The Impact of Data Quality Tagging on Decision Complacency

InduShobha N. Chengalur-Smith
ic307@albany.edu

Donald P. Ballou
dpb67@albany.edu

Harold L. Pazer
hlp66@albany.edu

School of Business
SUNY-Albany
Albany, NY 12222

Abstract

Data tagging is the process of attaching quality indicators to fields in a database in order to make decision-makers aware of the level of data quality. The underlying hypothesis in this study is that end-users of the database would make different choices in the presence or absence of information about data quality. This paper explores different formats for data tagging and investigates the effect of data tagging on decision complacency. We find an interaction between the format in which data quality information is presented and the decision-making strategy. Our preliminary results suggest that there is a significant difference in the use of data quality information between simple and complex decision environments. We find that unless care is exercised in presenting and formatting the data quality information, it may be ignored.

Introduction

Decision-making is an inexact science, and one of the reasons is the utilization of data of imperfect quality in the decision process. Because the exigencies of decision making may well preclude obtaining ideal data, one often must make decisions in spite of the data's imperfections. Effective decision-makers can compensate for various deficiencies the data may possess, especially if the decision maker is acquainted with the data's idiosyncrasies. This intuitive knowledge is lost, however, whenever data are used by various parties for purposes other than the original, which is happening with increasing frequency, a trend manifested by the prevalence of data warehouses. Those potential users of data not possessing an intuitive knowledge of the data's qualities are forced either to accept the data as is, which implicitly assumes that all data values are equally valid, or at the other extreme to avoid using data whose quality they cannot personally attest. Unfortunately but understandably the latter option often is chosen.

Common sense affirms that knowledge regarding the quality of the data one wishes to use would be beneficial. This proposition has not, however, been rigorously examined. Furthermore, the type of information regarding the data's quality that would be most helpful to users is not known. In addition it may well be that the kind of information about data quality that is most appropriate depends upon the nature of the decision process. Moreover, the effectiveness of the information regarding data quality depends to some degree on the sophistication of the user. The purpose of this paper is to begin the exploration of issues such as these.

Background

The systematic study of the quality of data and information is a relatively recent phenomenon. Initial research was conducted by members of the accounting profession and addressed accuracy issues in accounting and financial system (Cushing [1974]). Information system researchers have approached the issue of data quality from a broader perspective. For example, their work has encompassed not only hard data such as is found

in accounting type systems but also soft data needed, for example, by marketing systems. By soft data we mean data whose quality is inherently unverifiable.

One of the findings of information systems researchers is that the most effective way to think about data quality is via attributes or dimensions such as accuracy, timeliness, completeness, consistency, reliability, and so forth. This approach was originally posited by Ballou and Pazer [1985] and was confirmed and explored in depth by Wang and Strong [1996]. Another stream of research has as its starting point the quality of an information product delivered to a user. By information product we mean anything such as a report, spreadsheet, mailing list and so forth that was generated by processing data. In general we reserve information quality for the end product delivered to users and reserve data quality to refer to the quality of raw data and intermediate results. Improving the quality of the data upon which the information product is based and/or reengineering the processing of that data can be beneficial. Ballou and Pazer [1995] examined the value of information in the context of the tradeoff between the accuracy and timeliness of data. Recently, Ballou et al [1998] introduced concepts and developed a methodology to evaluate the impact of improvements to data quality and reengineering of the processing stages to the value, cost, timeliness and overall quality of information products.

A third research stream involves material on how to enhance the data's quality. An early, statistical based approach was described by Morey [1982]. Redman [1996] has done considerable work in the area of enhancing the quality of the stored data resource. Ballou and Tayi [1989] used an optimization procedure to target which data files should receive priority for data quality enhancement. An overall view of the status of data quality research as of 1995, can be found in the survey paper by Wang, Storey and Firth [1995]. They not only describe what research has been undertaken but also indicate a variety of problems requiring the attention of researchers in this field.

