Global Journal of Computer Science and Technology
Volume 11 Issue 9 Version 1.0 May 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
ISSN: 0975-4172 & Print ISSN: 0975-4350

A Rare Example of Pitfall in Corporate Data Modeling Practices for Information Systems

By Hae Kyung Rhee

Abstract- Although Entity Relationship Model have turned out to be de facto corporate data modeling vehicle, current exercise for its application to real-world business processes is disclosed to surprisingly face a pressing peril of an alarming level in terms of data consistency and the degree of unnecessary data redundancy. In this paper, the matter-of-fact legacy practice, once thought to be a confidential or secret, obtained from some major broadcasting company is articulately reported and its demerits as well as its antagonistic side effect due to abnormality are discussed in depth. We were able to remold such an ill-manifested case to a legitimate ER model vindicated through an endeavor of a couple of months long. The two cases then are compared both qualitatively and quantitatively. The result of analysis has shown that ill-formed data models contain more than a ratio of 40 percent of unnecessary data redundancy, which leads us to have an implication that it is heavily contingent to seek a corrective measure for cutting down the ratio.

Keywords: *Data modeling Entity Relationship model; Data consistency; Data obesity.*

GJCST Classification: *H.2.1*

Strictly as per the compliance and regulations of:

© 2011 Hae Kyung Rhee. This is a research/review paper, distributed under the terms of the Creative Commons Attribution-Noncommercial 3.0 Unported License http://creativecommons.org/licenses/by-nc/3.0/), permitting all non-commercial use, distribution, and reproduction inany medium, provided the original work is properly cited.

A Rare Example of Pitfall in Corporate Data Modeling Practices for Information Systems

Hae Kyung Rhee

Abstract- Although Entity Relationship Model have turned out to be de facto corporate data modeling vehicle, current exercise for its application to real-world business processes is disclosed to surprisingly face a pressing peril of an alarming level in terms of data consistency and the degree of unnecessary data redundancy. In this paper, the matter-of-fact legacy practice, once thought to be a confidential or secret, obtained from some major broadcasting company is articulately reported and its demerits as well as its antagonistic side effect due to abnormality are discussed in depth. We were able to remold such an ill-manifested case to a legitimate ER model vindicated through an endeavor of a couple of months long. The two cases then are compared both qualitatively and quantitatively. The result of analysis has shown that ill-formed data models contain more than a ratio of 40 percent of unnecessary data redundancy, which leads us to have an implication that it is heavily contingent to seek a corrective measure for cutting down the ratio.

Keywords: Data modeling; Entity Relationship model; Data consistency; Data obesity

I. Introduction

a) *Background- What We Have Dug Out*

Although this era of history is called information age, it seems to many of us that it meant to be the age of computerized or automated program rather than the age of data, since response time to data we need to retrieve is getting slower than ever. The slowness could be due to data inconsistency [FiKi2010, CiFrMa2009] or replication or redundancy for the pursuit of mere convenience or due to data deluge phenomenon [KaBoZe2010] or data overload phenomenon[ShRi2011]. Although the issue of data deluge has been recently brought into a focal point, the issue of data inconsistency is considered to be more imminent and fundamental in that it could aggravate the problem of deluge if not legitimately rectified and also in that the impact of deluge phenomenon to the issue of inconsistency is more or less inefficacious.

The issue of propriety of data modeling has not been diminished since the days of as early as early 1990s just under the presumption that it is more or less vague to debate whether a data item should be required to be classified either as an entity or as a relationship or even as an attribute. However, as the occasions of prolongation in response time to queries submitted by users have been reported from a number of sources these days, we judged this might be the right time to shed light again on the issue of corporate data modeling, since such a modeling is mostly carried out by persons of non-expertise and there seems to be a lot of serious flaws as they have often been confessed by a slew of field practitioners.

We still do not know the exact origin of problem inducing slowness of response, but in this report we are able to reveal some of the reasons that impede responsiveness with the entity-relationship models we obtained in the real-world 2

corporate data modeling practices. The so-called data maps, ER models [DaGrRo2006], are not easy to obtain, since most corporate is reluctant or hesitant to disclose their internals of blue-print of data design, particularly at the level of data attribute, as it is more or less considered to be a sort of secret due to confidentiality associated with database or a sort of shame due to field practitioners' lack of expertise in proper and legitimate data modeling.

