8-2010

# Methods to Measure Importance of Data Attributes to Consumers of Information Products

Christopher Heien
*University of Arkansas - Little Rock*, chheien@ualr.edu

Ningning Wu
*University of Arkansas - Little Rock*, nxwu@ualr.edu

John Talburt
*University of Arkansas - Little Rock*, jrtalburt@ualr.edu

# Methods to Measure Importance of Data Attributes to Consumers of Information Products

**Christopher Heien**
University of Arkansas, Little Rock
chheien@ualr.edu

**Ningning Wu**
University of Arkansas, Little Rock
nxwu@ualr.edu

**John Talburt**
University of Arkansas, Little Rock
jrtalburt@ualr.edu

**ABSTRACT**

Errors in data sources of information product (IP) manufacturing systems can degrade overall IP quality as perceived by consumers. Data defects from inputs propagate throughout the IP manufacturing process. Information Quality (IQ) research has focused on improving the quality of inputs to mitigate error propagation and ensure an IP will be fit for use by consumers. However, the feedback loop from IP consumers to IP producers is often incomplete since the overall quality of the IP is not based solely on quality of inputs but rather by the IP's fitness for use as a whole. It remains uncertain that high quality inputs directly correlate to a high quality IP. The methods proposed in this paper investigate the effects of intentionally decreasing, or disrupting, quality of inputs, measuring the consumers' evaluations as compared to an undisrupted IP, and proposes scenarios illustrating the advantage of these methods over traditional survey methods. Fitness for use may then be increased using those attributes deemed "important" by consumers in future IP revisions.

**Keywords**

Attribute importance, information product, consumer feedback

**BACKGROUND**

A continuing area of research in Information Quality (IQ) is tracking aftermarket user satisfaction, or as is traditionally defined, "fitness for use," of delivered information products to consumers. [5] Methods have been developed to improve the quality of outputs by focusing on inputs [1] [3], associating quality measurements with quality dimensions [7] [10], and determining optimal paths for these inputs in order to create a high quality output. [2][3] Methods to measure user satisfaction of the resulting IP have also been explored and techniques such as information quality assessments and surveying have been developed. [6][7][8].

IQ surveys and auditing tools seek to assign objective quality ratings through subjective evaluation of an IP from the consumer's point of view. These surveys may give insight into the quality of an output, but are costly in both time and resources [2], possibly inaccurate given the myriad of interpretation for quality measurements by consumers, and must be carried out on a voluntary basis. Consumers must be willing to commit quality feedback. Surveys can incorrectly represent a consumer's true opinion, most often in cases where the survey seems irrelevant, tedious, or not critically important to the consumer's continued consumption of the IP. Surveys themselves can be considered an IP and are therefore susceptible to possible data quality problems of their own.

The disruption of an IP's attributes can act as a "wake-up call" to consumers engaging in IP audits and survey. These disruptions may be applied. [9] However, implementing the methods discussed in this research may prove problematic given the very act of introducing and delivering disrupted IPs within an organization. Implementation in Academia is also problematic since this "wake-up call" effect from data disruption would need to be divulged to those surveyed, thereby nullifying the very effect of the disruption as consumers would be on the lookout for low quality attributes. Methods will be generalized and focus on use cases. Future research may seek to solve difficulty in application of these methods.

### Problem

When an individual commits a crime they become a criminal. However, they remain a law-abiding citizen only until such times as their crime is revealed. A criminal is made only when one fails at crime, is apprehended, and whose story is made public. However, most of the successful criminals are never detected, so there is no knowledge, and much less a news report, of their continuing success as a law-abiding citizen.

This analogy relates the same "wake-up" quality analysts and organizations receive in the aftermath of data quality disasters. [4] A small incorrect value in a database may one day result in a catastrophe, but until the value is known to be incorrect it will remain in the database, assumed to be correct and continually being used with no resulting adverse consequences.

