12 Combining and analyzing judicial databases

Susan W. van den Braak

Research and Documentation Centre, Ministry of Security and Justice, The Netherlands

Sunil Choenni

Research and Documentation Centre, Ministry of Security and Justice, The Netherlands

Sicco Verwer

Department of Computer Science, Katholieke Universiteit Leuven, Belgium

Research and Documentation Centre, Ministry of Security and Justice, The Netherlands

Abstract To monitor crime and law enforcement, databases of several organizations, covering different parts of the criminal justice system, have to be integrated. Combined data from different organizations may then be analyzed, for instance, to investigate how specific groups of suspects move through the system. Such insight is useful for several reasons, for example, to define an effective and coherent safety policy. To integrate or relate judicial data two approaches are currently employed: a data warehouse and a dataspace approach. The former is useful for applications that require combined data on an individual level. The latter is suitable for data with a higher level of aggregation . However, developing applications that exploit combined judicial data is not without risk. One important issue while handling such data is the protection of the privacy of individuals. Therefore, several precautions have to be taken in the data integration process: use aggregate data, follow the Dutch Personal Data Protection Act, and filter out privacy-sensitive results. Another issue is that judicial data is essentially different from data in exact or technical sciences. Therefore, data mining should be used with caution, in particular to avoid incorrect conclusions and to prevent discrimination and stigmatization of certain groups of individuals.

12.1 Introduction

In the Netherlands, many organizations work together to ensure the enforcement of law and public safety of people. Each of these organizations covers a specific area in the field of crime and law enforcement. For instance, the police focus on reported crime and hand over suspects to the prosecution service. The Public Prosecution Service then decides whether to prosecute or drop a case. The court can either convict or acquit a suspect, and may impose sanctions such as imprisonment. When a sentence is pronounced by court, the execution of sanctions follows. Together, these steps from reporting a crime to the execution of sanctions are referred to as the *criminal law chain*. This chain thus consists of four phases: investigation, prosecution, trial, and execution. The police, prosecution service, courts, and the organizations that execute sanctions collaborate in this chain. Each organization registers relevant data, for instance, about the case and the suspect, in its own data source.

To define an effective and coherent safety policy, policymakers have a practical need for statistical insights into the registered data [3, 4, 10, 11]. Such insights can only be gained by relating and integrating the data in a coherent manner. For instance, when data from the different co-operating organizations are integrated and compared, it can be investigated how specific groups of suspects or criminal proceedings move through the chain. Also, by monitoring flows within or between organizations in the chain, policymakers are able to observe whether there are potential problems in a certain part of the chain.

In the Netherlands, combined crime data have already been distributed offline in book form [9, 12] for several years. Although the statistical yearbook is very useful in its current form, there is a growing demand for online data from different groups of users. Therefore, several attempts have been made to develop tools or information systems that collect and process safety-related data from relevant sources and present them in an integrated and uniform way to the users [4, 5]. Such tools obviously have potential, but should be developed with care, as they may also provoke undesired effects. One of the core issues here is the protection of the privacy of individuals. Data should be processed, collected, and combined in a way that respects privacy law and regulations. In general, privacy has a subjective nature and is open to different interpretations depending on its context. In the context of public safety, privacy is primarily focused on the non-disclosure of the identity of individuals. A related issue is the discrimination of groups of individuals, that is, the prejudicial treatment of individuals because they belong to a certain group. To minimize the risk of discrimination or stigmatization, combined crime data should be presented and analyzed with caution.

In this chapter, it will be described how judicial data can be collected, combined, and analyzed such that the privacy of individuals in society is not violated. Although IT offers great potentials to automate the collection and combination of data, still a significant manual effort is required to ensure data quality and to avoid undesired effects. A dataspace approach is presented that allows one to efficiently

relate and exploit data from different sources. It is demonstrated how the information needs of judicial policymakers can be fulfilled using this approach. To analyze data , besides traditional statistical techniques, contemporary techniques such as data mining can be employed. However, it is argued that the straightforward application of such data analysis techniques on judicial data is not without risk. The main reason for this is that the nature of these data is essentially different from the nature of data in exact or technical sciences.