One prototypical example of data modeling practices is from one of the major world-class broadcasting companies. We could not reveal the name of the corporate, but the practices are not only limited to that company.

They are pretty much prevalent these days when readers of this paper have a chance to have a glimpse at the ER models like Fig. 1. The diagram like Fig. 1 is canonically called as the ER model, even though it fails to follow the simple and basic rule of connecting a relationship to an entity. In ER world, a relationship is practically either a binary, i.e. two-way relationship, or ternary. Anything more than ternary is in reality hard to come by and at the same time hard to think of.

About- Associate professor at Dept. of Computer Game in Yong- In Songdam college.(telephone:82-31-330-9234 email:leehk@ysc.ac.kr)

©2011 Global Journals Inc. (US)

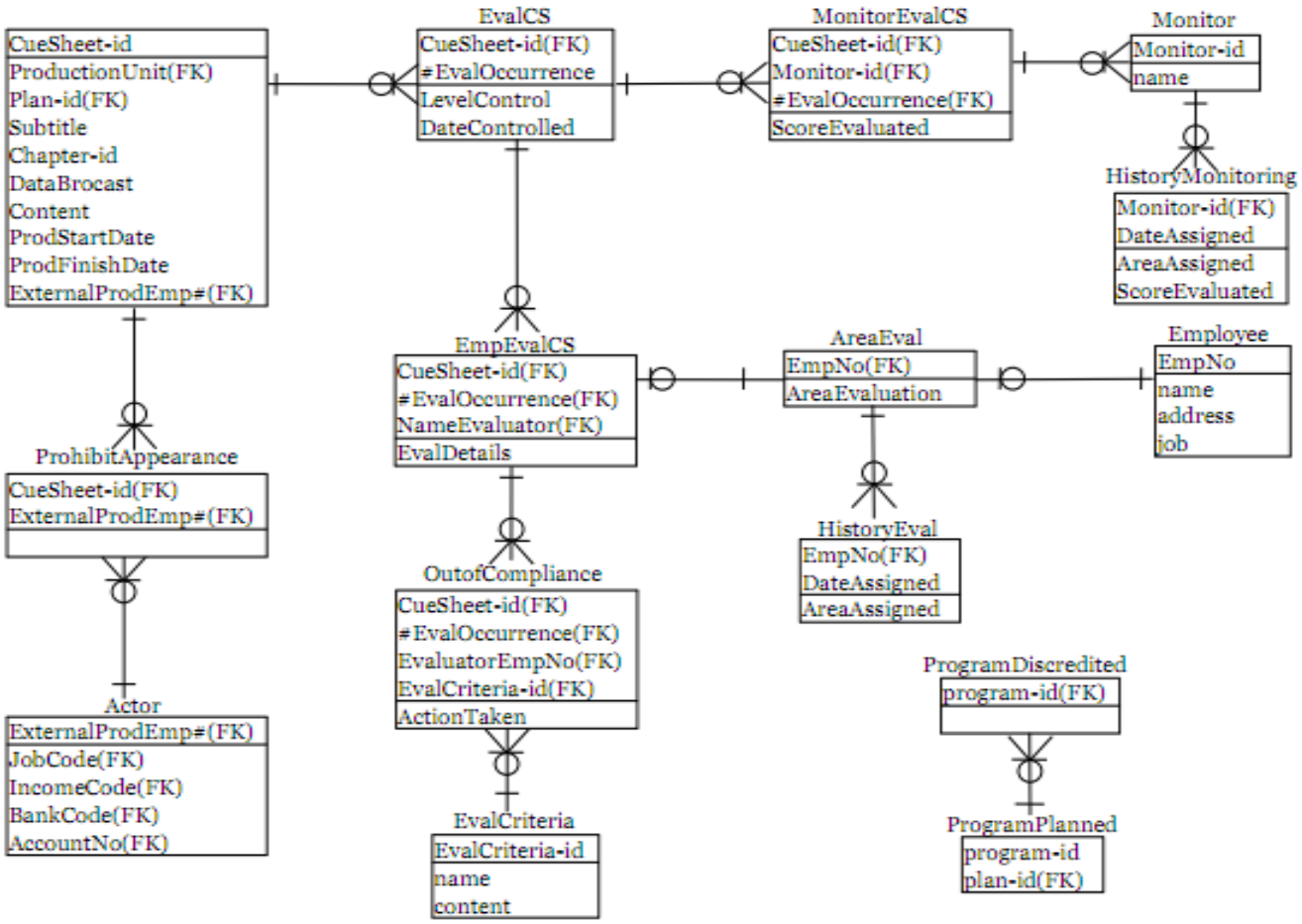

Fig. 1. Real-World Data Modeling Practice in Major Broadcasting Company

For instance, a binary relationship is used to describe an action, denoted by a verb, that identifies the subject of the action and the target object of the action in a way as in Fig. 2.