The detection of one poor quality attribute may lead to a chain effect revealing additional poor quality attributes, just as an investigation into a crime may reveal a criminal's additional crimes. Every law-abiding individual may have committed at least one offense in a lifetime, however to interrogate the world's population is simply not an option due to costs, individual rights, and acceptance thresholds defining "acceptable" crimes. This example is analogous to data quality improvement costs, data governance, and acceptable fitness for use thresholds in the IQ context respectively. The world's population cannot be randomly screened; however, in the IQ sense, data attributes (analogous to people) may be disrupted to a lower quality (analogous to charging of a crime) in order to check if the entire IP (person accused) contains additional crimes. This describes the "wake-up" effect, similar to a serial killer (catastrophic IQ disaster) being ultimately apprehended (data correction) for a simple speeding offense (data quality disruption).

### PROPOSED SOLUTION

An existing IP consisting of perceived high quality inputs can be disrupted by the replacement or injection of lower quality attributes prior to its construction. The IP manufacturer is aware of this lowering of quality, but the consumer is not, and the consumer is expected to respond to these injections of poor data quality attributes. The IP manufacturer can then trace responses and corrections by consumers to their sources to determine if intentionally disrupted, or even undisrupted data, has been corrected. Intentionally disrupted values can have a "trigger" on the consumer. Consumers may detect one attribute to correct that was intentionally disrupted, but may review the IP products as a whole to determine other values to be incorrect of which the manufacturer was unaware.

### Hypothesis

Research and methodology will support the hypothesis if consumers of an IP product will respond and review more data attributes on a final IP if an attribute that is known to be of high value to the consumer is disrupted to a lower quality. The hypothesis will be rejected if users fail to respond to, or correct, attributes presented on a delivered IP, even when attributes assumed to be important to the consumers are changed to incorrect or low quality values. The null hypothesis will hold if the disruption of attributes shows no change in response between the original IP and disrupted IP.

### Description of Quality Measurements

Traditional data quality measurements [10] are established prior to importance ratings of attributes. These measurements rate the quality of each attribute according to an appropriate data quality metric. The following table gives the notation used.

| Notation | Description |
|---|---|
| P | An information product manufactured by an organization |
| $A=\{a_1, a_2,\ldots, a_n\}$ | Set of n data attributes used to manufacture P |
| $M \subseteq A$ | The subset of attributes that the manufacturer of P considers most critical to its effectiveness |
| $C \subseteq A$ | The subset of attributes that the consumers of P consider most critical to its effectiveness |
| R | A discrete set of k ordered quality ratings, $R=\{r_1, r_2,\ldots, r_k\}$, $r_1 < r_2 <\ldots< r_k$ |
| $Q: A \rightarrow R$ | A function that assigns a data quality rating to each attribute of A , $Q(a_i) = r_{ij}$ |

**Table 1: IP definitions and functions**

It is not vital that each attribute share the same DQ metric, only that an initial rating must be established to show the data quality disruption is effective in lowering the quality rating.
Take as an example the case where P represents the names and addresses used to mail offers to customers.

| Notation | Description |
|---|---|
| P | Customer mailing |
| A | {last name, street address, city, state, zip code} |
| M | {street address, zip code}The manufacturer is primarily interested in deliverability and considers correct street address – zip code combinations to be most critical, or important, for that objective |
| C | {last name, street address} The consumer wants to be mailed pieces with his or her correct name and street address. An incorrect name may result in poor manufacturer - customer relations. |
| R | {Low, Medium, High} |

**Table 2: Example where P is customer mailing**

In the following example, we will assume a quality rating Q has been established for zip code accuracy and can result in any one of the R values above given the following mapping:

r1 = [0.0 – 33.3] (Low Accuracy)

r2 = (33.3 – 66.6] (Medium Accuracy)

r3 = (66.6 – 100.0] (High Accuracy)

We assume a data quality evaluation of zip code accuracy has taken place map it to a member of R:

Q(zip code) = 50% = r2

This calculation uses function Q to determine accuracy rating r of an attribute $a_i$. It is not important in this scope to explicitly define the algorithm of Q, only that a quality rating is achieved. Quality ratings for r, $r_j$, represent all possible outcomes for a quality assessment for r, defined by a quality definition and measure. [10] All measurements are bucketed in k number of bins, a discrete mapping of quality metric ranges into a single aggregate value to simplify analysis. In this example, accuracy measurements for $a_i$ are bucketed for three discrete values, j = {Low, Medium, High} denoted by $r_1$, $r_2$, and $r_3$ respectively for $a_1, a_2, a_3$.

