Some Methods of Data Improvement in EIS

Mladen Varga, Katarina Ćurko
Faculty of Economics & Business, Zagreb, Croatia
{mvarga,kcurko}@efzg.hr

Abstract. *The paper considers some of the challenging aspects of data improvement in an enterprise information systems (EIS), such as coping with data overload, data degradation, data integration, data ownership/stewardship, data quality, data privacy and data visualization.*

Keywords. Data management, data overload, data integration, data quality, data degradation, data ownership, data stewardship, data privacy, data visualization

1 Introduction

The underlying task of an enterprise information system (EIS) is to link processes on the operational, management and decision-making level so as to improve performance efficiency, support good quality management and increase decision-making reliability [1].

Data management involves activities linked with the handling of all organization's data as information resource. The Data Management Association (DAMA) Data Management Body of Knowledge defines data management [3] as "the development, execution and supervision of plans, policies, programs and practices that control, protect, deliver and enhance the value of data and information assets." Effective data management in an EIS is essential. The paper considers some, possibly not all, aspects of data improvement [14] in an EIS (Table 1.).

Table 1. Data improvement aspects

Aspect	Challenge
Data overload	Can we cope with all that data?
Data integration	How to integrate data? Could the data integrate the business? Do we know organization's data? Do we know how to use organization's data?
Data quality	Do we know what quality data is? Are we aware that the quality of data is degradable?
Data ownership/ stewardship	Who is the owner/steward of the data?
Data privacy	Is data privacy a concern?
Data visualization	Could we amplify cognition of data?

2 Data overload – exploding data

Challenge: Can we cope with all that data?

It estimates [5] that the digital universe in 2007 numbers 2.25 x 10^{21} bits, and by 2011 it will be 10 times the size it was in 2006. 70% of the digital universe is created by individuals, but enterprises are responsible for the security, privacy, reliability, and compliance of 87% of the digital universe. The amount of data content is increasing to the point that we individually and collectively suffer from data overload.

Decision making processes show that data overload negatively impacts decision performance. Decision makers face three problems: (a) *problem of timeliness*: high volumes of data constrain them if they are performing both sense making and decision making tasks, (b) *problem of throughput*: high volumes of data overwhelm and distract them during sense making tasks, causing "analysis paralysis" and lowering the overall performance, and (c) *problem of focus*: decision makers have a finite amount of attention that may be distributed across tasks.

Data fusion is an example of a technique to cope with data overload. By means of fusion, different sources of information are combined to improve the performances of a system [4]. An illustration of fusion is the use of various sensors, typically to detect a target. The different inputs may originate from a single sensor at different moments (fusion in time) or even from a single sensor at a given moment. In the latter case several experts process the input (fusion on experts).

As a response to the question "Can we ever escape from data overload?" it is suggested [16] that the problem may be resolved by cognitive activities involved in extracting meaning from data. They argue "that our situation seems paradoxical: more and more data is available, but our ability to interpret what is available has not increased". We are aware that having greater access to data is a great benefit, but the flood of data challenges our ability to find what is informative and meaningful.

Reduction or filtration of data does not seem

to solve the problem of data overload. Data which is evaluated as unimportant and non-informative and thus eliminated may turn out to be critically important and informative in another particular situation. Data are informative by "relationship to other data, relationships to larger frames of reference, and relationships to the interests and expectation of the observer. "Making data meaningful always requires cognitive work to put the datum of interest into the context of related data and issues" [16].

Consequently, all approaches to overcome data overload must involve some kind of positive selectivity. "Positive selectivity facilitates or enhances processing of a portion of the whole. In this form of selectivity, we use positive metaphors such as a spotlight of attention or a peaked distribution of resources across the field [16]". Current research on cognition suggests developing positive selectivity techniques by thorough exploration of the available data. Negative selectivity or filtering, which is commonly used, inhibits processing of non-selected areas or even delete unnecessary data. "The critical criterion for processes of selection, parallel to human competence, is that observer need to remain sensitive to non-selected parts in order to shift focus fluently as circumstances change or to recover from missteps" [16]. Furthermore, data processing should create context sensitivity rather than insist of data finesse.