Fig. 2. Action Meant to Be a Relationship between Two Entities

b) *Motivation*

i. Is Omission and Over-Simplification or Abstraction of Data Tolerable in Real-World Modeling?

It is obvious to find the drawing nature of Fig. 1 does not obey or intentionally dishonor the rule of binary relationship in that the diagram can be translated into Fig. 3 if we take it seriously in the perspective of relationship. It is extremely abnormal or insane in that a relationship frequents a direct connection to another relationship without going through a certain entity. In reality, there are many regular entities omitted, either unconsciously or consciously, in the real-world data model of Fig. 1.

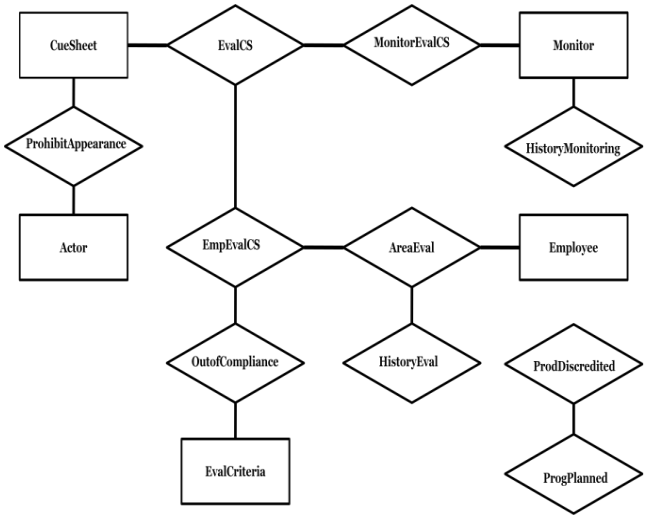

Fig. 3. ER Modeling Practice Out of Normalcy

This sort of abstraction or over-simplification might be originated from lack of knowledge associated with the philosophy and paradigm of conceptual data modeling, but unfortunately this customary misact is never tolerable, since it inevitably bring a serious blunder to the quality of data models. It must be carefully observed that it is obviously too far-fetched to the wrong direction from the basic guideline for entity-

©2011 Global Journals Inc. (US)

relationship modeling, which is the clear doctrine even the UML[ArElLa2006], the mostly favored data modeling tool these days, take its fundamental basis.

ii. *Is Real-World Model Merely Entity-Oriented Model?*

In this sense, the model is not relationship-oriented, e.g. business action-oriented or behavior-oriented [YaElOu2010], but rather it is fairly entity-oriented by omitting either sometimes a subject entity or other times a target entity. Not surprisingly, internal data models in most of the renowned ERPs are nonetheless no different from an ER model as in Fig. 1. Consequence of frequent adoptions of such ill-manifested data models by frequenting omission of entities shall cause a tremendous degradation of information systems, in terms of quality of response time and travail of maintaining consistent view to attribute-level data items, under which those type of data models prevail.

iii. *Objectives- How Much Far-Fetched is the Field Legacy Practice in Terms of Data Quality?*

Once the real-world practice is obtained, we were able to translate it to an ER model that could be considered to be normal and sane, albeit this

