AMCIS 2005 Proceedings

Americas Conference on Information Systems (AMCIS)

2005

Data Mining and the Five Pillars of Information Assurance: Where Does Society Draw the Line?

Mary Sparks
University of Detroit Mercy, marysparks@comcast.net

Antonio Drommi
University of Detroit Mercy, drommia@udmercy.edu

Dan Shoemaker
University of Detroit Mercy, shoemadp@udemercy.edu

Data Mining and the Five Pillars of Information Assurance: Where Does Society Draw the Line?

Mary Sparks
University of Detroit Mercy
marysparks@comcast.net

Antonio Drommi
University of Detroit Mercy
drommia@udmercy.edu

Dan Shoemaker
University of Detroit Mercy
shoemadp@udmercy.edu

ABSTRACT

The intent of this paper is to examine the basics of the legal, social and ethical issues implicit in commercial data mining ventures. With the advances in middleware and the enhancements to Business Intelligence tools, mining of the virtual data warehouses is expanding faster than the processes that control them. The question is, "How can organizations apply the 5 pillars of Information Assurance to this mining operation, while not alienating the individuals from whom the information is collected? What are the legalities of confidentiality, and how do we prevent the invasion of privacy. Who is truly the owner of the data being captured, stored, and interrogated?" The ethical questions with regard to the ability to collect versus the correctness of collecting available data, as well as insider snooping of the collected data will be explored. Lastly, the recent findings on the social impact of the data's integrity and authentication will be reviewed in the light of the Choice Point exploit.

Keywords

Information assurance, data warehouse, data mining, business intelligence, ethics.

INTRODUCTION

The thesis of this paper is that, data warehouses and the data mining operations that access them require an ethical model to ensure that security issues are understood and all threats, vulnerabilities and weaknesses associated with the retention and use of data are formalized as a code of conduct. Specifically, this entails itemization of acceptable behaviors to address the 5 pillars of Information Assurance. This paper will employ systems thinking as a means of exploring and integrating the myriad of issues created by the explosion of Data Warehouses and their use of Data Mining.

Systems thinking or systems intelligence attempts to make sense out of the interactions of events that are seemingly unconnected. It involves processing that information into a coherent picture. This paper will look at the question from the perspective of the legal, ethical and social principles that underlie the enforcement of proper use of data warehouses and data mining. Although there have been many papers published outlining the latest developments in this area, the formalized consideration of the IA accountabilities in Business Intelligence (BI) and the effect of the necessary controls has not been attempted before.

Because this entire phenomenon has developed so fast it constitutes un-traveled and uncharted territory for the people working in the field and it has passed completely under the radar of academia. This investigation seeks to review the short history of data warehousing and attempts to characterize emerging patterns in data mining. By doing so it is anticipated that substantive patterns of behavior can be identified. This would guide future data mining enterprises in secure behavior. As well, it is hoped that this research will provide the next generation of data warehouses and the methods used to access them with substantive guidance in the implementation of the five pillars of information assurance. In this investigation the behaviors that were uncovered appeared on the surface to be unrelated, but when viewed with a systems thinking perspective it was possible to demonstrate patterns and attributes that could guide further investigations.

BUSINESS INTELLIGENCE

Business Intelligence is the conversion of data into usable information. The term Business Intelligence (BI) covers many different aspects of analysis in the industry. These include Data Mining, Online Analytical Processing (OLAP), and ad-hoc queries against a data warehouse. The data source for most data warehouses is the production line known as everyday life. Every credit card transaction you make and every phone call you place will at some point or another end up in one if not several data warehouses. The BI industry is growing, as proven by the fiscal spending trend by businesses in North America and Europe. According to a research study done by the Forrester Research group it was determined that business intelligence surpassed security as the #1 expenditure item expected for 2005 (Kemp, 2005).