In this example, the set of critical attributes M are those chosen from A by the manufacturer as most important in fulfilling the goal of delivery. City and state are not included in M since these attributes may be determined by zip codes; however they are still maintained in P in cases of incorrect zip code or when additional address resolution is required. It is also assumed last names will have no bearing on the success of delivery; however it is important to note this as an assumption. There may be cases, or unexplored view of the attribute combinations, that provides even higher accuracy zip code calculations given last name. This may include a neighborhood postman's personal knowledge of knowing when a family changes address, yet these special cases will not be represented.

The set C contains name and street address, those values assumed by manufacturers to be most important to consumer of the mailing. A misspelled name may hinder the manufacturer and consumer relationship, in addition to other cases of name change in the case of marriage, divorce, or other event the manufacturer may fail to reflect in the mailing. The accuracy of names is not important in M, since the primary fitness for use is deliverability to an address. Name accuracy is of much higher importance once received, and represents a social correspondence of the manufacturer to the consumer.

P is formed from a subset of available attributes A. A may contain attributes of mixed quality ratings, a combination of low, medium, and high. However, it remains vital to note that these values are determined by the manufacturers' predefined quality functions Q for each attribute, not directly by the consumer and do not necessarily represent final usefulness of the IP. The manufacturer's assumption as to the contents of both C and M must be established, as was in Table 2, along with quality ratings for each member of A. Once the baseline attributes "assumed" important by manufacturers, M and C, are defined and R values (quality ratings) have been established for each attribute A, importance ratings can be tested through data disruption. This method seeks to determine if the attributes assumed to be of high importance to the manufacturer M, street address and zip code in this case, are in fact the attributes most desirable to be of high quality in P. Additionally, the manufacturer can determine if the assumption as to the contents of C are also valid. Attributes contained in both M and C should have precedence for quality improvement over all other attributes in A. However, it is not always clear as to which attributes belong to M and C. Increasing quality of attributes of no importance has no positive bearing on P. Failure to identify attributes that should be contained in M or C show loss of possible fitness for use improvement of P.

#### Description of Importance Ratings

The following method will analyze one attribute's pass through data disruption, or decrease in quality, as needed for testing for importance by the manufacturer. The actual importance to consumers is implemented as a separate value, since both manufacturers and consumers have a different view of an IP's fitness for use, relative in the above example to the importance of a manufacturer's delivery of a mailing and the importance of a correctly spelled name to a consumer.

Importance ratings of each member of set A will be surveyed, delivered as a whole in information product P. In a real world application, P may include representation quality aspects, defined by the physical form in which the P is delivered to the consumers (paper or electronic viewing, report layout, visual aesthetics). These representational aspects will not be accounted for in the scope proposed methods.

C is assumed to be in constant fluctuation, as the manufacturer may have little insight into those attributes important to consumers; this represents the IQ problems of determining a proper feedback function after delivery of an information product. M is assumed more accurate since the manufacturer is the producer of P, and knows what attributes are necessary to fulfill fitness of use for M. Both C and M may be revised during IP production. An optimal P would have M = C, in that all attributes used in C and M may be the subject for data quality improvement, thus increasing manufacturer and consumer fitness for use as well as preventing wasteful increases in quality of less important attributes.

This methodology holds M fixed for the given P and will attempt to determine revisions necessary in C to more efficiently target attributes requiring data quality improvement. M may then be revised in an attempt to better align with C. This evaluation of C will use passive surveying. Passive surveys will use P associated with a disruption function D. Passive surveys are defined as a survey in which the consumer is not told values are intentionally disrupted, only to review P and make any corrections necessary. This method of surveying is not appropriate for research given the ulterior motive of the

survey. However, revealing the ulterior motive defeats the very nature of the proposed research by revealing the presence of altered data. The consumer, knowing of these defects, may be more included to search for quality problems than casually reviewing P.