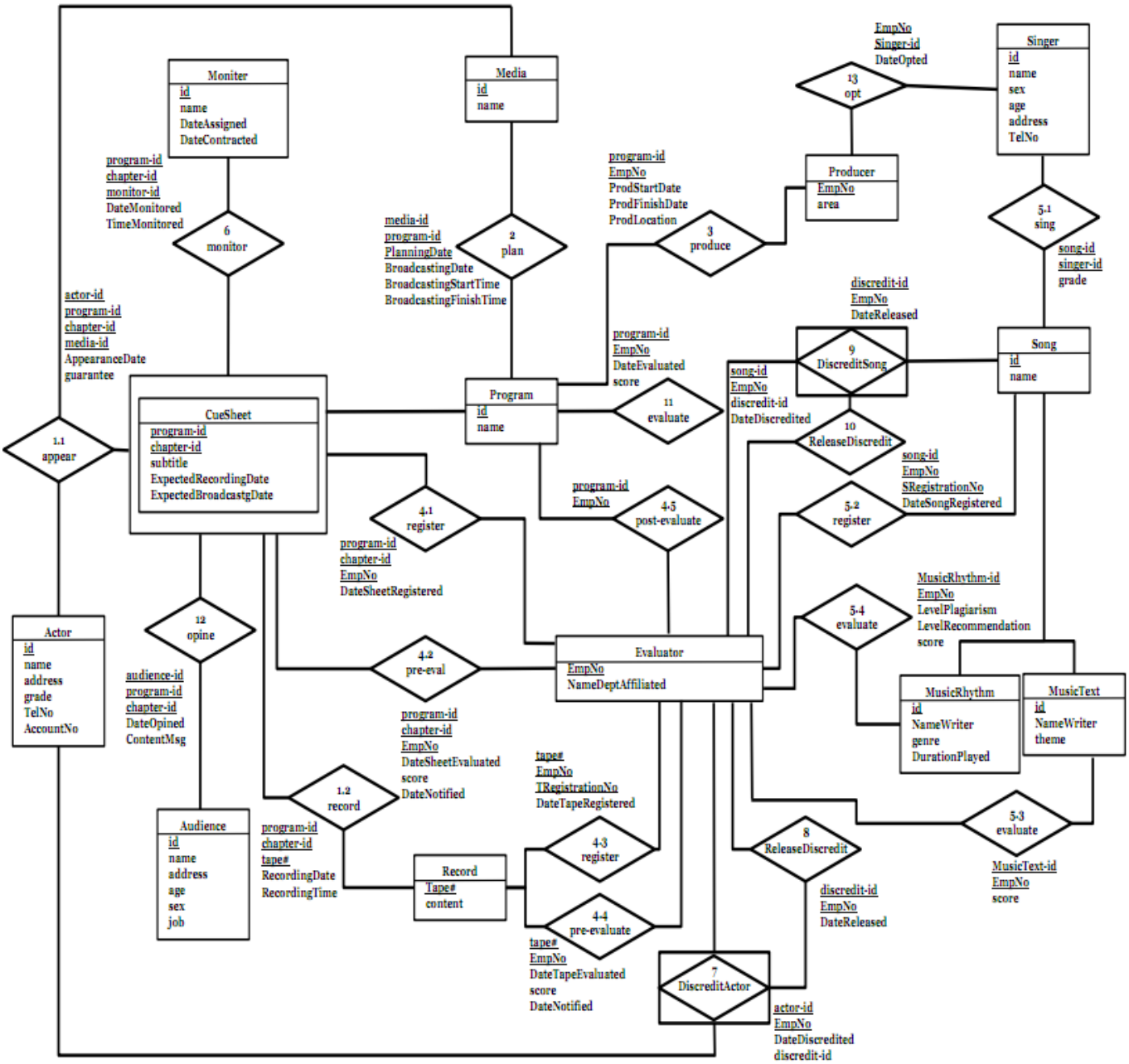

Fig. 4. The ER Model Translated with Scrutiny

©2011 Global Journals Inc. (US)

translation job took a labor of a couple of months long to discover how attribute-level data is associated with business processes. The final ER model obtained is depicted in Fig. 4. It seems to be classic in the sense that the philosophy of connecting an entity to another entity is fully honored in a way that any two entities are permitted to be linked always through a relationship that make two of them directly associated with each other. Notethat, in Fig. 4, key attributes are underlined for entities and relationships, and weak entities are denoted by double rectangles.

For instance, in Fig. 4, entity *CueSheet* is a weak entity deduced from a regular entity *Program*.

Relationship turned entity, i.e. entity upgraded from a relationship for the sake of the action which is directly subsequent to the relationship is denoted by diamond enclosed by rectangle. For instance, *DiscreditActor* is the action prerequisite to the action *ReleaseDiscredit*.