The remainder of this chapter is organized as follows. Section 12.2 is devoted to a brief description of the major databases in the Dutch criminal justice system. In Section 12.3, it is described how data from these databases are currently collected and combined. In this section two approaches to combining judicial data are presented: a data warehouse approach and a dataspace approach. Section 12.4 elaborates upon the problems that may occur in the data integration process due to the nature of crime data. Subsequently, in Section 12.5, potential privacy-related risks of integrating and presenting crime data are described and methods that enforce privacy law and regulations are listed. Section 12.6 explains how combined crime data may be analyzed and which risks are entailed by applying data analysis techniques to them. Finally, Section 12.7 concludes this chapter.

12.2 Databases in the Dutch criminal justice system

The Dutch criminal law chain consists of various organizations, each of which operates relatively autonomously and independently. This means that each organization registers data in its own way and in its own operational system. The most important databases of these organizations are described below.

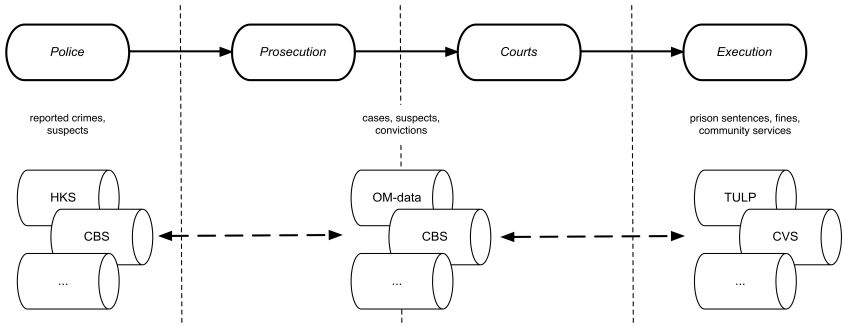
The national database of the Dutch police is called the Identification Service System (*Herkenningsdienstsysteem*, HKS). HKS contains information about crime reports and suspects. Additional information is provided by Statistics Netherlands (*Centraal Bureau voor de Statistiek* (CBS), a national institute that provides statistical information). The CBS Police Statistics also contains information about crime reports and suspects.

Information about judicial cases is stored in the registration system of the Public Prosecution Service (*Openbaar Ministerie* (OM); the information system is called OM-data) and in CBS Court Statistics. Note that these databases register information on a case level, while the police databases register crime reports. As more than one crime report may be handled in a single case, numbers obtained from these sources should be combined and compared with care.

Sanctions are registered by the different organizations involved in the execution of sanctions. Among these organizations are the Custodial Institutions Agency (*Dienst Justitiële Inrichtingen*, DJI), the Child Care and Protection Board (*Raad voor de Kinderbescherming*, RvdK), after-care and resettlement organizations, and the Central Fine Collection Agency (*Centraal Justitieel Incassobureau*,

CJIB). All of them have their own information system. For instance, DJI uses the Execution of Sanctions Program (*Tenuitvoerleggingprogramma,* TULP) to register the duration of detentions, while the Dutch Probation and After-care Organization (*Reclassering Nederland*, RN) uses a Client Follow System (*Cliënt Volg Systeem*, CVS). A schematic overview of the databases maintained in the Dutch criminal law chain is given in Figure 12.1.

Fig. 12.1 Databases in the Dutch criminal law chain



12.3 Collecting and combining judicial data

The database systems described above have in common that they contain data about individuals and their actions. Each of these individuals came into contact with the police or the criminal justice system. Each organization involved registers privacy sensitive attributes such as name, address, and identifying numbers, but also other data regarding a person. The different databases thus store the same, or similar, information and, therefore, they are partially redundant. Consider, for instance, the database systems of the police and the prosecution service that both contain data about people who are suspected of a crime and about the crimes they supposedly committed. If someone is suspected of a murder, both the database of the police and the database of the prosecution service will contain information regarding the date and place where the body is found and (if known) the date and place where the murder is committed. Other information, however, is registered in only one of the databases. For example, the police database contains detailed information about the suspect (such as whether he is first offender or not), while the database of the prosecution service contains detailed information about the case (such as the sections of the law that were violated). This is due to the fact that the police and the organizations involved in the execution of sanctions are individual-oriented, while the prosecution service and the courts are case-oriented.