### Disruption Function

A trivial assignment to algorithm D may be a shuffling, or reassignment of an entity's data to another entity. Other implementations of D include populating an entire attribute with blank values, inverting signs of numeric values, or randomly injecting characters in text values. The disruption function alters the P in such a way that C would be expected to respond with a lowered fitness for use value, as an attribute of high importance has been degraded. Failure of the C to change given disrupted data would reveal the low importance rating of those attributes disrupted, as the consumer either may not notice the degraded attribute or simply find the inclusion of the attribute irrelevant.

A simple disruption showing four possible outcomes of set A with two attributes, given R quality ratings binned {L, H}, representing a simple binary result of low and high quality. Only these two discrete quality ratings are used for simplicity of demonstration. Disruption can only lower the quality rating performed in Q, as it is assumed through logic that random entropy placed upon data elements cannot increase their quality. A transition from bin H to L will take place; therefore all initial ai values are ranked at high quality H, denoted as $a_i$H.

| Disruption | Original P | Disrupted P |
|---|---|---|
| Disrupt $a_1$ | P($a_1$H, $a_2$H) | P($a_1$L, $a_2$H) |
| Disrupt $a_2$ | P($a_1$H, $a_2$H) | P($a_1$H, $a_2$L) |
| Disrupt entire set A | P($a_1$H, $a_2$H) | P($a_1$L, $a_2$L) |
| No Disruption | P($a_1$H, $a_2$H) | P($a_1$H, $a_2$H), represents control mapping |

**Table 4: Disruption function examples**

D will always result in a number of possible disruption functions $d = k^i$, an important consideration when assigning additional bins to R as possible disruption combinations are created exponentially. A separate group must be surveyed on each resulting disrupted P. Biasing may be made based on M and some attributes in A may be excluded from testing as it may exceed the cost of surveying all disrupted P combinations, as $k^i$ number of survey groups would be needed to validate results. This becomes impractical paired with the exponential growth of disrupted Ps required as additional attributes are added.

A subset of D may be chosen if certain combinations of degraded attributes are preferred for testing. Theoretically, an exhaustive combination, using all members of A, would show the true nature of the purpose of this method: to make no assumptions as to the importance of attributes held in M or C. The control function, mirroring the original P, should always be used, and is worth noting also represents a typical consumer feedback survey since no distortion takes place.

## RESULT: IMPLICATION CASES

The final results are formed through analysis of the initial state of P, the disrupted states of P, the specific attributes corrected by users, and the implications that reveal which attributes should be contained in C. In the following table, only one attribute is taken into account, since the total number of implications follows the form $d^2$, again illustrating the exhaustive and costly nature of this method. An example with A of set size 2, with 2 bins, would result in d = 4 and 16 possible implications. Below is A of set size 1, 2 bins, and only four possible implications, requiring only two survey groups.

| Implications | $a_1$ Corrected by consumer | $a_1$ Not Corrected by consumer |
|---|---|---|
| $a_1$ Not Disrupted | 1: $a_1$ is important to user and not correct in data source, should increase quality control efforts for a1 | 3: $a_1$ may or may not be important to user, but is reported as correct |

| | | |
|---|---|---|
| $a_1$ Disrupted | 2: $a_1$ is important to user and has been detected as incorrect by correction | 4: $a_1$ is not important to consumer, as it is incorrect but reported as correct |

**Table 5: Truth table for disruption and correction for one attribute**

The most important result for the current topic of importance is Case 4. Even though an attribute is intentionally disrupted in a way unusable for the consumer, it is nevertheless included in the final IP. Through no audit by consumers, this attribute indicates a point of waste for inclusion in the IP. However, the truth table does show the results as would be distributed as a percentage. Each result in the table would be accompanied by a ratio, the denominator of which represented by the survey size. If a threshold is defined weighing cost of inclusion of an attribute against the number of consumers using the attribute, then a decision can be made regarding the future inclusion of that attribute. For example, the cost of including the Serial Number in a receipt is high, but 90% of consumers would correct an incorrect serial number for warranty and registration purposes. A product number's inclusion on the receipt may be of high cost, but found that only 5% of respondents correct this attribute. The product number may be considered to be of low importance.