The awkwardness of data models of type Fig. 1 leads us to contemplate on the cause and effect of technical blunders brought by data modelers of non-reputing the veracity of genuine ER modeling principles. The significance of them is then compared against the case of data models of type Fig. 4 to see how much deviation or damage has been brought by improper data modeling practices prevalent in legacy IT field. By conducting careful examinations to them, we are able to deduce a couple of insights that are considered to be very valuable to the readers who might be interested in corporate-wide data management through decent data modeling. The scope of this paper is to present such insights one by one to the degree as much as we can deliver them both qualitatively as well as quantitatively.

II. Notion of Corporate Data Map

Note that the case of Fig. 1 or Fig. 4 is one particular business process in the broadcasting company we contacted. The company is comprised of approximately 800 different such business processes. As it can be observed in Fig. 4, corporate data map is simply more or less like a road map in that where there is a way out should there be a way in as well in a way of forming circular paths as in Fig. 5.

As a road map of, say, a country is fully connected in a sense that there is always a connection to any isolated area, for instance, islands, a data map should has a property of full connection or perfect reachability in a way of never allowing any data silos[Moon2009]. Nevertheless, observe that there are data silos in Fig. 1. There are two. Allowing existence of silos incurs problems not only of data duplication but of slowness in response time due to difficulty in locating query response initiation spot. Once we designated a spot in a wrong silo, there would a waste of time proportional to the amount of unnecessary joins to get reached each individual one of terminal nodes in that silo. In contrast, Fig. 4 never allows data silos.

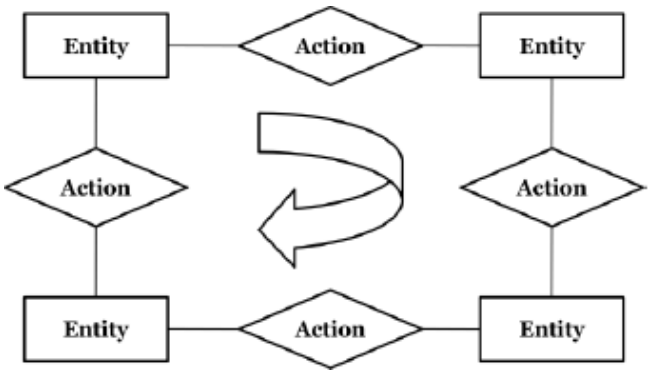

Fig. 5. A Circular Path in Corporate Data Map.

This property must be maintained and guaranteed even if we connect a slew of different data map segments altogether to get a single corporate-wide data map. Note that the major broadcasting company in this study contains 800 such separate data segments that can be connected altogether to get one single integrated corporate data map. Fig. 4 is just one of those segments.

It is evident that merely repeating data modeling practices as in Fig. 3 only contributes to forge of chaos in which ignominies of ER directives inadvertently prevail. Also note that there are 45 unique data attributes in Fig. 1, whereas 86 unique data attributes are included in Fig. 4. In this sense, acquisition of data map of full connectedness allows us to view data flow as well as process flow twice as much accurate and detail than the version of abstraction non-meticulously adopted in Fig. 1. The burden we need to pay with corporate data modeling fashion of Fig. 1 is huge. Let us compare this type of absent-minded approach, say of Fig. 1, as opposed to the type of thoughtful or considerate approach that is incumbent to take a stance of genuine and authentic approach, say of Fig. 4.

III. Skewedness Versus Balancedness

Absent-minded data modeling, AMD for short hereafter, is the canonical example of skewed design. Note that all the 6

links, i.e. relationships between entities, are just stretching out to the end, never returning or regressing to the point of origin of a link already disseminated. For instance, have a look at links from CueSheet to EvalCS to EvalCSMonitor to Monitor, then to HistoryMonitoring. There is only links that are forging out whenever a link is established. That sort of instance also appeared all over the data map of Fig. 1 in that the same observation

©2011 Global Journals Inc. (US)

holds from CueSheet to EvalCS to EvaluatorCS to OutofObeyance to EvalCriteria and from CueSheet to EvalCS to EvaluatorCS to Evaluator to Employee. The data map of Fig. 1 is just full of such cases.

Thoughtful or considerate data modeling, TCD for short hereafter, is the canonical example of balanced design. It guarantees the notion of balanced tree [RaGe2007] in that once a node is designated to be the root of a tree, then there is a threshold of depth of its left subtree and of depth of its right subtree as well. In case there is a circular loop in a component of data map, those two depths happened to be always the same. Look at, in Fig. 4, a case of such circular path from CueSheet to Program to post-evaluate to Evaluator to register and back to CueSheet. Once a node is at random chosen to be the root of a tree, then the maximum depth of both left subtree of it and right subtree of it is always limited to 2. Such cases are prevalent in Fig. 4. Note that there are 14 different circular paths in Fig. 4, whereas there is none in Fig. 1. This is clear-cut contrast between balanced data modeling versus skewed data modeling.