Data quality metadata

Data regarding the quality of stored data can be thought of as one type of metadata. At this time there are not rigid rules as to the level of granularity for which the metadata

should apply. At one extreme it is possible to have information regarding data quality at the level of the individual data item. This could be done by including relevant data quality dimensions such as accuracy, timeliness, etc. In certain situations this is warranted and can be achieved via the use of data tags (Wang and Madnick [1990]). At the other extreme one could have data regarding the quality of an entire file or relational table.

Our approach is intermediate between these two, as we postulate a certain level of data quality for each of the attributes (data fields) made available to the decision-maker. Essentially we are asserting for the purposes of this research that the quality of all data items in a particular domain is the same. (Different domains can have different levels of quality). Although this assumption is open to challenge, it does significantly simplify storage and retrieval issues. Since data about data quality are metadata, it is natural to include such information in the data dictionary. Although at this time current implementations do not include the kind of quality information that we envision, the needed modifications would be straightforward.

An important consideration is the type of metadata made available to users, more specifically how data quality information is recorded and presented. The way information is displayed (i.e. on a verbal or numerical scale) affects decision behavior [Stone and Schkade, 1991]. Even within a numerical format, decision making is sensitive to different displays, such as easy versus difficult fractions [Johnson et al, 1988]. Thus the availability of information is not enough, as it needs to be presented in a form that will promote effective decision making.

One way to address this issue is to stipulate that data quality be measured on a zero to one continuum. The value one represents data where quality is perfect or ideal with zero capturing the other extreme. These data quality measures should be thought of or relative rather than absolute. For example, an evaluation of 0.8, say, does not necessarily imply that 80% of the data values are correct. Rather it asserts that data possessing this evaluation is better than data with an evaluation of 0.7, say. In essence the approach assumes an interval scale and was utilized in Ballou and Pazer [1995] and Ballou et al [1998].

Another approach is to incorporate an n-point ordinal scale which could be mapped into categories such as excellent, good, average, and so forth. The extreme for this approach would be a two-point ordinal scale whose values could be interpreted as "above average" and "below average". For this study we have explored the impact on decision making of data quality metadata using an interval scale and a two-point ordinal scale.

Decision making paradigms

The most effective format for presenting data quality information could be a function of the decision making process or strategy. Over the past several decades a substantial literature has been developed regarding decision making. (An overview of the field can be found in Payne et al [1993]). For the purposes of this study we have adopted two approaches from the theory regarding decision making. The first can best be described as conjunctive decision making, the second as weighted additive decision making.

Conjunctive decision making assumes that the decision depends upon a known and specified set of criteria. For each of these criteria a minimum acceptable level is established. The decision-maker must choose from among several possible options. In order to make the decision, each option must be evaluated on each of the criteria. An option is acceptable so long as the evaluation on each of the criteria is at least as large as the specified minimum for that criterion. If the evaluation is below the minimum for even one criterion, then that option is not acceptable.

Weighted additive decision making also assumes that various criteria upon which the decision is based have been identified. In addition, a weight is assigned to each criterion to capture that criterion's relative importance. (Usually the weights for all the criteria sum to one). Again the decision-maker must choose from among several options. For each option each criterion is evaluated and the resulting score is multiplied by the weight. These values are summed to produce the overall score. The option with the largest score is selected.

Both weighted additive and conjunctive strategies are alternative based strategies, but weighted additive strategies are compensatory (allow for tradeoffs among attributes), whereas conjunctive strategies are non-compensatory. Further, weighted additive strategies form an overall evaluation for each alternative based on quantitative reasoning, whereas conjunctive does not form an overall evaluation and the basis for reasoning is qualitative (i.e. simple comparison of values). Thus the two decision making strategies we have chosen to investigate, provide a contrast in terms of their properties.

