From Police Reports to Data Marts: a Step Towards a Crime Analysis Framework

Fabrizio Albertetti and Kilian Stoffel

Information Management Institute, University of Neuchâtel, Switzerland
{fabrizio.albertetti, kilian.stoffel}@unine.ch [*]

Abstract. Nowadays, crime analyses are often conducted with computational methods. These methods, using several different systems (such as decision support systems), need to handle forensic data in a specific way. In this paper we present a methodology to structure police report data for crime analysis. The proposed artifact is mainly about applying data warehousing concepts to forensic data in a crime analysis perspective. Moreover, a proof of concept is carried out with real forensic data to illustrate and evaluate our methodology. These experiments highlight the need of such framework for crime analysis.

Keywords: Crime analysis, data warehousing, data marts, computational forensics, decision support systems, police reports.

1 Introduction

Over the last decades digital treatment of data and information related to criminal activities has gained drastically in importance. Usually, this information is stored in police logs or police reports (hereafter referred to as “logs” for the sake of brevity). Such information may serve as a basis for crime investigation or may provide important pieces of evidence in prosecution perspectives [1]. One specific type of forensic data we will consider in the remainder of this paper are documents dealing with criminal activities as well as police activities, containing mainly crime incidents, arrests, calls for service, and accidents. Generally focused on pragmatic and operational needs, the structure of these logs is not designed for decision support systems (DSS) [2], [3] —data mining, online analytical processing (OLAP), reporting, dashboard, etc.— and hence not for complex crime analysis. Despite this fact, in the case where logs are directly stored in a database, such forensic sources may still be directly used as an analytical tool by some analysts. To have such “turn key” solutions may have numerous advantages such as decreasing costs or reducing development time to name only two, but none of them have theoretical justifications.

In this paper, we propose a methodology to fill this gap. It leverages the inherent information in forensic data by specifying in a first step a wider an-

[*] This project is supported by the Swiss National Science Foundation (SNSF) [*project 135236*]

alytic framework. The proposed artifact enables this feature by applying well-established methods from the field of business intelligence to the computational forensics discipline. More particularly, the concepts we propose belong to the data warehousing area: corporate information factory, multidimensional modeling, and its underlying data marts, are the main ones. The proposed methodology, consisting of a process and some key success factors, is about designing data models. Once these models are implemented, the crime analysis framework is ready to be used by a crime analyst (i.e., the use of specifically required DSS will be supported/enabled by an underlying ad hoc architecture for data based on the logs).

Beside this, these following questions are addressed:

- do analytical biases decrease by having an ad hoc data structure,
- how to conduct an analysis based on multiple data sources,
- and, how to organize logs to get normalized and consistent data.

The remainder of this paper is structured as follows: Sect. 2 presents related research and highlights the need of the proposed artifact, Sect. 3 lays the groundwork by introducing briefly some data warehousing and data mart concepts. Then, the proposed methodology is described in Sect. 4 in a way such to be as generic as possible in respect to forensic issues and illustrated in Sect. 5. Sec. 6 evaluates the methodology through a case study based on real forensic data (police reports). To conclude, we postulate that an efficient data warehouse and its subsequent data marts contribute to a better crime analysis framework and discuss the idea of a potentially wider analytical framework including this artifact as a starting point.

2 State of the Art

Computational forensics is about applying *computational* methods from several disciplines in the forensic domain. In the sense of [4], these methods support forensic sciences in three ways: (a) they provide tools to overcome limitations of human cognitive ability, (b) a large volume of data is potentially usable for analyses and is not anymore limited to the human mind, and (c) human expert knowledge may be numerically represented to teach inference machines.

Many applications of computational forensics have been approached following these three basic ideas. From the evolution of automated fingerprint identification systems [5] or methods for hot spot crime detection [6] to a fuzzy extended BPMN (business process model notation) for modeling crime analysis processes [7] by an analysis of crime measurement and statistics with the national incident-based reporting system (NIBRS) [8] to name only a few.