The merit of well-balancedness is the guarantee of agile responsiveness, whereas the preference of skewedness only promotes a retard in response. Note that in TCD the number of joins required to get the desired answer to any query made is just 2, whereas in AMD it is 4. This suggests that response time in TCD is twice as much agile than AMD. Consequently, the cost we need to tolerate is untimely response accrued from the skewedness, whereas much more timely response would be envisioned if the balancedness of design is sought.

IV. Issue of Data Inconsistency

If we turn our attention to the issue of data inconsistency, it is self-evident that AMD obviously fails to provide a consistent view to even the same data. Look at, in Fig. 1, that notion of key attribute is definitely violated in that foreign keys, denoted in FK, appear their existence even in the non-key sections of entity or relationship. It is a strict rule that any key components should appear only in the key sections of an entity or a relationship, otherwise it is the case of overuse or tenet misuse.

Look at an entity *Actor* in Fig. 1. There are five different FKs as its non-key attributes. This is just absurd and it is an obvious sign of defamation to the notion of what the key is. The definition and principle of key must be preserved in the database at all times. It cannot be compromised or changed for mere convenience. The pursuit of this sort of ignominious act is definitely a sort of 'crime'.

It might be unconsciously committed due to misinterpretation of knowledge, but such commitment happens to bring very serious maintenance burden to corporate in that consistency is never guaranteed to be automatically preserved in such a database by the information system. Note that an information system is never the engine of superficial superpower that automatically insures data consistency. Accountability of such maintenance is entirely up to data modelers in charge. They are one hundred percent obliged to enforce data consistency. A proper and minimal use of foreign keys only waives them from manual enforcement of consistency.

Most of complaints with regard to reliance to ERP tools are instigated by the problem of data inconsistency even if intelligent ERP tools are considerately adopted. Information systems are just like any living fauna or flora in that they themselves in reality gradually evolve perhaps quarter by quarter. There is no guarantee of data consistency even though ERP approach is in use for whole information systems. Suppose a particular ERP tool has already been deployed for some part of IS for a certain business area of a company.

There would be no guarantee of placing the same ERP tool for some other part of IS, since technology evolves very rapidly in IT field. There is a quite a big probability of choosing some ERP tool other than the previous one in cases that the one that has been selected has more decent capabilities than any other candidates. Consequently, securing data consistency could be at serious peril as the degree of data replication would soon ascent to be prevailed in case that a certain part of data repository of IS happens to be replicated partly with some other parts of data repository.

V. Degree of Data Redundancy

In case data attributes are replicated, regardless of whether willingly or willy-nilly, maintaining data consistency remains careful craftsmanship of data modeler in order to keep track of where the replicated copies are and ensure their values to be identical only manually. Field practitioners usually tend to not willing to understand such a limit of capability of database engines. Engines do their role of automatically maintaining data consistency if foreign keys are used only sparingly in a way that only a relationship is permitted to borrow or import them via entities just directly connected to it.

For instance, in a relationship sing in Fig. 4 a foreign keys pair {song-id, singer-id}, which is comprised of foreign keys only, is forged. Note that song-id is borrowed from entity Song and singer-id is borrowed as such borrowed from Singer. This way of use of foreign keys sparingly is only appraised to be the genuine data modeling in the arena of relational modeling and object-relational modeling.

Any way of forging foreign keys other than this way of generation of foreign keys is considered to be out of normalcy in any circumstances. This form of

©2011 Global Journals Inc. (US)

import intrinsically induces data replication inevitably and this is dubbed inevitable redundancy [EllpVe2007] It should be bear in mind in corporate data modeling that any unnecessary redundancy can never be disguised under the name of inevitable redundancy whether it is intentional or unconscious.

Note that there are 42 occasions of data attribute replication in the data model of Fig. 1 of which 25 of them are obviously unnecessary. Since there are 61 attributes altogether in the data model of Fig. 1, about 40 percent of them are of unnecessary redundancy. Note also that the ratio of total data redundancy in the data model of Fig. 1 in reality amounts to 70 percent, which is computed from 42 divided by 61. This is surprisingly high and the major reason for this is due to misuse or overuse of foreign keys.