Payne et al [1993] proposed that people adopt a decision-making strategy on the basis of a cost benefit framework, i.e. individuals compromise between making the best possible decision and minimizing their cognitive effort in making that decision. Thus the impact of data quality information may also depend on the level of complexity in the decision making process. Task complexity is a function of several variables such as the number of alternatives, the number of attributes, and time pressure. A selective use of information or a shift to more simplistic strategies can compensate for increases in decision complexity. Grether et al [1986] found that consumers ignore less relevant information in complex information environments.

Even though the inclusion of data quality information has become technologically feasible, the collection of such information is still a challenging undertaking. Clearly, providing data quality information on a two-point ordinal scale is more manageable than providing it on a 100-point continuum. But if the latter option leads to more effective decision-making, then the considerable effort required may be worth it. We use an experimental approach to establish whether the use of information quality is worthwhile and if so, under what circumstances.

Experimental task and design

The choice of experimental task is an important one, in order for the results of the experiment to be valid and generalizable. Instead of building a completely new task, we modified existing tasks from the literature. One was adapted from an apartment selection task developed by Payne [1976]. For our purposes, we based the site selection decision on

five criteria: parking facilities, commuting time to work, floor space, number of bedrooms, and rent expense. The task had four alternatives with five attributes each, and was designed to be a relatively simple task with the choices being clearly distinct. For the conjunctive case, minimum acceptable levels for each criterion were specified, and for each alternative the evaluation for each criterion was provided. For the weighted additive case, the instrument listed the weights for each criterion instead of the minimum acceptable levels.

The other task was developed by Jarvenpaa [1989], requiring subjects to select a site for a restaurant from a set of candidate sites, on the basis of pre-specified criteria. This was modified to be a relatively complex task, with six alternatives and seven attributes per alternative. Each alternative was evaluated on the following seven criteria: area retail sales, traffic density, competition, average family income, land and building costs, population density, and population growth. The same type of information was provided to the test subjects as with the simple task. However, here the choices were more prone to interpretation, and this task had over twice the number of cells ($7\times6=42$) as compared to the simpler task ($4\times5=20$).

The tasks without data quality information had a preferred solution and hence provide objective evaluation measures. We used them as the base cases initially and then added a quality dimension to the criteria. The inclusion of data quality information made the ranking of the alternatives a subjective assessment, since there was no longer a demonstrably correct solution. In order to avoid confounding computational errors with judgments of alternatives, the weighted additive strategy was slightly modified by displaying the product of multiplying the weight by the value of the attribute. However, the operation was not completed by adding up the weighted scores and presenting the sum, since the primary interest was in finding out how decision makers incorporate the data quality information that was presented alongside.

The experiment was designed to elicit information about different decisions arising from differences in the data quality format for a given decision strategy. The three factors in the experiment are decision strategy (conjunctive or weighted additive), task (simple or

complex), and data quality information format. We developed twelve different instruments. The six instruments associated with the simple task all had the same set of evaluation scores. The same statement applies to the complex task. All the instruments were pre-tested on School of Business faculty and graduate students. Found in the appendix is the instrument of the simple task, weighted additive with interval scale information regarding the quality of the evaluations.

The experiment used undergraduate School of Business students enrolled in a senior level course as subjects. These students had previously taken course work that included decision support systems, systems analysis and design, data base management, simulation and statistics. They were randomly assigned to six groups. Each group was randomly assigned to one of the two decision making strategies. Thus each subject performed first the simple task followed by the complex task, using the same decision making strategy. We had a randomly chosen group of subjects perform the task with no data quality information and other subjects were randomly assigned to perform the same task with one of the two data quality formats. Thus they got the same data values (evaluation scores, minimum criteria, criterion weights) but in addition they had information about the quality of the evaluation scores. For the two-point ordinal case, the evaluation for each criterion was rated "above average" or "below average". Other subjects received data quality information in the form of interval scale values consisting of integers from 0 to 100. The actual values were chosen randomly but a stratified approach was utilized to ensure that some of the criterion evaluations would be reliable and that others would be relatively unreliable.