Case 1 represents a traditional survey, and indicates only those attributes manufacturers must improve regardless of any disruption or research results using the proposed framework for importance detection. Case 3 shows no indication of importance since there is no conclusive feedback from users. This may represent an overlooked data item, or a brief "once-over" of a report. Case 2 represents a roadblock in the area of academic research, or application of this framework in a real world case. A survey of consumers that hides the true intention of the research is generally considered an inappropriate research method. Once the consumer has insight of data in the IP that may be intentionally disrupted, the Case 4 results are expected to decline, as consumers would be more aware to the presence of incorrect data.

## USE CASES

The following use cases illustrate two separate points taking into account the proposed method. The first addresses a physical product, as opposed to an intangible information product, and addresses the difficulty and restrictions in passive surveying of consumers after disrupting data sources. The second uses a simple IP and reveals one additional implication case not addressed in the previous implication cases.

### Use Case 1: Physical Product

A simple cake analogy illustrates the application of this framework to determining attributes that are not necessarily important to consumers. A cake is a physical product, created from sources (ingredients) and constructed by manufacturers. The sources can be high or low quality; higher quality ingredients require higher costs. Inclusion of ingredients may create a better case, but require more time to integrate with other attributes. Whipped cream may be an included attribute on the cake, and increase its overall fitness for use (taste) by consumers. Cost comes not only with the purchasing of the cream, but also in processing (whipping) and integration with the cake.

Given high quality, fresh ingredients one would expect a high quality cake. However, even with the finest ingredients, consumers may have afflictions such as lactose intolerance, allergies, or general dislike for a flavoring used. This represents a feedback mechanism at the time of delivery resulting in uneaten cake, and showing a lack of fitness for use. Some consumers may scrape the attributes they don't enjoy from the cake, representing partial usefulness of an IP product. In this example, the amount of cake consumed reflects the enjoyment of the cake, and can be directly measured in physical form as fitness for use. The proposed framework would first determine a baseline of quality ingredients to create a cake; several cakes and survey groups would then be constructed to determine if revisions in data sources (ingredients) are needed or can be afforded.

Several cakes may be made, each with absence of an ingredient or with a lower quality substitution, representing the disruption algorithms in D. It's assumed the distorted cakes are made with lower cost if poor quality ingredients or exclusion of ingredients entirely resulting in saved time and resources. After delivery of the cakes, consumers may express enjoyment or disgust, as would correlate to the implication case results in Table 5. Corrections would be thought of as complaints, or poor view of the cake.

If a lower quality cake can be manufactured, with no consumers complaining of bad taste, then the entire product as a whole is still fit for use and can then on be manufactured at a lower cost. The response to altering certain cakes may indicate one ingredient that should never be omitted, or one that can be replaced with that of a lower quality if only a minority of consumers respond. As more cakes are baked, allergies may fade and preferences in taste may change, representing the continuing fluctuation of M(V) striving to match C(V).

### Use Case 2: Information Product

The following use case will reveal a fifth implication case, aside from those presented in Table 5. To create this additional case, more than one attribute must be used in the set A. The following set A represents a patient database for blood transfusion.

A = {Patient Name, Blood Letter, Blood Sign, Date of Birth}

Assume quality values have been given to each attribute and a baseline for creation of D has been established, with k bins of High and Low, resulting in a D of size 16. Only 2 of the 65536 possible implications will be analyzed in this example to illustrate a side effect of multiple attribute corrections. The following two disruptions each use the four mentioned attributes.

| Disruption | Original IP | Disrupted IP |
|---|---|---|
| Disrupt Date of Birth | $P(a_1H,a_2H,a_3H,a_4H)$ | $P(a_1H,a_2L,a_3H,a_4H)$ |
| Disrupt Blood Letter, Birth Date | $P(a_1H,a_2H,a_3H,a_4H)$ | $P(a_1H,a_2L,a_3H,a_4L)$ |

**Table 6: Special Implication Case Disruptions**

Let's assume the survey has been carried out on only these two disruption sets. As the number of attributes increase, the truth table increases in size to match each disruption as shown in Table 7.