In contrast, in the data model of Fig. 4, there are 38 occurrences of data replication, so the ratio of data redundancy is about 30 percent, but all this is of purely inevitable redundancy type. Note that there is no surfeit in deployment of foreign keys. In sum, the burden cost we need to pay for the misuse or overuse of foreign keys is much more than huge in that normally corporate database carry unnecessary burden of 40 percent of entirety of it.

VI. Conclusions

The major contribution of this paper is to disclose the ashamed privy parts in corporate data modeling that have been concealed and sedated so long with respect to serious pitfalls of their inmost details. It has been in a sense almost thoroughly out of conscience from even data modeling experts for the past couple of decades. Shedding a light on this issue is considered to be imminent in that so many discontents with information systems, without regard to whether they are on basis of ERP or not, have been reported in the real-world front [Cukier2010].

Unless the resolutions to the problem of data inconsistency and the problem of unnecessary replication of data are rendered, commencing to discourse about the issue of data deluge seems to be insensible. The degree of data redundancy is revealed to be enormously paramount. According to the outcome of a recent study, the level of data obesity [Rhee2010], measured in terms of the ratio of the total number of unnecessary data attributes to the number of all of the attributes resident in a corporate database, is reported to be about 50 percent in average at best.

This implies that more than 40 percent, which has been discovered in the case study in this article, in actuality goes with unnecessary parts of database. We believe that to help field practitioners design data models legitimately an automated tool [LeKiMo2010] enabling to corroborate data modeling labor by hand must be devised as soon as algorithmic solution to this problem would be discovered.

References Références Referencias

1. [FiKi2010] C. W. Fisher, B. R. Kingma(2010). Criticality of data quality as exemplified in two disasters. Information Systems, Vol. 39, 109-116.
2. [CiFrMa2009] C. B. Cinzia Cappiello, C. Francalanci, A. Maurino(2009). Methodologies for data quality assessment and improvement, ACM Computing Surveys Vol. 41, No. 3, Article 16, 52.
3. [KaBoZe2010] D. Katz, M. Bommaroti, J. Zelner(2010, March 1). The Data Deluge. The Economist print edition.
4. [ShRi2011] T. Shanker, M. Richtel(2011 Jan. 16). Data overload can be deadly. The New York Times print edition.
5. [DaGrRo2006] I. Davies, P. Green, M. Rosemann, M. Indulska, S. Galo(2006). How do practitioners use conceptual modeling in practice? Data and Knowledge Engineering Vol. 58, 358-380.
6. [ArElLa2006] E. Arisholm, S. Elisabeth Hove, Y. Labiche(2006). The impact of UML documentation on software maintenance: An experimental evaluation. IEEE Trans. Software Engineering Vol. 32, No. 6, 365-381.
7. [YaElOu2010] M. Yakout, A. K. Elmagarmid, H. Elmeleegy, M. Ouzzani, A. Qi(2010). Behavior based record linkage. Proc. VLDB Vol. 3, No. 1, 439-448.
8. [Moon2009] S. Moon(2009). Data Architecture, Hyungseol Publishing Co.
9. [RaGe2007] R. Ramakrishnan, J. Gehrke(2007). Database Management Systems , McGraw-Hill(Third Edition).
10. [EllpVe2007] A. K. Elmagarrmid, P. G. Ipeirotis, V. S. Verykios(2007). Duplicate record detection: A Survey. IEEE Trans. Knowledge and Data Engineering Vol. 19, No. 1, 1-16.
11. [Cukier2010] Cukier, K. (2010, Feb. 25). Data, Data Everywhere. A Special Report on Managing Information, The Economist. Retrieved May 1, 2010, from http://www.economist.com/specialreports/displaystory.cfm?story_id=15557443.html
12. [Rhee2010] H. Rhee(2010). Corporate data obesity: 50 percent redundant. Journal of Computer Science and Technology, Vol. 10, No. 5, 7-11.
13. [LeKiMo2010] S. Lee, N. Kim, S. Moon(2010). Context-adaptive approach for automated entity relationship modeling. Journal of Information and Science and Engineering, Vol. 26, 2229-2247.

©2011 Global Journals Inc. (US)