weight.

Table 2 the table of index value and the weight for vector data evaluation

Index	Projection parameter	Data coding	Data scale	Polygon close
value	100	100	100	98
Weight	0.3	0.15	0.1	0.2
Index	Document inspection	Name criteria	Data version	Evaluation ranking(R)
value	80	70	100	94.6
Weight	0.1	0.1	0.05	

With the above formula 1, this soil data evaluation result R is 94.6. So this soil vector data belongs to the perfect classification according to table1, and its quality control is also perfect.

7. CONCLUSIONS AND PERSPECTIVE

From this study, we can see in the current cognition level, this data evaluation method is properly for the data management, and could deal the data efficiently, which make a favourable base for the scientific data sharing project.

The future work would also be focused on the study of data management methods. And the conjunction of this method and the other method is also important for us.

References from Books:
Shi Wenzhong, 2005. *Principle of Modelling Uncertainties in*

Spatial Data and Analysis. The Science Press, Beijing, pp. 3-7.

Awrejcewicz J, 1989. Bifurcation and chaos in simple dynamical systems. Singapore: World Scientific.

Pawlak Z, 1982. Rough Sets. International Journal of Computer and Information Sciences, 11(5):341-356.

References from Other Literature:
The scientific data management group of the MOST, 2006, the data quality management discipline of the Scientific data sharing projects. Beijing, China.

The institute of Geographic sciences and natural resources research, CAS, 2009, Data quality management guidelines. Beijing, China.

References from websites:
China KDD, "About the data quality." http://www.dmresearch.net/research/shujucangku/2008/0805/124049_2.html (accessed 6 Nov. 2009)
Baidu. "data quality control of the spatial data". http://baike.baidu.com/view/125958.htm (accessed 28 Sep. 2009)

Acknowledgements and Appendix (optional)

The paper was funded by earth system science data sharing program(2005DKA32300); the National Natural Science Foundation of China(40801180、40771146); Open Research Funding Program of LGISEM; the National 863 plans projects(2006AA01A120); the forward project for youth researcher of igsnrr: study on the intelligent search for geo-metadata based on geographic ontology.