| Implications | | $a_1$ Corrected | | $a_1$ Not Corrected | |
|---|---|---|---|---|---|
| | | $a_2$ Corrected | $a_2$ Not Corrected | $a_2$ Corrected | $a_2$ Not Corrected |
| $a_1$ Not Disrupted | $a_2$ Not Disrupted | Case 1 | **Case 5** | Case 9 | Case 13 |
| | $a_2$ Disrupted | **Case 2** | Case 6 | Case 10 | Case 14 |
| $a_1$ Disrupted | $a_2$ Not Disrupted | **Case 3** | Case 7 | Case 11 | Case 15 |
| | $a_2$ Disrupted | Case 4 | Case 8 | Case 12 | Case 16 |

**Table 7: Implications resulting from two attributes under analysis**

The special cases implied from the 16 outcomes are shown when any one attribute is disrupted and more than one attribute are corrected. Cases 2 and 3 illustrate this scenario. For Case 2, $a_2$ (Birthday), is disrupted and both $a_1$ and $a_2$ (Blood Letter and Birthday) are corrected. If this Case support percentage is higher in the population survey than Case 5, in which only Birthday is corrected without being disrupted, we might propose the following postulate for future research: The presence of incorrect data, alongside data attributes that may or may not be correct, has an immediate effect on the perception of the IP map as a whole to be of low quality, passively eliciting a “double-check” from the user to take more time and care in revising and correcting those attributes presented. This, along with the previous methods of determining attribute importance, may be combined to disrupt one set of important attributes in order to receive feedback when presented alongside of unknown importance. However, this is only a suggestion at future applications of the methods discussed in this paper, and would be difficult to implement based on the very nature of surveys and their roles in research.

## CONCLUSION

The purpose of many methods to improve quality attempt to inject high value into a final output, value being those attributes most important to consumers, or those requiring most fitness for consumers' uses; however that is not the intent of the topics

covered in this methodology. Some view assumed by manufacturers of their consumer's use of an IP must be established, and most often is, prior to experimental research in fine-tuning this view. A disruption function can convert data to poor quality temporarily only to trigger a response and refresh focus on the IP as a whole, not simply as a set of attributes. Measuring the response to this disruption can reveal a chance for manufacturers to preserve resources through detection of unused or misused attributes that, in the final product, don't truly matter to consumers. In addition, this same methodology can be expanded in scope and generalized not only to surveys, but to any complex black box system that consumes IPs, since human consumers of data may be considered as such in the systematic scope of IP manufacturing systems. A consumer will continue to silently consume, but only in the presence of a disruption, either spoiled milk for a cake or an information quality disaster, do manufacturers gain most insight into the true utilization of their information products.

## REFERENCES

1. Ballou, D. P., Pazer, H. L. (1982) Modeling data and process quality in multi-input, multi-output information systems, *Management Science*, 31, 2, 160-162.
2. Ballou, D. P., Tayi, G. K. (1989) Methodology for allocating resources for data quality enhancement, *Communications of the ACM*, 32, 3, 320-329.
3. Ballou, D. P., Wang, R. Y., Pazer, H. L., Tayi, G. K. (1998) Modeling information manufacturing systems to determine information product quality, *Management Science*, 44, 4, 462-484
4. Fisher, C. W., Kingma, B. R. (2001) Criticality of data quality as exemplified in two disasters, *Information & Management*, 39, 2, 109-116.
5. Juran, J. M., Gryna, F. M., Bingham, R. S. (1974) Quality control handbook, 3rd edition, McGraw-Hill, New York.
6. Kahn B., Strong D., Wang, R. Y. (2002) Information quality benchmarks: product and service performance, *Communications of the ACM*, 45, 4, 184-192.
7. Lee, Y., Strong, D., Kahn, B., and Wang, R. Y. (2002) AIMQ: A methodology for information quality assessment, *Information & Management*, 40, 2, 133-146.
8. Pipino, L., Lee, Y. W., Wang, R. Y. (2002) Data quality assessment, *Communications of the ACM*, 45, 4, 211-218.
9. Talburt, J. R., Zhou, Y., Shivaiah, S. (2009) SOG: A synthetic occupancy generator to support entity resolution instruction and research, *International Proceedings of the Conference on Information Quality*, November 7- 8, Potsdam, Germany.
10. Wang, R. Y., Strong, D. (1996) Beyond accuracy: what data quality means to data consumers, *Journal of Management Information Systems*, 12, 4, 5-34.