Hypotheses

If including data quality information did not make a difference, the overall rankings of the sites would be the same under all three formats: no data quality information, two-point ordinal, and interval scale data quality information. When making a choice, however, decision-makers would be most interested in the top ranked alternative. Thus, rather than overall consistency, our focus is on the top choice. In particular, we focus on the issue of complacency. *Complacency* is a measure of the degree to which data quality information is

ignored. Knowledge of the data's quality may well affect one's selection of the preferred site. We design a set of hypotheses to determine if the number one choice changes if data quality information is provided. The following set of four hypotheses is designed to determine if the selection of the preferred site does indeed change.

H1A: Including data quality information changes the number of times the originally preferred site continues to be ranked the top site for the simple conjunctive task.
H1B: Including data quality information changes the number of times the originally preferred site continues to be ranked the top site for the simple linear task.
H1C: Including data quality information changes the number of times the originally preferred site continues to be ranked the top site for the complex conjunctive task.
H1D: Including data quality information changes the number of times the originally preferred site continues to be ranked the top site for the complex linear task.

A chi-squared statistic that keeps track of the changes in the number of times the originally preferred site is ranked the highest, under different data quality formats, is used to test this hypothesis. If, when groups with and without data quality information are compared, little change is noted in the proportion selecting the originally preferred site (resulting in a low value for the Chi-Square statistic) this is evidence of complacency.

Tables 1 (a) & (b) about here

As can be seen from Table 1(a), the level of complacency varied dramatically across the research design. The smallest level of complacency, corresponding to the greatest impact of data quality information, existed for the simple task when groups with interval scaled data quality information were compared to groups with no data quality information. This was true for both the conjunctive and the weighted additive tasks.

Table 1(b) presents the results for the most significant of these cases (None to Interval, Simple, Weighted Additive). Nine of the eleven members of the group with no data quality information chose Site 2 as the preferred choice. This site had a major relative advantage in terms of *commuting time*. The group with the interval scaled data quality information had the opportunity to observe that *commuting time* was the least reliable of the data inputs. Consequently only two of the eleven members of this group exhibited complacency and selected this originally preferred site.

Implications

Recalling that Complacency is a measure of the degree to which data quality information is ignored, one of the more interesting results of our study was the degree to which complacency varied across the research design. Table 1(a) presented these results of twelve paired comparisons between different groups of respondents.

When the data are subdivided by task it is observed that for the simple task, Complacency was rejected beyond the .001 level for three of the six comparisons. By contrast for the complex task Complacency was rejected in only one of the six comparisons at this level. This observed difference is primarily due to differences for the None to Interval comparisons for the two levels of complexity. For the simple task this comparison was significant at the .001 level for both the conjunctive and the weighted additive approaches while for the complex tasks the comparisons were clearly not significant for either approach.

Since the interval format is clearly the one which presents the most detailed information concerning data quality, it is not surprising that this detailed information would be more fully utilized in a task which is otherwise relatively simple rather than in one which already has a substantial level of complexity. While the results are not conclusive, they are strongly suggestive of an impact of information overload and suggest that providing too detailed information concerning data quality may be counter-productive in more complex decision environments.

The degree of complacency also differed substantially between the conjunctive and weighted additive approaches. Complacency was rejected at the .001 level for three of the six weighted additive comparisons but for only one of six comparisons when the conjunctive approach was employed. Once again information overload is a possible explanation. Since, for the weighted additive, the rating times weight computation was pre-performed and presented as the weighted score for each criterion, it may have been easier for the respondents to focus on one or two key criteria that exhibited the greatest impact. For the conjunctive approach, the criteria are implicitly equally weighted consequently it may have been more difficult to develop a clear view of the impact of the data quality information.