3 Data integration

Data integration is a process of integrating mutually related data which resides at autonomous and heterogeneous sources in an EIS, and providing a unified view of integrated data.

Challenge: How to integrate data?

Although the data integration problem has been theoretically extensively considered many problems still remain. Let us look at the basic types of data integration.

In *federation approach* each of data sources talks independently through wrapper to other data sources in order to be mutually "integrated".

Data warehouse approach is successfully used in many commercial systems oriented on data analysis and decision support. Information is extracted from source databases A, B and C to be transformed and loaded into the data warehouse D by the process of extraction, transformation and load (ETL). Each of data sources A, B or C has its own and unique schema. After extracting data from A, B and C, data may be transformed according to business needs, loaded into D and queried with a single schema. The data from A, B and C are tightly coupled in D, which is ideal for querying purposes. Nevertheless, the freshness of data in D constitutes a problem as the propagation of updated data from the original data source to D takes some time, called latency time. As business becomes more real-time, the system that supports it needs to be more real-time [8]. Some useful techniques are *"near real-time" ETL* (executing the ETL again), *direct trickle feed* (real-time data warehouse where the data warehouse is continuously fed with new data), *trickle and feed* (data is continuously fed into staging tables and at a given period the staging table is swapped with the actual table so the data warehouse becomes instantly up-to-date), and *external real-time cache.* Full integration of the data is achieved by the warehouse approach supported by an active data warehouse. The *active data warehouse* represents a single, canonical state of the business, i.e. a single version of the truth. It represents a closed loop process between the transactional (operational) system and data warehousing (analytical) system. The transactional system feeds the data warehouse, and the data warehouse feeds back the transactional system in order to drive and optimize transactional processing. Thus, the data warehouse is active if it automatically delivers information to the transactional system.

Recent trends towards *mediator approach* try to loosen the coupling between various data sources and thus to avoid the problem of replicated data. The main problem of replicated data is how efficiently keeping replicas synchronised, i.e. with the same content. This approach uses a virtual database with mediated schema and wrapper, i.e. adapter, which translates incoming queries and outgoing answers. A wrapper wraps [13] an information source and models the source using a source schema. The users of the integrated system, i.e. data source D, are separated from the details of the data sources A, B and C at the schema level, by specifying a mediated or global schema. The *mediated schema* is a reconciled view of data in sources A, B and C, so the user needs to understand the structure and semantics of the mediated schema in order to make a meaningful query [9]. The source databases are approachable

through a wrapper code which transforms the original query into a specialized query over the original databases A, B or C. This is a view based query because each of the data sources A, B or C can be considered to be a view over the virtual database D. The first step in the approach involves setting up the mediated schema either manually or automatically. The next step involves specifying its relationships to the local schemas in either Global As View (GAV) or Local As View (LAV) fashion. In the GAV or global-centric approach the mediated or global schema is expressed in terms of the data sources, i.e. to each data element of the mediated schema, a view over the data sources must be associated. In the LAV or source-centric approach the mediated or global schema is specified independently from the sources but the sources are defined as views over the mediated schema.

The *query language approach* propagates extending query languages with powerful constructs to facilitate integration without the creation of the mediated or global schema. SchemaSQL [7] and UDM [9] are examples of this approach. They serve as the uniform query interface (UQI) which allows users to write queries with only partial knowledge of the "implied" global schema.

An issue of growing interest in data integration is the problem of elimination of *information conflicts* among data sources that integrate. *Intensional inconsistencies*, referred to as semantic inconsistencies, appear when the sources are in different data models, or have different data schemas, or the data is represented in different natural languages or in different measures, such as "Is 35 degrees measured in Fahrenheit or in Centigrade?" In ontology based data integration it may be resolved by the usage of ontologies which explicitly define schema terms and thus help to resolve semantic conflicts. *Extensional inconsistencies*, often referred to as data inconsistencies, are factual discrepancies among the data sources in data values that belong to the same objects. Extensional inconsistencies are visible only after intensional inconsistencies are resolved. By [11] "all data are not equal", and "that data environment is not egalitarian, with each information source having the same qualification". In a diverse environment "the quality" of information provider's data is not equal. Many questions arise, such as: Is data enough recent or is outdated? Is the data source trustworthy? Is data expensive? The metadata, such as timestamp, cost, accuracy, availability, and clearance may help users judge the suitability of data from the individual data source [11].