This study surveyed 1,368, technology decision-makers in North America and Europe, and found that they plan to increase overall spending an average of 3.9%. However, regulatory concerns and an increasing need for data to support decision making kept business intelligence in the top spot among their planned expenditures, at 9%. As a result, a majority of the respondents were also optimistic about their business in the coming year: 54% had a positive outlook, compared with 44% last year (Kemp, 2005).

The numbers are barely available for 2004 but they are staggering. Terradata, a division of NCR, states in their white paper titled Enterprise Data Warehouse a Roadmap, that their major customers started out using 40 gigabytes (gb) and 2 terabytes (tb) are now using 3tb to 200tb for data warehouse storage (NCR, 2004). This continuous increase in the amount of data captured, along with the increase in technology has fallen outside of the capabilities of the traditional means of analysis. That has led to the increased usage of data mining techniques. Sami Hero, Senior Director, Product Marketing, for Hummingbird Ltd., states that data mining is a set of techniques used to uncover previously obscure or unknown patterns and relationships in very large databases (Hero, 2001).

> *The ultimate goal is to arrive at comprehensible, meaningful results from extensive analysis of information. He goes on to explain that the differences between data mining and other analytical tools support a verification-based approach in which the user hypothesizes about the specific data relationships and then uses the tools to verify or refute those presumptions. This verification-based process stems from the intuition of the user to pose the questions and refine the analysis based on the results of potentially complex queries against a database. The effectiveness of this analysis depends on several factors, not the least of which are the ability of the user to pose appropriate questions, the capability of the tools to return results quickly, and the overall reliability and accuracy of the data being analyzed"* (Hero, 2001).

Everything old is new again

Ad-hoc analytical tools have been in use in the IT industry since the mid-70's. University of Michigan's ILIR-MICRO was a 4th generation language used on their Merit Terminal System (MTS) network, SAS, RAMIS, and FOCUS were others that were used against VSAM and flat sequential files on large mainframe systems. What have changed over time are the types of files that these packages or their successors can now access. This type of access software came to be part of a descriptive category known as "Middleware". Middleware, in its Utopian state would be able to access, from anywhere, any data, stored on any platform, in conjunction with any analytical software. The business of access rights, fetch algorithms and messaging information is done "under the covers". The enhancement of this type of software along with the decrease in cost for data storage, and the increase in computing power has been influential in creating the ability to change 200TB of data into a competitive advantage in the business world.

Where is this boom going?

With this competitive advantage, the entire approach of management has changed, leading to the coinage of a new term, Business Process Management (BPM). Companies now understand the importance of enforcing the achievement of the goals defined by their strategy through metrics-driven management (Sveiby, 1997). Thus the new requirement of today's manager is to ensure that all processes are capable of continuous performance assessment. This is done through the use of Key Performance Indicators (KPI), dashboards and balance scorecards. The ability to inform operational mangers with indicators, which reflect strategic vision, is what is known as a closed loop system. A closed loop system allows for decisions to be made closer to the operational tasks and processes for reduced lag time and synchronization of goals.

This requires that the indicators be constantly fed and made available at the right time, at the proper level in the right form. This is where traditional Data warehouse applications end and a new technology known as Business Activity Monitoring (BAM) begin. BAM uses right-time integrator (RTI) to blend the data from the data warehouse with data from operational databases, and dynamic data stores, along with a KPI manager, a set of mining tools and a rule engine to monitor the right time integrator (Golfarelli, Rissi, Cella, 2004).

And therein lies our dilemma

Most of this data is accessed via the Internet. The use of the data warehouses via the Internet is, by virtue of its multi-tier architecture, more difficult to secure. Specifically:

Data originates in a *source* tier. The *derived* tier consists of information derived from the sources. The system should have an integrated security policy, so that each granule of data is protected consistently, whether obtained through the source or derived tier. Creating such a policy requires negotiations between administrators at different tiers (Rosenthal, Sciore, Doshil, 1999).