It is important to recall that the subjects of this experiment were seniors in an undergraduate MIS program. Information overload may be a relative concept. What is considered too complex and information intensive to undergraduates accustomed to simplified textbook examples and exercises may seem less daunting to analysts and decision makers accustomed to real world complexities.

In summary, the format of presentation of data quality information seems important with an indication that when a more complex format is used in a more complex environment, information overload may occur. Further research is currently underway to expand upon these results using more empirical experiments. Another possibility is to leverage the Internet to expand the scope and scale of participants in the experiment. A final point that remains to be addressed is the method by which data quality tagging could be implemented in data warehouses.

References

Ballou, D. P and Pazer, H. L (1985) "Modeling data and process quality multi-input multi-output information systems," *Management Science*, Vol. 31, No. 2, pp. 150-162.

Ballou, D. P, Pazer, H. L (1995) "Designing information systems to optimize the accuracy-timeliness trade off," *Information Systems Research*, Vol. 6, No. 1, pp. 51-72.

Ballou, D. P and Tayi, G. K. (1989) "Methodology for allocating resources for data quality enhancement," *Communications of the ACM*, Vol. 32, No. 3, pp. 320-329.

Ballou, D. P, Wang, R. Y, Pazer, H. L and Tayi, G. K. (1998) "Modeling data manufacturing systems to determine data product quality," *Management Science* (forthcoming).

Cushing, B. E. (1974) "A mathematical approach to the analysis and design of internal control systems," *Accounting Review*, Vol. 49, No. 1, pp. 24-41.

Grether, D. M., Schwartz, A. and Wilde, L. L. (1986) "The irrelevance of information overload: An analysis of search and disclosure," *Southern California Law Review*, Vol. 59, pp. 277-303.

Jarvenpaa, S. L. (1989) "The effect of task demands and graphical format on information processing strategies," *Management Science*, Vol. 35, No. 3, March, pp. 285 - 303.

Johnson, E. J., Payne, J. W. and Bettman, J. R. (1988) "Information displays and preference reversals," *Organizational Behavior and Human Decision Processes*, Vol. 42, pp. 1-21.

Morey, R. C. (1982) "Estimating and improving the quality of information in a MIS," *Communications of the ACM*, Vol. 25, No. 5, pp. 337-342.

Payne, J. W. (1976) "Task complexity and contingent processing in decision making: An information search and protocol analysis," *Organizational Behavior and Human Performance*, Vol. 16, pp. 366-387.

Payne, J. W., Bettman, J. R. and Johnson, E. J. (1993) "The adaptive decision maker," Cambridge University Press.

Redman, T. C. (1996) "*Data quality for the information age*," Artech House, Boston MA.

Stone, D. N. and Schkade, D. A. (1991) "Numeric and linguistic information representation in multivariate choice," *Organizational Behavior and Human Decision Processes*, Vol. 49, pp. 42-59.

Wang, R. Y. and Madnick, S. E. (1990) "A polygon model for heterogeneous database systems: The source tagging perspective," *Proceedings of the 16th International Conference on Very Large Databases*, pp. 519-538, Brisbane, Australia.

Wang, R. Y., Storey, V. C. and Firth, C. P. (1995) "A framework for analysis of data quality research," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 7, No. 4, August, pp. 623-640.

Wang, R. Y. and Strong, D. M. (1996) "Beyond accuracy: What data quality means to data consumers," *Journal of Management Information Systems*, Vol. 12, No. 4, pp. 5-34.