Challenge: Could the data integrate the business?

The data could integrate the business across the organization by Enterprise Resource Planning (ERP) system where integrated data registers all business events instantly and only ones and thus reflect a single version of the truth of the business. A good EIS relies on ERP which uses a multitude of interrelated program modules that process data in individual functional areas, such as supply, manufacturing and sales. They stretch from the top to the bottom of the pyramid. If a program module covers the whole function, it runs through all management levels: operational, tactical and strategic. ERP is often supplemented by program modules for analytical data processing in order to support decision making process. According to business process approach, some typical process lines are introduced: Business Performance Management (a set of processes that help organizations optimize their business or financial performance), Customer Relationship Management (a set of processes utilized to provide a high level of customer care), and Supply Chain and Operations (a set of processes involved in moving a product or service from supplier to customer).

A vertically integrated EIS connects activities on the lowest level of a particular function (e.g. a process of transactional retailing) with data analysis and data representation on the decision-making level (e.g. reports on sales analysis for the general manager). If the vertical coverage of data from a particular functional area is insufficient, data does not comprise all levels of this functional area. This is the case when, for example, only operational sales data exist, whereas sales data for the tactical and strategic level of decision-making are missing due to the lack of appropriate sales reports or if there is no possibility to interactively analyse sales data.

A horizontally integrated EIS enables a systematic monitoring of specific business process from end to end. For instance, as soon as a purchase order arrives, an integrated EIS can receive it, forward it "automatically" to sales and delivery process which will deliver the goods to the customer and send the invoice to his EIS, open a credit account; record the quantity of goods delivered in warehouse evidence; in case

that goods first need to be produced in a manufacturing plant the system will issue a work order to manufacture the required quantity of goods; the production process can be supplied by a production plan; sales results will be visible to the sales manager and analysed using analytical tools of the EIS; etc. If data in an EIS is not sufficiently integrated, the EIS will be comprised of isolated information (data) islands. An EIS is not integrated unless individual functional areas are mutually integrated with other functional areas by data. For instance, sales data may not be integrated i.e. connected with production data; or they may be integrated in indirect and complicated ways.

Challenge: Do we know organization's data?

If we do not know organization's data well enough, we will use them rarely and poorly. Users often do not know what information is available and they do not use the EIS frequently enough. Users should be briefed about the EIS and the ways to use it. A (central) *data catalogue* should be provided to present the complete data repertoire.

Challenge: Do we know how to use organization's data?

If we do not know how to use the EIS well enough, we will not use it sufficiently and we will not exploit all of its capabilities. Even if users know that the organization has data, they may not know how to use it. The EIS should be documented and ways of using it described in a *catalogue of information processes*.

4 Data quality

Challenge: Do we know what the quality data is?

[6] defines data to be of high quality if the data is fit for their intended uses in operations, decision making and planning, i.e. if it correctly represents the real world to which it refers. Among various categories of attributes describing data quality the most commonly used are *accuracy* (degree of conformity of the data to the actual or true value to which it refers), *completeness* (if nothing needs to be added to it), *relevance* (the data is pertinent and adequate to the context and the person who use it) and *timeliness* (the data is given in timely manner). It is desirable that the data supported by EIS fulfil the mentioned quality characteristics. The end user requires quality data in quantities that support the decision-making process. Too much data can be a burden for the person who needs to make a decision.

Most companies view data quality control as a cost problem, but in reality it can drive revenue. The problem of data quality drives the companies to set up a data governance function whose role it is to be responsible for data quality. Executing a data quality project may help improve overall data quality in the EIS [10].

Challenge: Are we aware that the quality of data is degradable?

Many databases behave as growing organism. Data is added, updated or deleted. Due to new applications the structure of a database may be altered to meet new business requirements. During the database lifetime many alternations keep the database functional. But over time, with continuing alterations to its structure and content, the logical integrity of the database can be so degraded that it becomes unreliable. It often happens that data integration is negatively affected by the lack of proper documentation regarding changes, the fact that the people responsible for the application have moved on, or that the vendor of the application no longer exists. In many cases the organization does not realize that the database has been degraded until a major event, such as data integration, occurs.