In addition, the first of many problems begin to occur because of the divergence from the traditional Operational Data Stores (ODS). The timeliness requirements impact data quality and integrity. This means that the fact that operational decisions are being made on hastily retrieved data and data vulnerable to attack is a cause for concern. Total lockdown is not an option because that would impact the availability principle. Therefore, even though there are issues of confidentiality and especially integrity, business must be allowed to continue.

INFORMATION ASSURANCE

So where is this heading? In the case of the Federal Government, the formation of Information Assurance has evolved through three phases, starting with the Communications Security (COMSEC), then Information Systems Security (INFOSEC), and most recently Information Assurance (IA). The website for NSA and NSIP website state that IA is focused on the need to protect information during transit, processing, or storage within complex and/or widely dispersed computers and communication systems networks. According to the National Security Telecommunication and Information Systems Security Committee (NSTISSC) there are five main counter-measures of information assurance (IA) known as the 5 pillars of IA. They are availability, integrity, authentication, confidentiality, and non-repudiation. These pillars and any measures taken to protect and defend information and information systems, to include providing for the restoration of information systems constitute the essential underpinnings for ensuring trust and integrity in information systems.

3 out of 5 isn't bad

Three of these five pillars (availability, integrity, confidentiality) were originally discussed by John McCumber in his paper Information Security: A Comprehensive Model, presented at the 14th National Computer Security Conference, in Baltimore, MD in 1991. McCumber's cube also showed the states in which data resides (transmission, storage, processing) and the counter measures that can be applied (technology, policies and procedures, training). The counter measures have evolved into the Defense in Depth philosophy of the Federal Government. Since 1991, the explosion in usage of the Internet as a means to extend the internal network using virtual private networks (VPN) has caused two additional security services to be added to McCumber's model. The additions are authentication and non-repudiation (Maconachy, Schou, Ragsdale, Welch 2001).

Information Security Controls

The formal process of applying security to IT applications is a well-defined discipline. Several standards exist containing security controls to implement. The most prominent of these is the ISO 17799 Code of Practice for Information Security Management, which was fast-tracked into an international standard from BS7799.1. Waiting in the wings for fast-track status is the BS 7799.2 Implementation Guide, which will turn these principles into certifiable security systems. In addition there is the NIST 800-26 Security Self Assessment guide for Information Technology Systems, the GASSP and COBIT to name a few. In January 2005, NIST placed a draft copy of the Recommended Security Controls for Federal Information Systems special publication 800-53 on the web for public review.

The intent of this publication was to provide guidelines for selecting and specifying security controls for information systems supporting the executive agencies of the federal government. The appendix G of this document contains a mapping of the security control sets to the ISO 17799 and 800-26 guides. In short the draft document suggest the process to select and specify security controls for a system, and assists in the specification by breaking the security controls down into families and classes and gives supplemental guidance for each.

A lot of controls but none quite right

The proliferation of security standards for information systems is an acknowledged fact. What is (understandably) lacking are the specifications for how to specifically secure web-applications, or services sometimes referred to as Web Services. Although there has been phenomenal growth of Web-based applications, access control issues related to Web security have never been made explicit in any standards document (Joshi., Aref, Ghafoor, Spafford, 2001). Moreover, the transformation of our enterprise computer environments to fully distributed, networked client/server or web-based systems dramatically endangers the security of the owners of data warehouses. The authors of the Data Warehouse Lifecycle Toolkit manual state this clearly.

> *"As an industry we are sitting on a time bomb. Our data is grossly exposed and at risk. We face significant legal, and professional challenges in understanding our vulnerabilities and responding to them effectively" (Kimball, 1998).*

To this end, consideration should be given in formulating the next generation of controls. These must be specifically geared toward the use of Internet facing Business Intelligence tools, data warehouses and data mining methodologies.