Other pieces of research emphasize on the role of data and its importance to be well organized. In [9] it is reported that forensic case data is poorly integrated into crime investigation and crime analysis. They advocate the use of a framework for putting in advance forensic case data, in order to go beyond the single production of this latter data for court evidence. In [10], we can see how companies were able to answer more quickly to business issues when using data warehousing. [11] present an overview of data mining tools, and emphasize on the need of an appropriate data warehouse. A general framework for crime data mining has been briefly approached [12]. An analytical methodology is proposed by [13] to apply data mining focused on a criminal detection perspective.

However, none of these studies are explicitly combining computational forensics with methods to structure data in an analysis perspective. For instance, crime analysis is very seldom introduced by specifying an underlying structure forensic data requires. This is the gap we want to fill within this paper.

3 Data Structures and its Systems

In this section, a brief definition and an overview of the covered subjects is presented. As a starting point, crime analysis can be defined as "the systematic study of crime and disorder problems as well as other police-related issues (including socio-demographic, spatial, and temporal factors) to assist the police in criminal apprehension, crime and disorder reduction, crime prevention, and evaluation" [14]. In this paper, we will focus on a specific aspect of that definition, i.e., crime analysis in conjunction with the use of information technologies and computational methods. The main objective is to understand that the data to analyse might be structured in several ways.

3.1 Transactional/Operational Systems

A transactional/operational system, dealing with day-to-day transactions, is usually using data stored in a database supported by a database management system (DBMS). This kind of environment is called online transaction processing. Not really designed for long term analyses or strategic purposes, the data is considered to be structured and normalized (usually at the third normal form or Boyce-Codd normal form) to avoid redundancy and to preserve consistency. Conceptual entity-relationship (E-R) or relational models (contrasted with multidimensional models) are techniques used to draw these systems.

3.2 Data Warehousing Systems

Whereas operational systems support business processes, data warehouse (DW) systems support the evaluation/analysis of the process. A widely accepted definition of a data warehouse is "a subject-oriented, integrated, time variant and non-volatile collection of data used in strategic decision making" [2]. We will contrast this statement by defining that a DW is not directly integrating decision making and is not subject-oriented: these feature-s are transferred to another level, the data marts. So the DW is not directly serving or supporting a DSS, but the data mart is. Actually, the general purpose of such data architecture is to centralize data stemming from several sources into a unique and conformed warehouse. In a nutshell, using a DW environment has proved to supply data for any form of analytical system within the business [15], and therefore the data should be generic and atomic.

3.3 Data Marts and Decision Support Systems

In the previous part, we explained the role of a data warehouse, and emphasized on the fact that a DW is designed to be unique and generic. The conceptual difference with data marts is that they are a subset of a DW, built to answer specific business questions [16]. Whereas a DW is specific to an enterprise, a data mart is specific to a business problem. Therefore, each business issue (i.e., forensic issue) potentially needs an ad hoc data mart.

The structure of the data mart is not necessarily in a normalized form; according to the required structure of the analytical tool, it might be in any form [17]. Multidimensional modeling may produce star, snowflake, or galaxy schemas for an OLAP environment whereas entity relationship modeling may produce normalized, flat, or hybrid schemas for data mining applications [18].

Decision support systems are a set of tools fed by data marts to conduct analysis. Some examples of DSS techniques are OLAP cubes or data mining algorithms [19].

Basically, the "data" we are dealing with is polymorphous according to the level of detail and interpretation we give to it: in a simple raw log or in a database, it is considered as *data* (as no metadata describes it and no interpretation is given); when it comes to the DSS and is viewed by analysts, the *data* turns into *information* because a context is given to it; eventually, if we know how to use this *information* into actionable business rules, then we consider some useful *knowledge* may be extracted.

4 Proposed Methodology

The essence of the methodology we propose here pertains to the art of data warehousing design; however, the specificities resulting from the application to the computational forensics domain are taken into account and have a considerable impact on the proposed artifact. It is not intended to explicitly define technical details or implementation issues in this paper, however, some concepts will be illustrated in the next section (Sect. 5).

Conceptually, the methodology is a recipe helping in switching from a data acquisition environment to a data delivery environment (i.e., a framework helping in converting operational data into business intelligence or knowledge).

As mentioned before, police event logs are assumed to be stored in databases. These databases use as the finest grain of data a criminal incident, meaning each incident/event occurring turns into a new line (or *transaction*) within the database. These incidents have often the same structure among different police logs and therefore can be usually summarized by [14]:

- an incident number (acting as a transaction identifier),

- the date of the report,
- the location of the crime,
- the date of the incident,
- the method of the crime (a.k.a. *modus operandi*),
- and sometimes a description.

4.1 The 5-Steps Iterative Process

The 5-steps iterative process we propose uses an operational data structure as a starting point and ends by the creation of data marts. Based on a process used to design a data warehouse to represent the enterprise perspective [15] according to the corporate information factory (CIF), it has been simplified and adapted to a shorter iterative cycle. This process is iterative and therefore the cycle has to be performed several times to converge towards the most appropriate solution. The resulting required steps we identified are depicted in Fig. 1.

The first task is to identify the data needed to conduct crime analyses. As part of an iterative process, it is not sought to set exact boundaries from the first time. A set of business questions needs to be raised by crime analysts. Next, we must identify the data sources required to be able to answer to these questions. In our case, the only source might be the logs or a subset of the logs. A clear idea of the main elements or topics to be included in the DW is important.

A subject area model depicting the main entities and their links is the principal output of this step. This model serves as a reference to go further and as a communication basis among stakeholders.

Then, the second step refers to a conceptual data model. This detailed model, representing at a more specific level of detail the relationships and the attributes of the subject area model, is helpful to understand the business. This model is called the *business data model* and is generally using the entity-relationship (E-R) modeling technique. If information stemming from logs is already stored in a database, then the structure is already defined through the system and the business data model will look the same except technical details. Otherwise, a model has to be designed to lay the ground for the next steps. The main advantage is that this model helps people envision how these different parts fit together. This model is not meant to be directly used (neither in a transactional nor an analytical environment), it is purely an intermediate step.

The third task consists in designing the data warehouse. Designing a DW is the most challenging part of this methodology, and may be the subject of an entire book. Many conceptual choices have to be made, and these decisions are mainly part of the information systems domain knowledge. Nevertheless, as we are dealing with a specific case —an ad hoc DW about police reports for the purpose of crime analysis—, some “short cuts” can be taken and consequently the level of abstraction decreases.

The fourth step entails the creation of data marts. Whereas the purpose of a DW is to be generic and normalized, data marts are shaped by requirements and have to be suitable for the DSS, which often implies a multidimensional modeling approach [2]. Actually, a data mart is specific to a subject, to a business problem, and to an analytical tool. The recommended methodology is to design a data mart for each of the possible combinations. Data marts can be viewed as replicated subsets of a data warehouse, and the data within is not meant to be updated (the only atomic and consistent *source* being the DW).

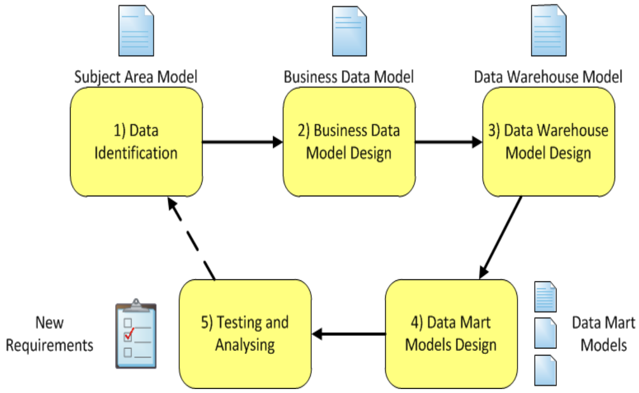


Fig. 1. 5-Steps Iterative process of the proposed methodology

Table 1. Key Success Factors

Key Success Factor	**Steps**	**Effects**	**Based on**
Design the DW with a generic approach	1,2,3	Reduces the development time for a marginal DSS, decouples DW from DSS and technologies	[15]
Manage the data quality upstream	1,2,3,4	Avoid redundant work, preserves data consistency	[2]
Guide conceptual choices with forensic answers	1,2,3,4	Business driven approach, more realistic	[16]
Use an iterative approach	All	Lighter development cycle, easier to integrate new requirements	[15], [2]

The last step is all about implementing, testing, and analyzing what is missing for the next process iteration.

Concerning the implementation, it is mainly about extracting, transforming and loading (ETL) data from a system to another. The ETL technique will not be presented here, but we would like to underline that one should not underestimate (in matter of time) this step.

Testing the system means checking the integrity of the data (Is the data of the new system consistent and coherent with the data of the source system?), checking the interoperability with the DSS, and measuring its effectiveness (i.e., the capacity of the system to answer to business questions mentioned by analysts according to the time and the resources needed). From these tasks, considering the system is never perfect, should emerge new requirements, being future inputs of the next iterations of the process.

We want to emphasize that users of such systems are crime analysts first and foremost, and technician second. Once the ETL project is defined, pushing a single button should be enough to activate the necessary mechanism to update the data flow through all the steps.

4.2 Key Success Factors

Key success factors are a list of requirements that need to be fulfilled to guarantee a high success rate. However, there is no absolute guaranty that whether the key success factors are respected the methodology will be systematically successful.

Each of the specified factors (detailed in Table 1) includes the concerned steps of the process, the positive potential effects on the methodology and their theoretical groundings. They have to be seen as general guidelines to follow during the implementation of the process, or as meta-information about the process.

5 Proof of Concept

The main purpose of this section is twofold: (1) to demonstrate the capability of the proposed methodology to enhance crime analyses through the use of decision support systems and to show the benefits (i.e., reducing marginal costs/time of building specific data marts or ad hoc environments for each specific crime analysis) of building a DW as an intermediary layer ; and (2) to illustrate the methodology through a case-study with data stemming from a real environment.

As a starting point, we will describe the data used in the remainder of this section. In partnership with the Police de sûreté du Canton de Vaud (Switzerland), we gathered a subset of data stemming from their police event logs. Mainly focusing on burglaries and being event centric, this data has several purposes within the organization: keeping a trace of all events for court evidence, serving as a directory for simple exploration, feeding an operational tool used by crime analysts in order to support police operations and to better understand crime series, etc. The input we used to anchor the first step of the proposed methodology is composed of several files encoded in the comma separated values (CSV) format. This is generally the effect of exporting data from another existing tool, such as the one used to store the logs.

This data, because gathered in a police reporting context, is not suitable for advanced analyses. For example, the date of the event is not detailed (the day of the week is not directly given, it is not stated if the date is a public holiday, etc.) and is not accurate (the moment of an event is not a unique date: it is an interval given by the victim —usually corresponding to the interval from when the victim leaves one's house to when he comes back). Moreover, the time of the event has no context, in the sense that neither the daylight saving time nor the time zone is mentioned. Another example are the gps coordinates of the event: the format of the location is using swiss cartesian coordinates (a specific X-Y-Z system). Depending on the import function of the geographic information system (GIS), the data might need to be converted. More generally, these problems can be summarized by this statement: *there is no metadata giving a context to interpret/analyze these events with computational methods.*

Another fact to notice is that each event may have several modus operandi (and vice-versa), therefore a many to many relationship is required. This kind of relationships (usually occurring for *categories* in general) may be a problem when it comes to modeling the DW. If the DW were not generic, then we would denormalize these categories into an array of fixed sized (i.e., if the maximal cardinality of an event is of N categories, then an array of N categories will be denormalized in the event table). This choice is not appropriate because it will limit the use of other potential useful DSSs (certain requiring normalized data).

According to the first step of the methodology, in order to identify the data required for crime analysis, some business questions have to be raised by the end users conducting an analysis in the future DSS. These questions might be essential to delineate the scope and to emphasize the added value of a crime analysis framework. Here are a few examples of questions that could be considered by crime analysts focusing on burglary issues: "Can we group some events together by specific burglary phenomena in order to better understand them? What kind of crime methods occur often together? Are some crime methods following a predictable trend? What is the seasonal trend of burglaries using a window to enter into a house as modus operandi? How many events during summer imply many offenders and last more than one hour?" We may already assume that these questions require data mining techniques and perhaps an OLAP system, due to the inherent difficulty to provide answers using simple SQL requests. A subject area diagram describing the main entities may be drawn from the these questions. Entities such as event, modus operandi, and location of the crime are obviously needed.

Then, the second step of the process is to design the business data model for describing how these entities are constituted and how they interact. This business data model, normalized and using a relational modeling technique, must handle all needed attributes for the analysis. For example, the location of a crime will be defined in a way to handle all the attributes present in the log, but in a normalized way (e.g., if the policeman had reported the modus operandi of the crime as free text, a normalized entity regrouping these values has to be included in the model).

To illustrate some issues of the DW design, let us consider a date describing the moment an event took place. In the previous steps, the structure of the data was probably a single field stating the day, the month and the year with a separator (e.g., *03-06-11*), which was totally sufficient. Whereas in the DW, another structure is needed according to our business questions: if we want to conduct an analysis based on the seasons or to forecast a trend related to weekdays or public holidays, a detailed *calendar* has to be implemented to make the navigation across temporal aspect easier. This can be translated by the creation of entities including these metadata (see Fig. 2: entities *Dates, Days, Months and Years* have been created to represent the concept of a calendar, the derived field *duration_h* has been added to easily retrieve the duration of an event, etc.). The creation of metadata can be leveraged in the same way to consider other aspects such as the daylight time saving, holidays, location, or any other hierarchy.

The fourth step is about designing data marts. According to the numerous DSS requirements, their respective ad hoc data marts will be designed. To analyze our data, we decide to use the following DSSs: one for a data mining system and another for an OLAP system. While the mining mart needs a flat structure (a single denormalized table) using an E-R model, the OLAP mart needs a star schema (modeled with the multidimensional technique).

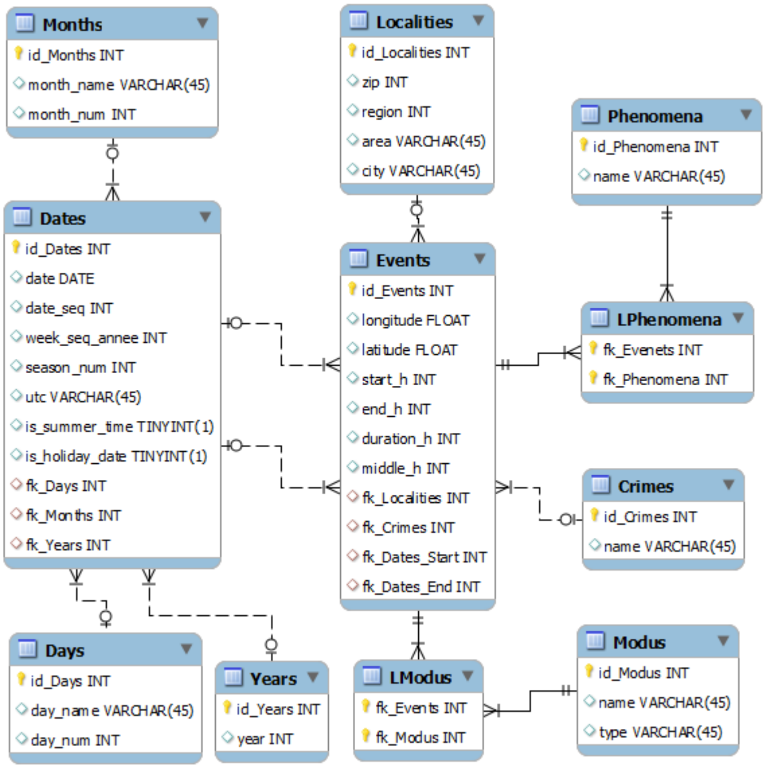
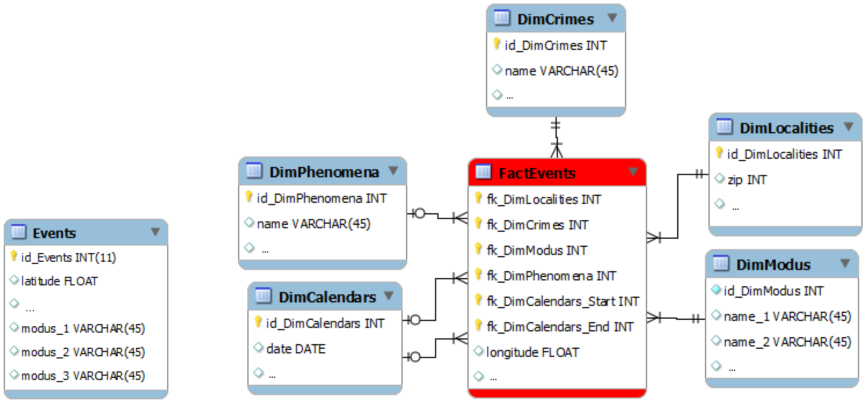


Fig. 2. A data warehouse diagram representing all required entities for burglaries analyses in our case study.

The mining mart (see Fig. 3) is designed in accordance to the specifications of the data mining algorithms included in the data mining software chosen (*Weka*



(a) (b)

Fig. 3. Excerpts of data mart diagrams resulting from the latter DW: (a) a data mart for a data mining DSS (Weka) and (b) an OLAP mart for the fact *Events*.

in our case). As most of the work had been done during the design of the DW, only little adaptation is needed: one of the most difficult task is to denormalize the N-N relationship between the table *Events* and the table *Modus*. The same holds for the OLAP mart: the calendar is already designed and only a few changes (such as denormalization) are required.

The last step entails the implementation and the testing of the framework. Concerning the implementation, it is mainly about using extraction, transformation and loading (ETL) techniques to move gradually the data from the raw logs to the DW and finally to the data marts supporting the DSSs. Within the ETL part, unifying and cleaning the data is very important to guarantee that the integrity and the conformity of the data is ensured all along the processes (e.g., the event *evt306-fis* which occurred in Lausanne at 9 pm on Friday 03-09-11 described in the log needs to be recognizable such throughout all the steps). Lastly, the testing part has to be done by crime analysts. In our case, some data mining techniques were used to answer a part of the preceding questions. These tests have defined new needs as well as new business questions. As such they fulfill the requirements of an iterative process.

All the software used within this part, is open-source software. The DBMS is based on a MySQL server (from Oracle), the ETL part is using Spoon (being a subpart of Kettle PDI from Pentaho), the data mining DSS is Weka (from Pentaho) and the diagrams were created by MySQL Workbench (from Oracle).

6 Evaluation

In this part, we aim to evaluate the usefulness of the proposed methodology. To do so, we suggest to use the following criterion: what is the difference between a scenario when analysis is conducted without implementing an advanced structure (i.e., analyses conducted from raw logs) and a scenario when the analysis is conducted with the use of our artifact.

To conduct the evaluation, we will consider this basic scenario: “A crime analyst for burglaries wants to understand the significance of a day to be a public holiday (for the year 2011)”. Hence he probably needs to calculate (a) the probability that a burglary was committed during a public holiday and (b) the probability that a burglary was committed when that day is not a public holiday.

The latter calculation is not difficult, and only three measures are needed: the total number of burglaries (for 2008), the number of burglaries which happened during public holidays, and the number of burglaries which happened during working days. Considering finding these numbers with data stemming from logs (i.e., without the use of the framework), a few steps will be needed (as extracting the day of the date and looking up the date in a calendar including public holidays). Whereas with the framework implemented, the latter work had been done already and can be leveraged in all further similar analyses: we just need to select the events by filtering them according to the column *is-holiday-date* and to count them.

Of course, in both cases we may be able to reach the objective, but as the number of business questions increases and as the computations become more complex, the first scenario will be de facto inadequate and will be too laborious compared to the simplicity of the questions the crime analyst may raise (low efficiency). The second scenario will have facilitated the task of the end user by delegating these technical tasks to another person (the person designing and implementing the data warehouse) and by avoiding redundant work. Another important fact has to be noticed namely that the risk to introduce errors into an analysis process is reduced as most of the work is zoned directly by the DW, always following the same procedure.

7 Conclusions

The main purpose of this paper is to propose an artifact describing how to lay the groundwork for crime analysis by specifying a way to manage forensic data.

By using an intermediate layer to store the data in a structured way (i.e., a data warehouse), we illustrated how crime analysts can avoid redundant work, reduce the risk to introduce cognitive biases, move from a “stovepipe” view to a transversal view helping in inter units communication and therefore converging towards an intelligence-led policing approach, and decrease the cost and the time of implementing marginal DSSs.

We also evaluated the usefulness and the applicability of the proposed methodology by comparing different scenaria. The methodology was successfully implemented and the main finding was that it made it much easier and less tedious to execute the needed steps to conduct the data analyses in order to answer business questions.

Being a first step towards a wider crime analysis framework by structuring forensic data serving as a basis for computational methods, we obviously encourage additional work or discussions.

Acknowledgements. The authors would like to thank the Police de sûreté du Canton de Vaud and its chief Alexandre Girod for their cooperation and the data provided.

References

1. Hess, K., Orthmann, C., Cho, H.: Police Operations: Theory and Practice. Cengage Learning (2010)
2. Inmon, W.: Building the data warehouse. Wiley (2002)
3. Kimball, R.: A dimensional modeling manifesto. DBMS and Internet Systems (1997)
4. Franke, K., Srihari, S.: Computational forensics: An overview. IWCF 2008 **5158** (2008) 1–10 10.1007/978-3-540-85303-9-1.
5. Maltoni, D., Maio, D., Jain, A., Prabhakar, S.: Handbook of Fingerprint Recognition. Springer professional computing. Springer (2009)
6. Grubesic, T.: On the application of fuzzy clustering for crime hot spot detection. Journal Of Quantitative Criminology **22**(1) (MAR 2006) 77–105
7. Stoffel, K., Cotofrei, P.: Fuzzy extended bpmn for modelling crime analysis processes. SIMPDA 2011 (2011)
8. Maxfield, M.G.: The national incident-based reporting system: Research and policy applications. Journal of Quantitative Criminology **15** (1999) 119–149 10.1023/A:1007518620521.
9. Ribaux, O., Walsh, S.J., Margot, P.: The contribution of forensic science to crime analysis and investigation: forensic intelligence. Forensic Sci Int **156**(2-3) (January 2005) 171–181
10. Ferguson, N.: Data warehousing. International Review of Law, Computers & Technology **11**(2) (1997) 243–250
11. Mikut, R., Reischl, M.: Data mining tools. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **1**(5) (2011) 431–443
12. Chen, H., Chung, W., Xu, J., Wang, G., Qin, Y., Chau, M.: Crime data mining: A general framework and some examples. Computer **37**(4) (APR 2004) 50+
13. Westphal, C.: Data Mining for Intelligence, Fraud & Criminal Detection: Advanced Analytics & Information Sharing Technologies. Taylor & Francis (2008)
14. Boba, R.: Crime Analysis with Crime Mapping. Thousand Oaks, CA: Sage (2012)
15. Imhoff, C., Galemmo, N., Geiger, J.: Mastering data warehouse design: relational and dimensional techniques. Timely, practical, reliable. Wiley Pub. (2003)
16. Kimball, R.: The Data Warehouse ETL Toolkit. Wiley, New York (2004)
17. Jukic, N.: Modeling strategies and alternatives for data warehousing projects. Communications of the ACM **49**(4) (2006) 83 – 88
18. Moody, D.L., Kortink, M.A.: From enterprise models to dimensional models: A methodology for data warehouse and data mart design. Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW'2000) (2000)
19. Bao, Y., Zhang, L.: Decision support system based on data warehouse. World Academy of Science, engineering and Technology **71** (2010)