The right way to determine the true nature of data is through a *data audit*, involving thorough data analysis or profiling. There are manual and automated methods of data profiling. In manual data profiling analysts assess data quality using the existing documentation and assumptions about the data to anticipate what data problems are likely to be encountered. Instead of evaluating all data they take data samples and analyze them. If such an analysis discovers wrong data, the appropriate programs have to be written to correct the problems. The manual procedure can be a very costly and time-consuming process requiring a number of iterations. The more efficient and accurate way of examining data is to use an automated data profiling tool that is able to examine all the data. Data profiling tools can identify many problems resident in the data and provide a good picture of data structure including metadata descriptions, data values, formats, frequencies, and data patterns. Automated data profiling solutions offer to correct data errors and anomalies more easily and efficiently.

5 Data ownership/stewardship

Challenge: Who is the owner/steward of the data?

Data ownership refers to both the possession of and responsibility for data. Ownership implies power over and control of data. Data control includes the ability to access, create, modify, package, sell or remove data, as well as the right to assign access privileges to others. This definition probably implies that the database administrator is the "owner" of all enterprise data. Nevertheless, this is not true. The right approach to ownership aims to find the user who will be responsible for the quality of the data.

According to [12], telling a user that he/she "owns" some enterprise data is a dangerous thing. The user may exercise the "ownership" to inhibit the sharing of data around the enterprise. Thus, the term "data stewardship" is better, implying a broader responsibility where the user must consider the consequences of changing "his/her" data. Data stewardship may be broken into several stewardship roles. For example, definition stewardship is responsible to clearly define the meaning of data and may be shared between analyst, database administrator and key user. Access stewardship is responsible to permit or prevent access to enterprise's data. Quality stewardship is responsible to fulfil the broad term "quality data" and may be shared between analyst and key user. Thus, data stewardship may be seen as distributed responsibility, shared between IT people (analyst, database administrator) and key users. Responsibilities assigned to all data stewardship actors, i.e. data stewards, must be registered in a data catalogue.

6 Data privacy

Challenge: Is data privacy a concern?

Data privacy is important wherever personally identifiable information is collected and stored. Data privacy issues arise in various business domains, such as healthcare records, financial records and transactions, etc.

The challenge in data privacy is to share data while protecting personally identifiable information. Since freedom of information is propagated at the same time, promoting both data privacy and data openness might seem to be in conflict. Data privacy issue is especially challenging in the internet environment. Search engines and data mining techniques can collect and combine personal data from various sources, thus revealing a great deal about an individual. The sharing of personally identifiable information about users is another concern.

Data protection acts regulate the collecting, storing, modifying and deleting of data which identifies living individuals. This data is referred to as *personal data*. The legal protection of the right to privacy in general, and of data privacy in particular, varies greatly in various countries. The most regulated are data privacy rights in the European Union (EU) where Data protection Directive 95/46/EC defines personal data as "any information relating to an identified or identifiable natural person". Personal data must be "(a) processed fairly and lawfully; (b) collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes; (c) adequate, relevant and not excessive in relation to the purposes for which they are collected and/or further processed; (d) accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that data which are inaccurate or incomplete, having regard to the purposes for which they were collected or for which they are further processed, are erased or rectified; and (e) kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed."

Privacy and data protection are particularly challenging in multiple-party collaboration. If heterogeneous information systems with differing privacy rules are interconnected and information is shared, privacy policy rules must be mutually communicated and enforced. Example of a platform for communicating privacy policy in the web-services environment is Web Services Privacy (WS-Privacy). P3P is a system for making Web site privacy policies machine-readable. It enables Web sites to translate their privacy practices into a standardized, machine-readable XML format that can be retrieved automatically and easily interpreted by a user's browser.

7 Data visualization

Challenge: Could we amplify cognition of data?