Begin at the Beginning or Better Yet the Middle(ware)

While the definition of standards for traditional information systems was a daunting task, it was made easier due to the nature of the system universe. There are logically defined data collection and reporting points at the application level and defined user access rules at the infrastructure level. With the advent of the Internet, all boundaries have become blurred. The data in a data warehouse can originate from several sources, and the middleware is the control to the warehouse. The final user is usually unknown to the data warehouse. Once the data is returned it can be converted to an Excel spreadsheet, E-mailed to a customer list or sent as an alert to the production manager's cell phone. In each of the examples cited here the access authority to the data warehouse is defined by the middleware, not the individual user requesting the data. It is the middleware that controls the access, and the vulnerability to the middleware is paramount. Therefore, risk mitigation and not risk annihilation should be the goal when the pillars of accessibility and authentication are placed in the security policy for data warehouses and data mining.

Web Services security must integrate security across infrastructure and applications and be able to dynamically negotiate the security access level rights at run time of loosely coupled applications. *"The term "web services" refers to the delivery of software as services and components across the internet. A web service is a modular application used to describe, publish, locate and invoke services across the web" (Bhatia, 2005).* Standardization initiatives started, but until they are completed client-server technology will continue to be deployed for security that will be able to support a few hops in web services (Bhatia, 2005).

LEGAL ISSUES

As in any other circumstance, where vulnerabilities exist, exploitation occurs. When exploitation occurs, lawyers show up. The courtroom brings plaintiff and defendants together, and in any court, proof is the key to justice. Of the five pillars, non-repudiation is that proof. Non-repudiation confirms that what was transmitted was received and vice versa. In the discussions of legal issues as they relate to data processed from the data warehouse or mined, the ability to trace the data flow transmission must be available. One of the most prominent legal issues with regards to data warehouses and data mining is the intrusion of privacy and the chance for abuse of the trust given in exchange for the information collected.

That ties directly to the one "counter-pillar" which is availability. Availability ensures that authorized individuals can access data. Its goal is to get the maximum data into the hands of the users. That obviously has to be controlled by something, which is the authorization function. The problem is that, thanks to the Internet too much private data is accessible to operations whose authority is questionable and it, in turn, is too accessible to unauthorized uses. Thus increased access to personal data by technology and humans has intensified privacy issues, raised consumer awareness and resulted in legislation with penalties for abuse (Romney, Romney, 2004). Many companies present privacy policies on their Web sites, but these policies are often offensively paternalistic and one-sided. If you want to continue using their Web-site you must agree to their policy. Many examples of what is referred as predatory privacy policies exist at the most prominent companies. These policies can actually harm sales. When prompted for pre-purchase demand of personal data 80% of Internet shoppers routinely abandon their orders.

Same song, different beat

With respect to Confidentiality and Integrity, one tends to assume that the issue of privacy is something brought on by the information age, but one of the most influential articles regarding the discussion of privacy occurred in 1890. It was published in the Harvard Law Review, and defined privacy as it is still defined today "to be let alone" (Brandies, Warren., 1890). This article stated that privacy was under attack by the recent inventions and business methods. Still, much new legislation in Europe and the United States has been created, is aimed at protecting the privacy of the information being collected, stored, manipulated and dispersed. The need to become familiar with the international and national laws governing data privacy is necessary to be able to create valid security controls.

EUDPD, PIPEDA, HIPAA and the GLB

The Privacy Act of 1974 and modified by the Computer Matching Act of 1988, which was established to protect the individual's privacy from the federal government's collection of personal information. It requires the federal government to give notice to and consent from individuals when the government collects, or shares information about them, unfortunately the act does not apply to commercial databases (Dempsey, Flint, 2003). In addition to legislation to protect privacy there is legislation to limit privacy, The Patriot Act, passed in October of 2001, which authorized for both wiretap and grand jury information to be provided to "any federal law enforcement officer, intelligence, protective, immigration, national defense or national security official" for the performance of his official duties (Markle Foundation, 2002). In 1995 the European Union Data Protection Directive (EUDPD), specified the kind of data that can be collected, and for what purpose, along with how that information must be handled in transmission and storage. This directive caused the self-policing "safe harbor" to take effect to offset a possible trade war. Europe's protection of consumer privacy is not unique. In April of 2002 Canada put into law the Personal Information Protection and Electronic Document Act (PIPEDA) which contains a set of rules for Canadian and non-Canadian businesses regulating the processing of personal information during commercial activities.