To perform their tasks in an effective and efficient manner, the police and justice organizations not only require access to their own data; they also have a great demand for a combination of relevant data from other organizations in the criminal law chain. Organizations with operational tasks (such as the police and the prosecution service) require combined data at an individual level, while organizations with strategic or knowledge transfer tasks (such as policymakers and criminologists) require data at a higher aggregation level.

As an example of the former, assume that a Public Prosecutor wants to prosecute a suspect for his actions, then all relevant data (from different sources) that pertain to this suspect should be collected and combined. In this way, the prosecutor can build the strongest case possible, because all information about the suspect is gathered; including evidence for the fact that he is a criminal. Thus, integrating data on an individual level involves data reconciliation, that is, the identification of data in different sources that refer to the same entity. In Subsection 12.3.1 it is shown how this can be established using a data warehouse approach.

Alternatively, policymakers need combined data at an aggregate level. They want to gain insight into the criminal law system as a whole, for instance, to answer the question of which kinds of suspects are brought to court and which kinds of cases are settled out of court. Such insight may be relevant to them in order to be able to define an effective policy. To provide them with this information, the different databases also have to be combined, but not on an individual level, in this case a higher level view is more useful as will be shown in Subsection 12.3.2. In this subsection, a dataspace approach will be presented in which aggregate data are related.

12.3.1 A data warehouse approach to combining judicial data

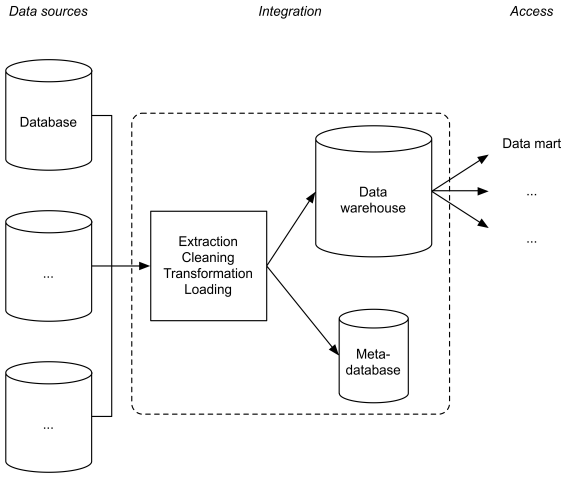
A data warehouse [13] is a central repository of data collected from different sources. These data are stored and structured in such a way that querying and reporting are facilitated. It provides a uniform data model for all data regardless of their source. Generally, a data warehouse consists of three layers that provide storage of the original data sources, integration, and access (see Figure 12.2). First, the raw data from different databases are extracted. Subsequently, these data are cleaned, transformed, and loaded into the data warehouse. The data warehouse then contains data from different databases that are combined and ordered. In addition, information about the data in the data warehouse is stored in a metadatabase. This database contains information about the sources and history of the data. Finally, as a last step, data from the data warehouse are provided to endusers through data marts. The key step in developing a data warehouse is data integration; therefore, data reconciliation is of crucial importance [3].

The main problem with combining and integrating crime data is that only a few organizations with an operational task are allowed (by law) to combine data on the

basis of unique identifiers or a set of privacy sensitive attributes. For this reason, before making crime data available for research purposes, privacy sensitive attributes are stripped from the databases. Hence, for data reconciliation other overlapping information in the to-be-combined databases has to be exploited. This can either be information about the database schemata or information that is extracted from the database content. Furthermore, in order to be able to utilize this information, domain knowledge from experts is needed.

In practice, to establish whether two records from different database system denote the same object, the following general rule of thumb can be applied [3, 6]: the larger the number of common attributes with the same values for two records from two different systems, the higher the chance that the records relate to the same object in reality. Note that this rule of thumb requires that the selectivity factors of the common attributes are small [2].

Fig. 12.2 An overview of the data warehouse approach



Example: An offender-oriented data warehouse

An example of a data warehouse in the Dutch criminal law chain is the offender-oriented data warehouse [3, 6]. In this data warehouse data from different judicial databases (HKS and OM-data) were integrated by applying the rule of thumb explained above. Additionally, the data was structured and combined in such a way that all data relate to individuals.