The results of data analysis of large data sets, such as textual, numerical or multimedia databases, are not attractive to the end users without good techniques of interpretation and

visualization of information. Information visualization [2] is "the use of computer-supported, interactive, visual representations of abstract data to amplify cognition. Cognition is the acquisition or use of knowledge. The purpose of the information visualization is insight, not picture. The goal of the insight is discovery, decision making and explanation. Information visualization is useful to the extent that it increases our ability to perform this and other cognitive activities."

Information visualization exploits the exceptional ability of human brain to effectively process visual representations. The purpose of visualization is to communicate information clearly and effectively through graphical means, so that information can be easily interpreted and used. Based on the scope of visualization, there are different approaches to data visualization. According to [15] the visualization types can be structured into seven groups: *sketches* (help visualize a concept), *diagrams* (precise, abstract and focused representation of relationships at times using predefined graphic formats), *images* (visualize impression, expression or realism), *maps* (represent individual elements in a global context), *objects* (exploit the third dimension), *interactive visualizations* (computer-based visualizations that allow users to access, control, combine, and manipulate different types of information or media), and *stories* (mental images are imaginary and non-physical visualizations). Each of mentioned visualization types has its specific area of application.

The visual elements should reveal the analytical or logical relationships in the data and gain the insight in the data. The data must be carefully visualized because wrong data visualization may misconcept the reader.

8 Conclusion

Aspects of data improvement are numerous and diverse. The most challenging aspects of data improvement deal with problems of data overload, data integration, data quality, data degradation, data ownership or stewardship, data privacy, and data visualization. Solving the problems offers innovative, better, more efficient and more effective data usage.

9 References

[1] Brumec, J. A contribution to IS General Taxonomy. *Zbornik radova*, Vol 21 No 1(22), 1997, pp 1-14.

[2] Card, S. K., Mackinlay, J. D., Shneiderman, B. *Readings in Information Visualization: Using Vision to Think*, Morgan Kaufmann, San Francisco,USA, 1999.

[3] *DAMA-DMBOK: Functional Framework*, ver. 3., DAMA International, 2008.

[4] Hall, D., Mcmullen, S. H. *Mathematical Techniques in Multisensor Data Fusion*. Artech House, 2004, Norwood, USA.

[5] IDS White Paper: The Diverse and Exploding digital Universe, IDC, Framingham, 2008.

[6] Juran, J M., Godfrey, A. B. *Juran's Quality Handbook*. McGraw-Hill,1999.

[7] Lakshmanan, L., V., Sadri, F., Subramanian, S. SchemaSQL: An extension to SQL for multidatabase interoperability. *ACM Transactions on Database Systems*, 26(4), 476–519, 2001.

[8] Langseth, J. Real-Time data Warehousing: Challenges and Solutions. http://DSSResources.com/papers/features/langseth/langseth02082004.html [29.08.2008]

[9] Lu, J. J. A Data Model for Data Integration, *Electronic Notes in Theoretical Computer Science* 150, 3–19, 2006. http://www. sciencedirect.com [29.08.2008]

[10] McGilvray, D. Executing Data Quality Projects. Morgan Kaufmann Publishers, 2008.

[11] Motro, A., Anokhin, P. Fusionplex: Resolution of Data Inconsistencies in the Integration of Heterogeneous Information Sources. *Information Fusion* 7, 176-196, 2006.

[12] Scofield, M. (1998). Issues of Data Ownership, *DM Review Magazine*. http://www.dmreview.com/issues/19981101/296-1.html [29.08.2009]

[13] Ullman, J. D. Information integration using local views. *Proc. of the 6th Int. Conf. On Database Theory* (ICDT'97), pp. 19-40, Delphi, 1997.

[14] Varga, M. Challenges of Data Management in Always-On EIS, in *Always-On EIS for Business Continuance*, ed. Bajgoric, N., Business Science Reference, Hersley, 2010.

[15] Visual Literacy: An E-Learning Tutorial on Visualization for Communication, Engineering and Business. http://www.visual-literacy.org/ [03.10.2008]

[16] Woods, D. D., Patterson, E. S., Roth, E. M. Can We Ever Escape from Data Overload? A Cognitive systems Diagnosis. *Cognition, technology & Work*, 4, 22-36, 2002.