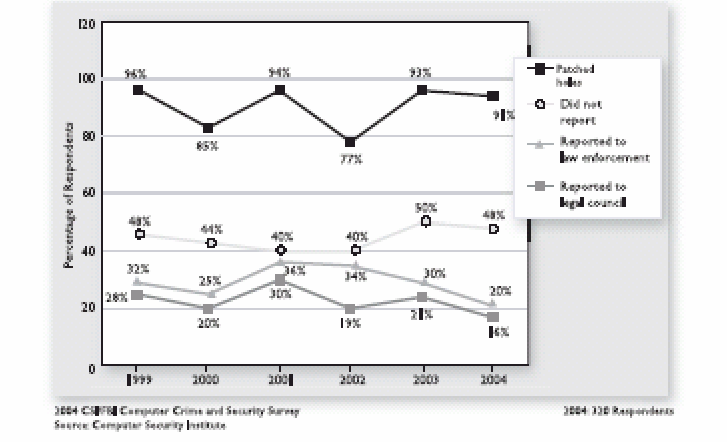
The United States has passed legislation to protect certain types of data. In the area of healthcare, the Health Care Portability and Accountability Act (HIPAA), includes provisions for limitations of use of personal medical information, restrictions on marketing medical information; and confidentiality of doctor patient communications. In the area of finance, the Gramm-Leach-Bliley Act (GLB) forces financial institutions to send out to customers (not consumers, defined as not having a continuing relationship to a financial institution) (Romney, et al., 2004). As a result of each of these legislations, U.S. companies have been brought into court. The pillar of confidentiality as it applies to data warehouse and data mining and use with Business Intelligence tools must definitely take these legal issues into account when defining the tasks associated with preparing and maintaining the security controls for this endeavor.

SOCIAL ISSUES

These laws are a beginning but the individual is still left vulnerable in regards to personally identifying data collected by commercial enterprises. What is the social impact to the individual whose privacy is lost? Privacy is important for the development of the individual; because privacy has been described as the basis for self-identity and friendship, intimacy, and trust all need privacy in order to develop. In extreme cases, others may have more information about you than you have yourself. This could in fact jeopardize you ability to develop an identity (Severson, 1997, p.65). Of higher concern would be what if the data collected was in error? The social stigmatism of being incorrectly listed on a Web site as a pedophile has definite negative social aspects. Data Integrity as an Information Assurance pillar enforces the policy of standards of validation and verification. These must be included in all ETL design for loading of the data warehouse, and the methodologies for data mining must be accountable for logic verification.

What is the impact when the details of a security incident are published? In Information Security Policies in Large Organizations, the topic is addressed such that… "Organizations are reluctant to publish details of security incidents which have occurred, partly because of the effect this might have on shareholders, customers, and public confidence, and also the embarrassment of making such events public knowledge" (Doherty, Fulford, 2004).

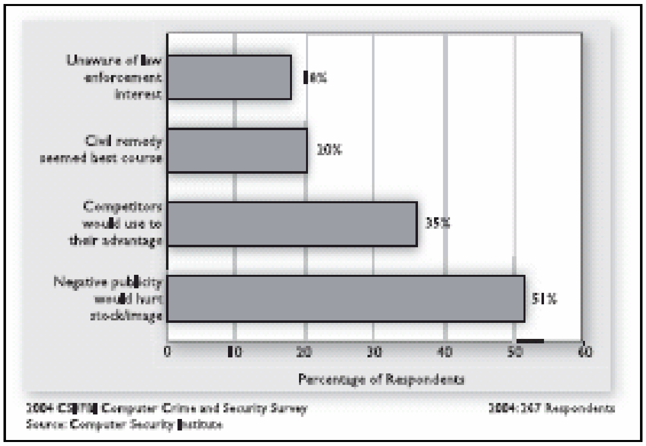
This is more recently proven true by the 2004 CSI/FBI Computer Crimes and Security survey, which stated that the number of incidence reported to law enforcement officials was down and the reason most often given was the negative publicity (CSI/FBI, 2004). This neglectful behavior is certainly not illegal, but borders on the un-ethical.



2004 CSI/FBI Computer Crime and Security Survey
Source: Computer Security Institute

2004: 320 Respondents

Figure 1. If your organization has experienced computer intrusion(s) within the last 12 months, which of the following actions did you take? (reproduction permission received)



2004 CSI/FBI Computer Crime and Security Survey
Source: Computer Security Institute

2004: 267 Respondents

Figure 2. Reason organization did not report intrusion to law enforcement: Percentage of respondents identifying as important

According to the same study, the number of insider abuse incidents reported was steadily declining. Although the number of attacks from inside the organization still out number those from outside... *"The issue is trust, insiders must be trusted to do their jobs; applications must be trusted to perform their tasks. The problem occurs when insiders –be they users or applications- intentionally or unintentionally extend trust inappropriately* " (Thompson, Ford, 2004).

User and administrator naivety contribute to the insider threat. Most applications and network architectures offer a all-or-nothing trust model. This model facilitates the extending of privileges to a malicious user or application.

The responsibility of the corporation to train its employees is a given. However, unfortunately the money spent on employee training is sometimes taken from the same budget that is needed to secure the operation. The departmental budget wars within organizations are notorious. In almost all cases the bottom line is the final answer. From whose budget will training dollars come, once departmental data is sourced to the data warehouse, which department continues to own it? And who then is responsible for incurring cost associated with it misuse?

ETHICS ISSUES

Money is always a deterrent to ethics. Mary Ann Davidson, Chief Security Officer for Oracle makes that point very well in describing what higher learning can do to help instill ethical behavior for its students. " *If business schools select people who value money over everything else, they're not going to be able to teach these students in two short years, that doing the right* *thing is more important than making money."* She goes on to state that ethical behavior as it pertains to computers, and computer resources should be part of the curriculum, and that colleges and university have codes of conduct that should be more often enforced.

Collection of data is a major ethical issue in the data-mining arena. An example of the amount of data that is currently being collected at WalMart (500 terabytes) is 10 time what the Internal Revenue Service accumulates (50terabytes) (Romney, et. al., 2004). While on the surface the fact that WalMart can track every purchase in all of their stores for an ever growing time dimension may not make you nervous. What if that information is shared with the government or other organizations? In fact the sharing of data is apparent in the use of credit scores to calculate automobile insurance premiums. Just because the data is available to collect, should it be collected and shared? The temptation to use the data for other purposes is so great. In fact with the expansion of the use of biometrics to secure one's personal data and files it can in fact decrease privacy by more clearly identify the person and be used for other purposes (van der Ploeg, 2001).

CONCLUSION

The five pillars of Information Assurance, Availability, Integrity, Confidentiality, Authentication, and Non-Repudiation are excellent checks in the security plan of data warehouses and the data-mining methods that use them. As mentioned earlier in the paper, the middleware must be the first priority. Several initiatives are underway but the end is 3 to 5 years down the road. The extract, transform and load software (ETL) used to generate the source data must be guarded for security breeches at all times, not just during development. The ad-hoc nature of the development environment must be stifled, and the process that controls the source data for the warehouses must be placed under normal configuration management. As mentioned in regards to the BAM closed loop systems, this immediate feed process is only going to become a higher priority. Practices must be defined, refined, and optimized for this purpose. We have today at our fingertips, the freedom to explore the largest expanse of information ever available in the history of mankind, but with that freedom brings responsibility. We, as the security professionals standing at the precipice of this great expanse must be vigilant to act ethically and socially responsible in all matters and chart the course for those who come after us.

REFERENCES

1. Amon Rosenthal, Edward Sciore, Vinti Doshi1, (1999) Security Administration for Federations, Warehouses, and other Derived Data IFIP Conference Proceedings; Vol. 171, Proceedings of the IFIP WG 11.3 Thirteenth International Conference on Database Security: Research Advances in Database and Information Systems Security.
2. Bhatia, M. Web Services Security, Information Systems Control Journal, Volume 1, 2005.
3. Brandies, L.D., and S. Warren, S., "The Right to Privacy. The Implicit Made Explicit," in Harvard L.R., 4; (1890) 193-220.
4. Crawford-Mason, C. (2004) Watergate's Deep Throat-A Systems Thinker, in Quality Progress 37, no. 11, 61-67.
5. CSI/FBI 2004 Computer Crimes and Security Survey, 2004, Computer Security Institute.
6. Dempsey, J., Flint, L., Privacy's Gap: The Largely Non-Existent Legal Framework for Government Mining of Commercial Data, Center for Democracy and technology, May 28, 2003.
7. Doherty, N., Fulford, H. (2004), Information Security Policies in Large Organizations: The Development of a Conceptual Framework to Explore their Impact, Information Security & Ethics: social and organizational issues, 238-260.
8. Enterprise Data Warehouse Roadmap Modeling - A state-of-the-art method for planning and prioritizing your EDW implementation, NCR Corporation, 2004.
9. Golfareli, M, Rizzi, S., Cella, I, (2004), Beyond Data Warehousing: What's Next in Business Intelligence? DOLAP'04, November 12-13, 2004, Washington, DC, USA.
10. Hero, S., (2001) Hummingbird Ltd. What Works: Volume 11, May 2001, The Need for Enterprise Data Mining Solutions, What Works? Volume 11.
11. Information Assurance, What is IA? www.nsa.gov, accessed January 15, 2005.
12. Kemp, Ted, February 10, 2005. Business Intelligence Market Blooming , February 19, 2005. http://www.compliancepipeline.com/60300146, page 2
13. Kimball, R., Reeves, L., Ross, M., Thornthwaite, W., The Data Warehouse Lifecycle Toolkit, expert methods for designing, developing, and deploying data warehouses, Wiley & Sons, Inc. New York, NY, 1998.

14. Maconachy, W., Schou, C., Ragsdale, D., Welch, D., A Model for Information Assurance: An Integrated Approach, Proceedings of the 2001 IEEE Workshop on Information Assurance and Security United States Military Academy, West Point, NY, 5-6 June.
15. Markle Foundation (2002), Protecting Americas Freedom in the Information Age, A report on the Markle Foundation Task Force.
16. NIST Special Publication 800-53 Recommended Security Controls for Federal Information Systems
17. Romney, V., Romney, G., Neglect of Information Privacy Instruction – A case of Educational Malpractice?, SIGITE '04, October 28-30,2004 Salt Lake City, Utah, USA
18. Stahl, B.C. (2004), Responsibility for Information Assurance and Privacy: A problem of Individual Ethics?, Journal of Organizational and End User Computing, Vol. 16, no.3.
19. Severson, R. J., (1997) The principles of information ethics. Armok: M.E. Sharpe.
20. Sveiby, K.E., The New Organizational Wealth: Managing and Measuring Knowledge Based Assets. Berret Koehler Publishers, San Francisco, CA 1997.
21. Teradata a division of NCR (2004), Enterprise Data Warehouse Roadmap Modeling: A state-of-the-art method for planning and prioritizing your EDW implementation.
22. Thompson, H., Ford, R., The Insider, Naivety, and Hostility: Security Perfect Storm?, Queue, June 2004.
23. van der Ploeg, i. (2001). Written on the body: Biometrics and identity. In R.A. Spinello & H.T. Tavani (Eds.), Readings in cyberethics (pp.501-514). Sudbufy, MA: Jones and Barlett.