Table 1 (a): Complacency

	Data Quality Information	*Conjunctive*	*Weighted Additive*
Simple (4 choices)	None to Ordinal	$\chi^2 = 1.73$ p = 0.1879	$\chi^2 = 18.96$ *** p = 0.0000
	None to Interval	$\chi^2 = 21.23$ *** p = 0.0000	$\chi^2 = 29.94$ *** p = 0.0000
	Ordinal to Interval	$\chi^2 = 8.13$ ** p = 0.0044	$\chi^2 = 1.14$ p = 0.2864
Complex (6 choices)	None to Ordinal	$\chi^2 = 2.79$ p = 0.0951	$\chi^2 = 13.78$ *** p = 0.0002
	None to Interval	$\chi^2 = 1.24$ p = 0.2658	$\chi^2 = 0.46$ p = 0.4984
	Ordinal to Interval	$\chi^2 = 0.36$ p = 0.5479	$\chi^2 = 8.76$ ** p = 0.0031

Table 1(b). Most significant: (None to Interval, Simple, Weighted additive)

Site	*Interval*	*None (Expected)*	$(F-E)^2/E$
2	2	9	5.44
Others	9	2	24.5
Chi-squared statistic (1 df)			29.94

Appendix: Sample of Task (Simple, Weighted Additive, Interval)

Your job requires you to move to a new city. You have a friend who lives there and you request her help in locating an apartment. You provide a list of criteria that are important to you and you have weights in mind for each criteria, which reflect their relative importance to you. Your friend gathers information about 4 potential apartment complexes and passes it along to you. She scores each apartment complex on each factor on a 50-point scale, such that a higher number is always more desirable. For example, a rating of 40 for rent expense is more desirable than a rating of 30.

Your objective is to choose the complex that overall performs the best. However, you realize that the data she obtained may not be completely accurate. For instance, she estimated commuting time by looking at the map. Also, the complex managers indicated that the rent quoted could increase at any time.

You decide to incorporate this uncertainty into your decision making process by using a 0-100 reliability measure where a score of 100 indicates perfectly reliable data and 0 scores imply completely unreliable data. The following tables display the reliability of the information provided about each criterion and the weights assigned to each criterion.

Criterion	*Reliability*	*Criterion*	*Weight*
Parking facilities	57	Parking facilities	1
Commuting time to work	23	Commuting time to work	2.5
Floor space	76	Floor space	2
Number of bedrooms	68	Number of bedrooms	1.5
Rent expense	44	Rent expense	3

After multiplying the ratings by the weights for each criterion, you obtained the following results. The weighted scores are shown on the next page (for example, a score of 70 for commuting time is the result of multiplying its rating of 28 by its weight of 2.5).

Given the weighted scores, the objective is to choose the apartment complex that overall is the best (has the highest overall sum). Rank the apartment complexes in order of preference with 1 corresponding to the complex you would most prefer and 4 to the one you would least prefer. (Use all the information given to break any ties.) Recall that the reliability refers to the data and not to the weights. Next to each complex write its rank, along with a brief explanation of how you arrived at the rank.

A

Criterion	*Reliability*	*Rating*	*Weight*	*Weighted scores*
Parking facilities	57	22	1	22
Commuting time	23	28	2.5	70
Floor space	76	20	2	40
# of bedrooms	68	32	1.5	48
Rent expense	44	40	3	120

Rank = ___

Explanation:

B

Criterion	*Reliability*	*Rating*	*Weight*	*Weighted scores*
Parking facilities	57	25	1	25
Commuting time	23	32	2.5	80
Floor space	76	31	2	62
# of bedrooms	68	36	1.5	54
Rent expense	44	36	3	108

Rank = ___

Explanation:

C

Criterion	*Reliability*	*Rating*	*Weight*	*Weighted scores*
Parking facilities	57	28	1	28
Commuting time	23	27	2.5	67.5
Floor space	76	33	2	66
# of bedrooms	68	29	1.5	43.5
Rent expense	44	26	3	78

Rank = ___

Explanation:

D

Criterion	*Reliability*	*Rating*	*Weight*	*Weighted scores*
Parking facilities	57	27	1	27
Commuting time	23	25	2.5	62.5
Floor space	76	35	2	70
# of bedrooms	68	38	1.5	57
Rent expense	44	37	3	111

***Rank* = ___**

Explanation: