Data Sharing: A Fundamental Need for Black Sea Region and European Security

Olcay KURSUN*, Cabir ERGUVEN*, Kenneth M. REYNOLDS*

Abstract: With the increases in the computer technology, many law enforcement agencies store their data electronically in databases. However, every agency requires keeping the ownership of their data for obvious reasons, such as data security, data up-to-datedness, synchronization and reliability, simplicity of data management and policy making, and implausibility of centralized/global data control mechanisms. Police agencies find the computerized systems useful for simplifying bookkeeping and speeding up simple local searches; however, they do not find it very useful in alerting them to potential terrorists or other miscreants. Proponents point out that local crime data can lead to big breaks when shared. September 11 mastermind Atta was stopped for a traffic violation in Delray Beach, Fla., in mid-2001 -- but he was let go because officers did not know of a bench warrant for him the next county over due to lack of information sharing capabilities. Data sharing is of grave importance to law enforcement and it must be understood with the issues it brings along, such as distributed and compound queries, dirty data, and assessment and evaluation metrics.

Keywords: Homeland security, terrorism, shared databases, information sharing, approximate name matching, dirty data, evaluation metrics, compound query, distributed query, FINDER.

1 Introduction

With the advances in computer technologies, large amounts of data are stored in databases that need to be efficiently shared, searched, and analyzed [1]. The terrorist attacks of "9-11" have lead to an increased emphasis on the use of information technology to facilitate information sharing among and between local, state, federal, and international agencies [2].

Information sharing initiatives should include the integration of key data from internal systems to enable easy access to the needed data, ideally through one standardized user-interface. This finding has implications for national security and antiterrorism systems. As these systems grow in breadth to support security and antiterrorism efforts, they will include local, statewide, national, and even international data, thus adding complexity to both data and searches [3, 4]. For example, with the increased number of records that organizations keep the chances of having "dirty data" within the databases (due to aliases, misspelled entries, ethnic factors etc.) increases as well [6, 7].

* Department of Computer Engineering, International Black Sea University, Tbilisi, Georgia, okursun@ibsu.edu.ge

* Department of Computer Engineering, International Black Sea University, Tbilisi, Georgia, cabir@ibsu.edu.ge

* Department of Criminal Justice and Legal Studies, University of Central Florida, Orlando, FL, USA, kreynold@mail.ucf.edu

In this paper, we present a case study, FINDER, a working data sharing system in Florida and soon to be including other states of the United States. We believe in the Black Sea region and its prospective relations with the European Union data sharing will play an important role in taking necessary security measures. We describe FINDER in Section 2. In the subsequent sections, we stress the issues emerging as a result of sharing data, such as dirty data, compound queries and text mining, determining assessment and evaluation metrics. We conclude in Section 6.

2 The Operational Data Sharing Environment -- FINDER

FINDER – the Florida Integrated Network for Data Exchange and Retrieval – has been a highly successful project in the state of Florida that has addressed effectively the security and privacy issues that relate to information sharing among more than 120 law enforcement agencies as of May 2006. It is operated as a partnership between the University of Central Florida and the law-enforcement agencies in Florida sharing data – referred to as the Law Enforcement Data Sharing Consortium. Utilizing a federated query framework, this software platform enables officers to search for information and reports from any agency within the system. The system uses a GJXDM-compliant (GJ stands for Global Justice standard) query middleware tier that is scalable, flexible, and low cost. The FINDER architecture is both dynamic (data is added to it continuously) and distributed (data and resources are distributed over a number of "FINDER nodes" = agencies). Figures 1 and 2 depict an overview of the FINDER system and a detailed configuration of a FINDER node. The existing FINDER system allows the law enforcement agencies to exchange data in efforts to prevent criminal activity and more efficiently solve crimes.

As of May 2006, FINDER has solved close to 400 documented criminal cases including burglaries, armed robberies, and attempted murder. Detectives report that without FINDER, these cases may have gone unsolved and violent and dangerous career criminals might still be on the streets. The interoperability that this system offers supports traditional crime suppression objectives and is crucial in this era of heightened domestic security. This system will allow access to an unmatched amount of information that was previously inaccessible. It can be utilized by every member of the agency. This information provides an opportunity for agencies to address crime control issues that cross jurisdictional boundaries. It can also save countless man-hours by allowing agency personnel to query a system to obtain information that they otherwise would attempt to obtain by making numerous and time consuming phone calls.

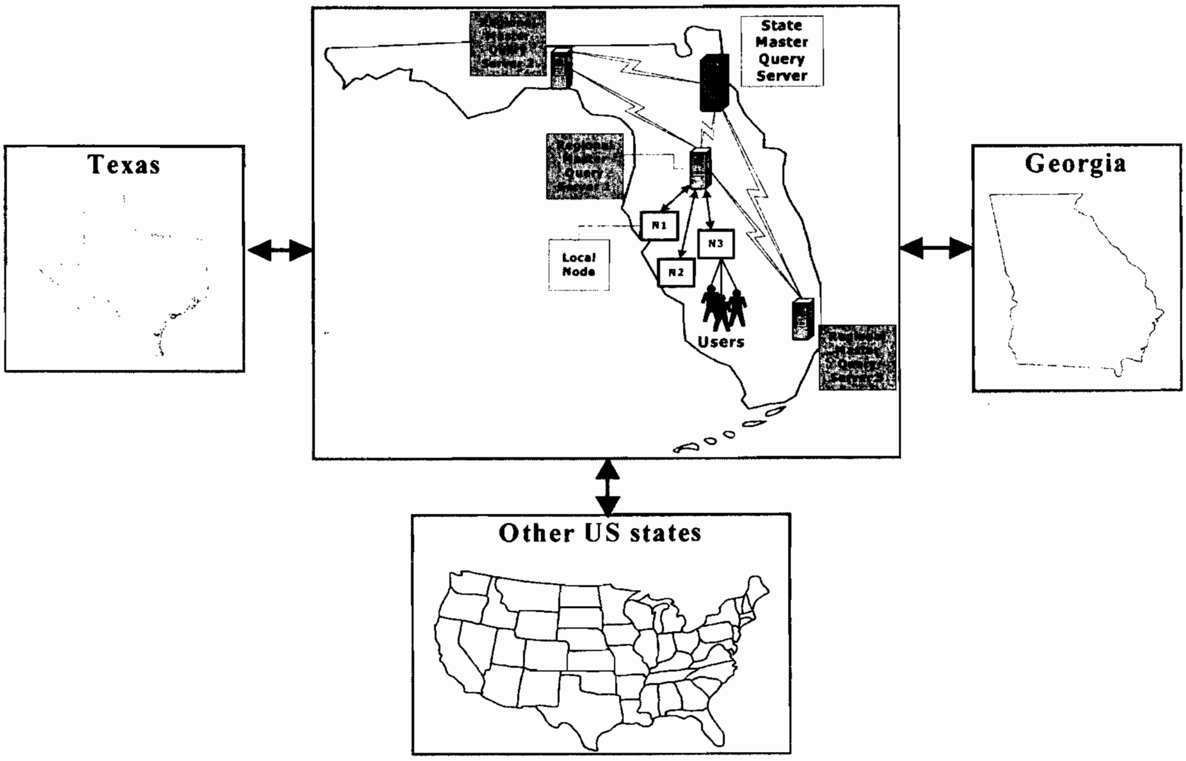

Regional Master Query Server 2
State Master Query Server
Regional Master Query Server 1
N1
N3
N2
Local Node
Users
Regional Master Query Server 3
Texas
Georgia
Other US states

Figure 1: The general overview of the FINDER network in Florida and expanded to other states

A major achievement of FINDER is the successful targeting of property offenders who routinely operate across city and county lines. For years, individuals would burglarize homes and businesses in one jurisdiction and then sell the stolen property in another city or county. With effective information sharing across jurisdictions, property detectives have the increased ability to locate stolen property and identify and apprehend these repeat offenders. Today, it is not uncommon for a burglar arrested through the use of FINDER to tell detectives that he thought he would go undetected by stealing in one county and selling in another. To date, over $1 million in stolen property has been recovered through successful investigations driven by FINDER.

Detailed information about the organization of the data sharing consortium, success stories, and the FINDER software is available at http://finder.ucf.edu.

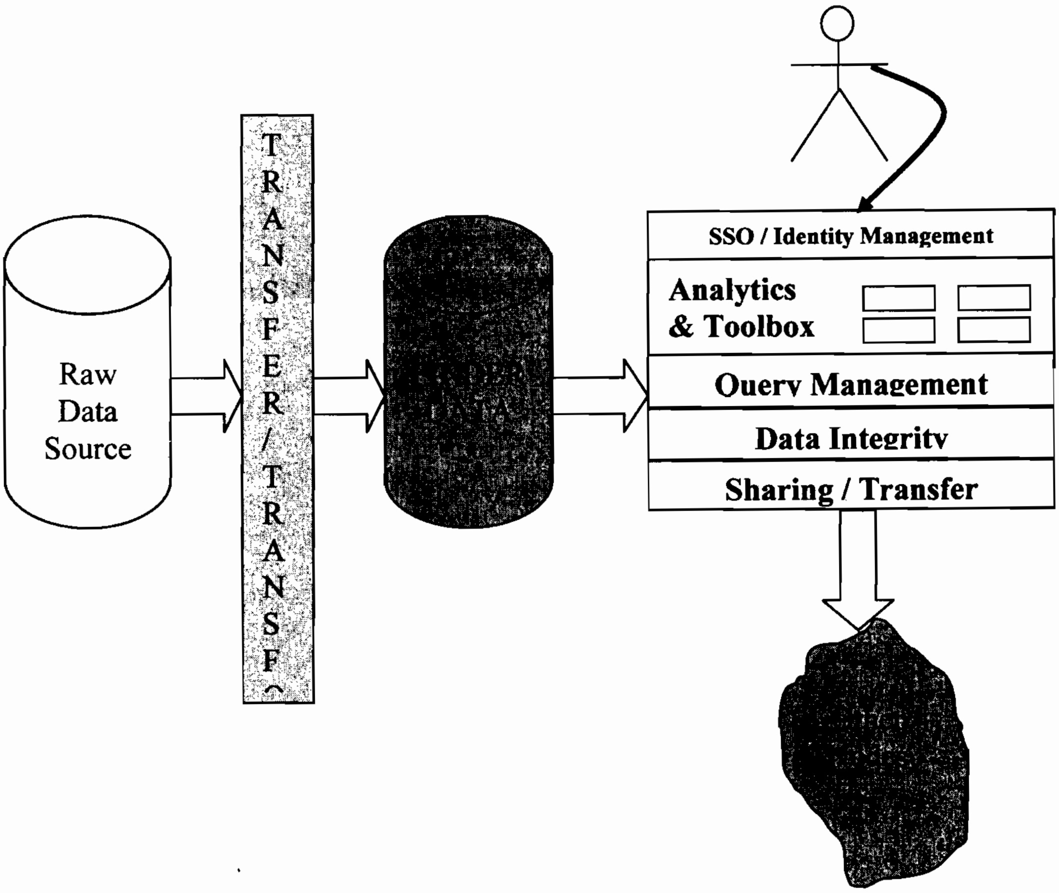


Figure 2: Detailed components of a FINDER node

3 Dirty Data: Practical and Ethnic Factors

Part of the constraints of the FINDER system and also most law enforcement records management systems is that once the data has entered into the system it must remain intact in its current form. This includes data that have been erroneously entered, and consequently they contain misspellings. In particular, in the presence of dirty data, a search for specific information by a standard query (e.g., search for a name that is misspelled or mistyped) does not return all needed information. This is an issue of grave importance not only in homeland security and criminology, but also in medical applications, GIS (geographic information systems), customer services, and so on. This problem was identified by the FINDER team and has also been substantiated in the literature[6, 8, 9, 10]. Therefore, prior to the implementation of any algorithm to analyze the data, the issue of determining the correct matches in datasets with low data integrity must be resolved. The problem of identifying the correct individual is indeed of great importance in the law enforcement and crime analysis arenas. For example, when detectives or crime analysts query for individuals associated with prior burglary reports, they need to be able to examine all the records related to these individuals, otherwise they might miss important clues and information that could lead to solving these cases.

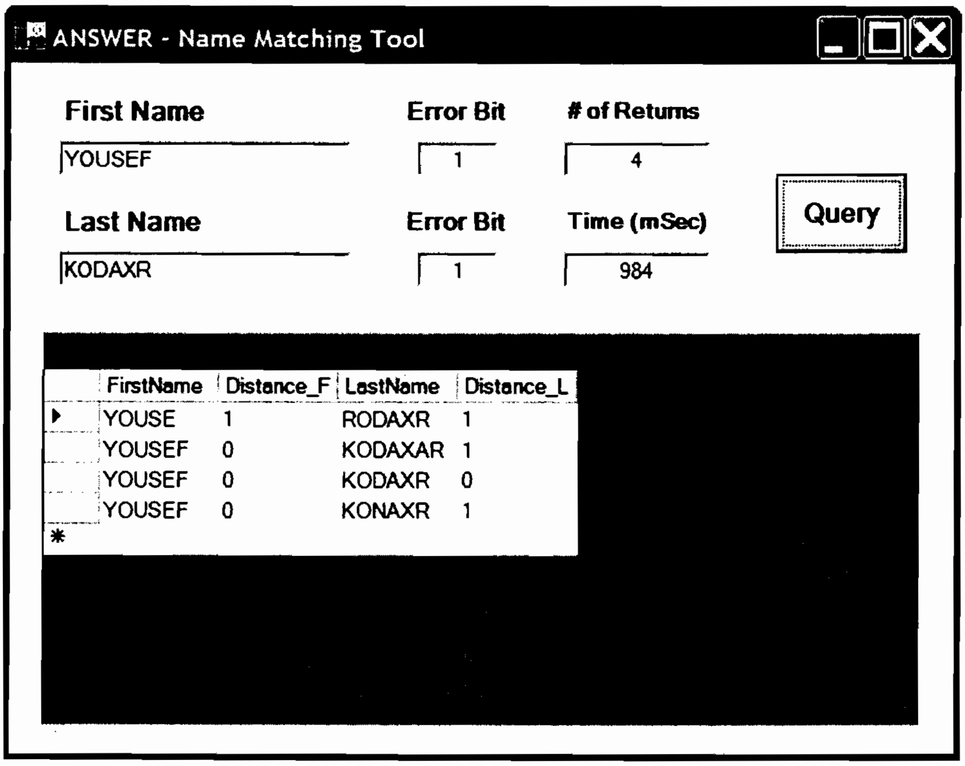
Incorrect data entries occur more frequently due to language differences in international data sharing or when the data consist of names of individuals from diverse ethnicities and languages; not to mention that criminals try to slightly modify their names and other information in order to deceive the law enforcement personnel and evade punishment, which is easier to manage when the suspects are foreign. Another reason is that for a large number of cases, the name information might come from witnesses, informants, etc., and therefore this information (for example the spelling of a name) is not as reliable as when identification documents are produced. This turns out to be an important issue in the field of counterterrorism, where a lot of information comes from sources that might be unreliable, but which still needs to be checked nevertheless. It is evident then that it is imperative to have an efficient and accurate name matching technique that will guarantee to return all positive matches of a given name.

A simple illustration related to name matching, utilizing dirty data available in the FINDER system, is shown in Table 1, which emphasizes both the level of data integrity and the challenges of using standard SQL queries to retrieve records from a law enforcement database (also known as merge/purge problems [10]). In Table 1, we are depicting the results of an SQL query on "Joey Sleischman". An SQL query will miss all the records but the first one. The other records could be discovered only if we were to apply an edit distance algorithm [11] on all the existing records in the database, an unsuitable approach though, due to its high computational complexity, especially in large databases. In particular, the rest of the records (besides the exact match), shown in Table 1 were identified by comparing the queried record ("Joey Sleischman") against all records in the database (by applying the edit distance approach). The Last Name, First Name, DOB (Date of Birth), and Sex were used as parameters in this search. In order to detect the matching records, we assigned weights to the fields: Last Name (40%), First Name (20%), DOB (30%), and Sex (10%). We used the edit distance algorithm [11] for determining the degree of match between fields.

Table 1. Example of the Data Integrity Issues within the FINDER data.

Last Name	First Name	DOB	Sex	Match
INPUT QUERY:				
SLEISCHMAN	*JOEY*	*1/21/1988*	*M*	≥ *85%*
MATCHING RECORDS:				
SLEISCHMAN	JOEY	1/21/1988	M	100%
SLEICHMAN	JOEY	7/21/1988	M	91%
SLEISCHMANN	JOSEPH	1/21/1988	M	88%
SLEISCHMANN	JOSPEH	1/21/1988	M	88%
SLEISHMAN	JOEY		M	87%
SLEISCHMANN	JOEY		M	87%
SLEISHCHMANN	JOSEPH	1/21/1988	M	86%
SLESHMAN	JOEY		M	85%

As it can be seen in Table 1, the edit distance algorithm provides an excellent level of matching, but the algorithm requires a full table scan (checking all records in the database). This level of computational complexity makes it unsuitable as a technique for providing name matching in applications, such as FINDER, where the number of records is high and consistently increasing. For detailed analysis of name matching techniques and a method proposed to alleviate this complexity, refer to Kursun et al. 2006 (see Figure 3).



First Name: YOUSEF
Error Bit: 1
# of Returns: 4

Last Name: KODAXR
Error Bit: 1
Time (mSec): 984

FirstName	Distance_F	LastName	Distance_L
YOUSE	1	RODAXR	1
YOUSEF	0	KODAXAR	1
YOUSEF	0	KODAXR	0
YOUSEF	0	KONAXR	1

Figure 3: A simple implementation of ANSWER (Approximate Name Search With ERrors) in FINDER environment. *Error Bit* specifies how many edit errors can be tolerated in the search. *Distance* tells how many edit errors are actually present in the returned names for both *First Name* and *Last Name*, respectively.

4 Compound Queries: Theoretical Factors and Database Design

The advanced search capabilities must allow a user to query the persons, vehicle and pawn data bases with a single query. This query must allow the user to make inquires regarding a person, vehicle or pawned item and receive a return that identifies possible suspects and/or relationships to each other. This must include how each is related to each other. Example: A user can make a query on a Hispanic male, 5'-6", tattoo right arm, driving a red pickup truck and receive a return of all Hispanic males, 5'-6" with a tattoo on the right arm and associated with a red pickup truck.

Currently, in many law enforcement databases, if not all, a big problem associated with mining the data for terrorist activities or cells is that there are no explicit data fields that indicate whether an incident relates to terrorism. While there are efforts in law enforcement agencies to close this gap of not collecting the useful data in regard to terrorism, or collecting the data inefficiently into unstructured text in the narratives, in the literature there are research efforts to make the best use of whatever data available in narratives by moving the unstructured text into structured form (i.e. fields of the database), which involved natural language processing and understanding. However, generalized autonomous natural language understanding is beyond current computational methods. The problem is currently undecidable in the general case, and intractable even for well defined corpus of any scope. There are three basic methods used to overcome this problem: limit word meanings, limit the corpus size, and limit sentence structure. Limitations on word meaning reduce ambiguity and allows for more precise sentence parsing and extraction of meaning from the sentence. Limitations on corpus reduce the time required to establish the structural and semantic role of individual words within statements. Limitations on sentence structure reduce the ambiguity of sentence structures (so called structural and deep structural ambiguity) and allow more precise comprehension of semantics by extricating syntactic and semantic forms.

5 Assessment and Evaluation Metrics: Bureaucratic Factors

In order to assess the performances of a data sharing system, metrics must be determined. These measures have emerged as results of not only academic interpretation but also with discussions with police agencies and criminal justice experts in the consortium. Below, we list a number of metrics, some of which are already employed and some are yet to be implemented in our FINDER system to measure the success of the analytical and data integrity tools developed. They are distinguished in three categories: Process Measures, Performance Impact/Outcome Measures and Usability Measures. Process measures are statistical measures that are relative easy to collect and simply require the manpower to perform the collection tasks. Performance Impact/Outcome measures require significantly more thought in their design in order to be appropriately measured. In many cases, indirect measures may have to be employed to effectively gauge these measures. Usability measures can be measured by collecting direct user feedback, both through the system and through user surveys and focus groups.

i) Process Measures

1. Number of search tasks completed
2. Amount of time for completing search tasks
3. Number of journal entries detailing search experiences
4. Number of jurisdictions/agencies using the system
5. Type of data available on system
6. Training provided on system
7. Number of cases with missing data when data should exist

ii) Performance Impact/Outcome Measures

1. Change in user job performance (e.g., cases cleared, investigations conducted)
2. Change in productivity level (e.g., decrease in time to search for information)
3. Change in accuracy of information obtained
4. Effectiveness of information (e.g., in reducing crime, solving crime)
5. Change in time to obtain information
6. Change in time for case processing
7. Change in ability to apprehend suspects or clear cases or make arrests

iii) Usability Measures

1. Level of satisfaction concerning interaction with the system
2. Efficiency of computer screen design use for task completion
3. Organization of information on the computer screen
4. Ability to find information
5. Level of effort required to use system
   (e.g., the amount of time taken to complete a task)
6. Level of ease in learning how to use the information sharing system
7. Navigation ease for obtaining information
8. Time to complete a task

6 Conclusions

In this paper, we presented a data sharing case study, FINDER, that is currently employed in Florida and has real data from law-enforcement databases. FINDER (the Florida Integrated Network for Data Exchange and Retrieval) has been a highly successful project in the state of Florida that has addressed effectively the security and privacy issues that relate to information sharing among more than 120 law enforcement agencies as of May 2006. It is operated as a partnership between the University of Central Florida and the law-enforcement agencies in Florida sharing data – referred to as the Law Enforcement Data Sharing Consortium. Data sharing is a must-have technology in today's world; however, there are bureaucratic (e.g. local or central data ownership, evaluation and assessment), theoretical (e.g. topology of the network, the node and query architectures), and practical issues (e.g. dirty data) need to be thought through. In this paper, we addressed some of these issues in the light of our experiences with FINDER. We believe that the Black Sea region will soon need to develop and use efficient data sharing systems and address these issues.

References

1. Brown, M. M. (2001, June). The Benefits and Costs of Information Technology Innovations. Public Performance and Management Review. Vol. 24, Number 4. 361-366.
2. Bureau of Justice Assistance. (Feb, 2002). Mission Possible: Strong Governance Structures for the Integration of Justice Information Systems. US Department of Justice, Office of Justice Programs, Bureau of Justice Assistance. NCJ 1922278.
3. Baird, Z & Vatis, M.A. (2003). Creating a Trusted Network for Homeland Security: Second Report of the Markle Foundation Task Force. The Markle Foundation.
4. Taipale, K.A. (2003). Data Mining & Domestic Security: Connecting the Dots To Make Sense of Data. The Columbia Science & Technology Law Review, vol. V. Pages: 42-44.
5. Zaworski, M. (2004). Assessing an Automated, Information-sharing Technology in the Post "9-11" Era: Do Law Enforcement Officers Think it Meets their Needs? (Doctoral dissertation, Florida International University, Miami, Fla.)
6. Kim, W. (2002) "On Database Technology for US Homeland Security", *Journal of Object Technology*, vol. 1(5), pp. 43–49.
7. Taipale, K.A. (2003) "Data Mining & Domestic Security: Connecting the Dots to Make Sense of Data", *The Columbia Science & Technology Law Review*, vol. 5, pp. 1–83.
8. Wilcox, J. (1997) "Police Agencies Join Forces To Build Data-Sharing Networks: Local, State, and Federal Crimefighters Establish IT Posses", *Government Computer News*, Sept. 1997.
9. Maxwell, T. (2005) "Information, Data Mining, and National Security: False Positives and Unidentified Negatives", *Proceedings of the 38th Hawaii International Conference on System Science.*
10. Hernandez, M., and Stolfo, S. (1998) "Real-world Data is Dirty: Data Cleansing and the Merge/purge Problems", *Data Mining Knowledge Discovery*, vol. 2, pp. 9-37, 1998.
11. Levenshtein, V.L. (1966) "Binary codes capable of correcting deletions, insertions, and reversals", *Soviet Physics*, Doklady, vol. 10, pp. 707–710.
12. Kursun, O., Koufakou, A., Wakchaure, A., Georgiopoulos, M., Reynolds, K., Eaglin, R. (2006) "ANSWER: Approximate Name Search With errors in Large Databases by a Novel Approach Based on Prefix-Dictionary", International Journal on Artificial Intelligence Tools (accepted and to appear).