In the data warehouse the 'intersection' of the to-be-combined databases is exploited. HKS stores information on three entities: suspects, the official reports about them, and the offences of which they are suspected. OM-data also records information about suspects and offences. Additionally, it registers case-related information. Thus, the databases are integrated based on the attributes concerning the two common entities, that is, suspects and offences. To do so, the databases are compared to each other and the probability that two records relate to the

same person based on common attributes is determined. While doing so, domain knowledge is considered, for instance, the fact that an offence is usually reported to the police on the same day as it is committed. The date of an official report in HKS is, therefore, considered to be the same as the date of an offence in OM-data.

As an example, assume that HKS contains a record relating to a person who resides in Amsterdam and in respect to whom an official report has been filed on September 1, 2010. Additionally, assume that OM-data contains a record of a person residing in Amsterdam who committed an offence on September 1, 2010. Then, it is likely that both records concern the same person. Alternatively, if HKS would show that the date of the official report is unknown because it is not entered correctly, this probability would be considerably lower. Note that in the example the residence of the suspect is not very selective and that it is surely possible that multiple residents of Amsterdam commit an offence on the same day. If this is the case, additional or different attributes are needed to ensure that the records are combined properly. After all, if more attributes overlap, the probability that the two records denote the same person increases.

The data in a data warehouse can be made available through data marts. An example that is based on the offender-oriented data warehouse is the Drug Crime Data Mart [6, 14] which consists of a selection of the data concerning drug-related crime. This data mart can be used for analysis and reporting purposes, such as National Drug Monitor publications.

12.3.2 A dataspace approach to combining judicial data

In a dataspace approach [7], also three layers are distinguished (see Figure 12.3): a dataspace layer, a space manager layer, and an interface layer. The dataspace layer contains a set of (cleaned) databases that are complement to each other and may be related. Although these databases are related there is no need for data reconciliation. Alternatively, the relations that exist between the databases are stored in a relationship manager in the space manager layer. This layer maintains data quality (the plausibility and consistency of the data) by providing rules to which the data must adhere. For this purpose the relationship manager contains different types of rules:

1. Rules to handle similar data coming from different sources.

2. Rules to deal with missing data.

3. Rules to allow for incomplete or tentative data.

4. Rules to record semantic changes in attributes.

5. Rules to filter out results that should not be shown to the user.

6. Rules to determine whether large deviations exist between past and future data or between values from the same or different databases.

All in all, a combined set of rules in the relationship manager serves to complete incomplete data sets, determine whether they are acceptable, and warn users when they are less reliable. The relationship manager also serves to minimize the chances of misinterpreting data. To do so, this layer maintains the relations between attributes in the different databases and keeps track of changes in the meaning of these attributes. Based on this history of changes, the space manager may decide to reorganize or convert a database, in particular if the semantics of major attributes changed considerably over the years.

Another task of the space manager (besides providing a relationship manager) is to serve as a communicator between the database and the user interface. As users define questions at the interface level, the query scheduler of the space manager decides which databases to query in order to answer each question. Once the answer is retrieved from the databases, it is displayed to the user through the interface. Before presenting the output, rules (of the fifth type) may be applied to check whether it can be shown to the user. This can, for instance, be used to preserve the privacy of individuals as will be explained in Section 12.5.

The interface layer not only contains mashups of crime data, but also provides features that are more tailored to the needs of specific users. An example is a publishing on demand module which provides users with the possibility to generate and print reports. Such a module should insert automatically updated tables and graphs in a preformatted report. A feature of this kind is particularly useful for standardized research reports (see, for instance, [14]).

Fig. 12.3 An overview of the dataspace approach

Example: A public safety monitor

An example of the dataspace approach at work in the Dutch justice system is the public safety monitor [5]. This monitor shows the development of the input and output of cases in the different organizations in the criminal law chain. In addition, it provides a comparison of the actual data with forecasts. The data in the monitor's dataspace is extracted and aggregated from the various databases in the field of crime and law enforcement described in Section 12.2.

With respect to the relationship manager of the monitor, it should be clear that the data in the monitor are closely related. The output of one organization in the criminal law chain often serves as the possible input of another organization. For instance, the input of cases into the prosecutorial level largely depends on the number of suspects handed over by the police. Therefore, a plausible rule in the manager would be: number of suspects handled by the police ≥ number of case handled by the prosecution service; meaning that the police usually do not send all cases to the prosecution service. Using such rules the plausibility and consistency of the data is maintained. Similar rules can be formulated in order to handle variables coming from different sources. Take, for instance, the number of community services agreed with the Public Prosecutor. This information comes from two organizations in the chain: the prosecution service and the organization responsible for executing sentences. As a rule, the number of community services registered by the executing organization is lower than the number of community services registered by the prosecution service, as suspects may die or 'disappear' before the sentence is executed. Such rules are typically based on historical data and domain knowledge and can be extended with error values to allow for incomplete or tentative data (see [5]).

The primary goal of the monitor is to alert users when there are large differences between input and output, or between the actual input or output and the forecasted input or output. In this way, policymakers are able to indentify potential capacity problems at an early stage. Therefore, rules are added to the relationship manager that detect large deviations. Based on these rules, three types of alerts are provided to the users:

1. *large deviations in the proportion of organization X's output to its own input;*

2. *large deviations in the proportion of organization X's input to organization Y's output;*

3. *large forecasting errors.*

The user interface presents the user with an overview of the input and output data and the corresponding alerts in either table or graph format, depending on the user's preference. In this way a quick overview of the irregularities in the data is provided. In these views, the user is able to zoom in on specific parts of the criminal law chain by selecting a subset of data categories. More specifically, the user can subdivide the data into various categories including age, gender, and region of the suspect, and type of crime committed by the suspect. Thus, the user can, for instance, choose to only show the input and output of male suspects who are older than 18 years or the input and output in a specific region. Additionally, the monitor periodically produces written reports through a printing on demand module as described above.

In this section it was shown how data from various judicial databases may be combined and integrated using two different approaches: a data warehouse and a dataspace. In the first approach, data is linked explicitly on an individual level. In the second approach, more dynamic relations or rules are established to link data and maintain data quality. Thus, a dataspace differs from a data warehouse in the sense that a common data model is not required and that there is no need to link data based on unique identifiers. As a result, in a dataspace approach not only micro data but also aggregate data may be used. This does not alter the fact that a dataspace layer may contain a data warehouse as a data source.

The worked-out examples from the Dutch criminal justice chain illustrate that data integration can be executed in a variety of ways. For instance, depending on the needs of the users or the availability of the data, parts in this process may have to be altered. In the next section it is shown how potential problems associated with linking (crime) data affect the data integration process and the choices made in it.

12.4 Challenges in combining judicial data

The main problem with data integration is that, although it can be automated for a large part, a significant amount of manual effort is still required. The main reason for this is the nature of crime data: redundancy, inconsistencies, dependencies, and semantic changes are not uncommon. In the remainder of this section, these potential problems and their consequences for the data integration process are described in detail.

Taking care of quantitative and qualitative dependencies
One of the problems with reconciling judicial data is the fact that quantitative dependencies between different data sources exist. For example, the date on which a crime is reported is usually the same as the date on which the crime is committed or the output of the police is usually greater than the input into the prosecutorial level. Though some of this knowledge may be exploited for data reconciliation (to compare records from different sources), it requires manual effort and the participation of domain experts.

Qualitative dependencies also exist within databases. For instance, it is generally assumed that the value of a certain attribute does not change dramatically in a few years. Therefore, it is recommended to compare the value of an attribute in a certain year to its value in preceding years in order to detect large deviations.

Thus, when data from different sources are combined, both quantitative and qualitative dependencies have to be managed in order to avoid unreliable data. In a data warehouse this has to be done manually by domain experts. In a dataspace approach it can be automated fully using dynamic rules that check the reliability of the data and detect deviations.

Managing semantic dependencies
Besides quantitative and qualitative dependencies, also semantic dependencies exist in and between judicial databases. These arise because different organizations in the criminal law chain store data about the same events, but often label or classify these data differently. For example, in case of a robbery a victim may classify it as a violent crime, while the police may classify it as a crime against property. Additionally, for a single case in court that contains several offences, the severest case is taken as the classification criterion. As a result, less severe offences ‘disappear’ in the data reported by the court.

It is important that existing semantic dependencies between attributes (if any) are preserved while integrating data. Therefore, in a data warehouse domain experts need to keep track of semantic dependencies. In a dataspace these may be captured in rules.

Resolving inconsistencies
The different judicial databases have overlapping or redundant attributes. Redundancy may introduce inconsistencies that have to be detected and solved manually based on domain expertise. Take for example the nationality of a suspect that is

recorded by different organizations. It is known that, in practice, foreigners tend to provide a wrong nationality when they are not able to show identification papers. As a result, inconsistencies may arise between different databases of different organizations. This can be resolved by utilizing the domain knowledge.

Prior to loading data into a data warehouse, inconsistencies have to be indentified and resolved. This means that all values of overlapping or redundant attributes have to be in agreement with each other. In a dataspace approach inconsistencies can be detected automatically and on the fly using rules that check attributes coming from different sources. .

Handling semantic changes

Data evolve over time as rules and regulations are changing. Therefore, certain values on certain attributes may have gotten a different meaning over time. For instance, due to municipal reorganizations in the Netherlands, names of municipalities and cities have changed, while the old registered names were not always updated. Over time, the meaning of the old names may become unknown. Moreover, in case cities are expanded, their names mean something different before the reorganization than after. If these changes are not recorded, data may be combined improperly or wrong conclusions may be drawn based on them. To keep track of the 'history' of the attributes, semantic changes have to be recorded. In a dataspace this can be done in the relationship manager.

Concluding example

In general, a dataspace approach may be considered to be more efficient and practical than a data warehouse approach, because in the former it is easier to combine data and add new sources, as there is no need for data reconciliation. Additionally, using a dataspace approach dependencies, inconsistencies, and changes can be managed more effectively.

As an illustration, assume that one wants to know how many of the suspects questioned by the police are handed over to the prosecution and how many of them are actually prosecuted. To answer this question, the databases of the police (HKS) and the prosecution (OM-data) have to be integrated. However, OM-data only contains data of cases that are handled by the prosecution. This means that not all individuals in HKS are present in OM-data and, therefore, combining on an individual level, which is needed in a data warehouse approach, is impossible for these individuals. In a dataspace approach, however, aggregate data can be used, so the database may contain the total number of suspects questioned by the police (aggregated from HKS) and the total number of suspects (cases) handled by the prosecution (from OM-data). Then, a comparison can be made between the two totals, and the difference between output and input can be calculated. This task can be performed easily by the public safety monitor (described in Subsection 12.3.2). Thus, for this type of questions, a dataspace is more efficient as the heavy

computational and troublesome task of uniquely linking individuals does not have to be performed.

12.5 Protecting privacy when combining judicial data

Tools or information systems that collect, relate, and present safety-related data, pose a serious privacy threat as the identity of individuals or groups of individuals may be exposed. For instance, assume that in the public safety monitor (see Subsection 12.3.2) the number of sex offenders is presented, and that it is possible to categorize them by age, gender, and city. If there is only one female sex offender in a certain city, then the age of this female is exposed. Depending on the additional information that is shown about her, or the information that can be gathered from alternative sources, it is likely that her full identity is exposed. If this is indeed the case, privacy laws are violated.

In the data integration process several precautions can be taken to respect the privacy of individuals and to minimize the risk of exposing someone's identity. First, a data source that contains crime data should only record attributes that are in line with the Dutch Personal Data Protection Act (PDPA). This act defines a set of sensitive attributes that should be handled with care, namely data on someone's religion or life conviction, ethnic origin, political opinions, health, sexual orientation, and memberships of (trade) unions [15]. Such sensitive attributes should not be stored. Second, aggregate data has a clear advantage over micro data as data on a higher aggregation level does not provide personal information. Therefore, for privacy reasons, it is recommended to use aggregate data instead of micro data when possible. Finally, whenever there is a risk of exposing the identity of an individual to a user of a tool, the result of the user's question or selection should not (or only in part) be shown. For example, if a user wants to view the number of sexual offenders per region, and if there are just two offenders in a certain region, this number should not be presented to the user. After all, in this case there is a reasonable chance that with additional information, the identity of the offenders concerned can be deduced.

When all three precautions are followed, the risk of disclosing personal data and thereby violating the privacy of individuals are minimized.

The preceding sections focused on ways to combine and integrate data from various judicial databases. Combined crime data may help in gaining insight into the criminal law chain and in developing new policies. An even deeper under-

standing of crime and delinquency may be acquired by applying data analysis techniques to such data. In this way, profiles of criminals or offenders may be constructed. In the next section, potentials and challenges of analyzing combined crime data will be described.

12.6 Risks of analyzing judicial data

Statistics may be considered as a standard tool for the analysis of police and justice data. However, as in many organizations, the amount of data collected and stored by the judicial organizations has grown exponentially. In many fields, especially technically oriented fields, data mining [16] has been proven to have an added value over statistics in analyzing large amounts of data [1, 8] (see [1] for a summary of the differences between statistics and data mining). Data mining is the process of searching for statistical relations, or patterns, in large data sets. It is often used to gain a different perspective on the data and to extract useful information from them. Commonly used methods include rule learning (searching for relationships in the data), clustering (discovering groups in the data that are similar), and classification (generalizing known structures to new data). Thus, data mining is able to reveal useful knowledge that is hidden in a large amount of data. Therefore, there is a growing interest in applying data mining techniques to crime data.

However, the straightforward application of statistical techniques, and data mining in particular, may be risky. As has been pointed out in the literature [8], data mining results need to be evaluated by experts to determine whether they hold in the real world. The main reason for this is that data mining is based on induction and, therefore, the results may be true given the data, but not in the real world. For example, assume that all swans in a given databases are white, then it may be induced from the database that all swans are white. However, it is very well possible that only features of white swans are stored in the databases and that the very small group of black swans is neglected. As a result, the induced knowledge with regard to swans does not hold in the real world. Therefore, it is of vital importance to evaluate the truthfulness of data mining results.

For police and justice data, evaluation is even more important and that because of the following reasons. Opposed to findings in exact or technical sciences, findings in social sciences may be subject to change in the course of time. For instance, Newton's laws of motion were true decades ago and do still hold today, while the age-crime distribution in crime science is changing over time. For instance, in 2000 minors were responsible for roughly 17% of the committed crimes (that is, of all interrogated suspects, 17% was between 12 and 17 years old), while in 2007 they were responsible for around 19% of the committed crimes [9].

Another reason to be cautious with data mining results in social sciences is the fact that, since data collection is a time-consuming and difficult process, often legacy databases are used for data mining. Such databases contain large amounts

of data that were collected and stored in the past; sometimes decades ago. As a result, these databases mostly reflect the situation in the past, so mining these databases results in knowledge about the past. Evaluation of such data mining results is important for three reasons:

- It has to be determined whether this knowledge corresponds with the real world of the past; and
- it has to be determined whether the knowledge still holds in the real world of today; and
- it has to be determined whether it is useful to apply the obtained knowledge (for instance, in developing new policies).

As an example, assume that data mining is applied to a database containing data about juveniles and nuisance offences from 1975 to 2005. By doing so, profiles of youngsters who cause annoyance may be found. A hypothetical result may be that young men born in a particular country have a higher probability to cause nuisance. However, it may be the case that this was true in the seventies, but not today, as since then they may have adapted their behavior to the Dutch society and norms. Thus, although the result corresponds to the real world of the past, it does not correspond to the real world of today. It is surely possible that nowadays young men from other countries show nuisance behavior. In this case, the fact that young men in general cause nuisance does hold in today's world, and can be usefully applied, but using the country of origin of these men is dangerous.

Contrary to data mining, the chances that such issues are encountered when applying statistics on crime data are small. Statistics requires carefully formulating hypotheses that are tested on newly collected data. Thus, the data used for standard statistical analyses always reflect the real world as it is today and do not involve the issues relating to legacy data.

Another important issue is that, since data mining tools are developed to find patterns based on any correlation in data, they can find patterns that use personal characteristics of groups of individuals. This may lead to discrimination and stigmatization of these groups. For instance, assume that data mining algorithms are employed on a database of sex offenders that is enriched with demographical and economical data. A likely data mining result may be that unemployed white men are responsible for 80% of the sex offences. There are two problems with such a statement. First, it could lead to stigmatization as the relation to the total population of unemployed white men is not made clear. Second, using it to discover new (unknown) sex offenders leads to discrimination because suddenly all unemployed white men are suspects, while only a few of them are actual sex offenders.

In sum, in this section it was shown that although applying data mining techniques to crime data seems promising, there are some issues regarding the applica-

bility and generalizability of the obtained results. Additionally, data mining may lead to discrimination and stigmatization. Therefore, data mining methods should be used with caution.

12.7 Concluding remarks

In this chapter it was illustrated how data from judicial databases in the Netherlands are currently processed, combined, and analyzed. It was explained how precautions in the data integration process should be taken to better respect privacy law and regulations. When such measures are taken, the risks of exposing the identity of individuals are minimized. Subsequently, it was shown that applying data analysis methods to judicial data is not straightforward and that data mining results should be considered with caution. When these reservations are taken into account and the precautions mentioned are taken, applications that exploit combined crime data and provide statistical overviews are valuable tools for judicial policymakers in developing new and effective policies. An example is the recently developed public safety monitor. This monitor fulfills the information needs of policymakers and advisors and allows them to timely identify potential capacity problems.

12.8 References

1 Choenni S, Bakker R, Blok H, de Laat R (2005) Supporting technologies for knowledge management. In: Baets W (ed) Knowl Manag and Manag Learn: Ext the Horiz of Knowl-Based Manag vol 9, part 2 of Integr Ser in Inf Syst, pp 89-112. doi:10.1007/0-387-25846-9_6

2 Choenni S, Blanken H, Chang T (1993) Index selection in relational databases. Comp and Inf, 1993 Proc ICCI '93 Fifth Int Conf on Inf and Comp 491-496. doi:10.1109/ICCI.1993.315323

3 Choenni S, van Dijk J, Leeuw F (2010) Preserving privacy whilst integrating data: Applied to criminal justice. Info Pol 15(1,2):125-138. doi:10.3233/IP-2010-0202

4 Choenni S, Leertouwer E (2010) Public safety mashups to support policy makers. In: Andersen K, Francesconi E, Grönlund Å, van Engers T (ed) Electron Gov and the Inf Syst Perspect vol 6267 of Lect Notes in Comp Science, pp 234-248. Springer, Berlin. doi:10.1007/978-3-642-15172-9_22

5 Choenni S, Kalidien S, Ariel A, Moolenaar D (2001) A framework to monitor public safety based on a data space approach. Proc of EGOV 2011 In: Janssen M, Macintosh A, Scholl HJ, Tambouris E, Wimmer MA, de Bruijn H, Tan Y-H (ed) Electron Gov and Electron Particip Jt Proc of Ongoing Res and Proj of IFIP EGOV and ePart 2011 vol 37 of Schriftenreihe Inform, pp 196-202. Trauner Verlag, Linz.

6 Choenni S, Meijer R (2011) From police and judicial databases to an offender-oriented data warehouse. Proc of the IADIS Int Conf on e-Soc 98-105

7 Franklin M, Halevy A, Maier D (2005) From databases to dataspaces: a new abstraction for information management. SIGMOD Rec 34(4):27-33. doi:10.1145/1107499.1107502

8 Hand DJ (1998) Data mining: statistics and more?. The Am Stat 52(2):112-118

9 de Heer-de Lange NE, Kalidien S (2010) Criminaliteit en Rechtshandhaving 2009: ontwikkelingen en samenhangen [Crime and Law Enforcement 2009] vol 279 of Onderz en beleid. Boom Juridische uitgevers, The Hague
10 Kalidien S, Choenni S, Meijer R (2009) Towards a tool for monitoring crime and law enforcement. Proc of ECIME 2009 the 3rd Eur Conf on Inf Manag and Eval 239-247
11 Kalidien S, Choenni S, Meijer R (2010) Crime statistics online: potentials and challenges. Proc of the 11th Annu Int Digit Gov Res Conf on Public Adm Online: Chall and Oppor (dg.o '10) 131-137
12 Kalidien S, de Heer-de Lange NE (2011) Criminaliteit en Rechtshandhaving 2010: ontwikkelingen en samenhangen [Crime and Law Enforcement 2010] vol 298 of Onderz en beleid. Boom Juridische uitgevers, The Hague
13 Kimball, R, Ross M (2002) The data warehouse toolkit: the complete guide to dimensional modeling 2nd ed. John Wiley & Sons, Inc, New York
14 Meijer R, van Dijk J, Leertouwer E, Choenni S (2008) A drug crime data mart to support publication on demand. Proc of the 2nd Eur Conf on Inf Manag and Eval 277-286
15 Sauerwein LB, Linnemann JJ (2001) Guidelines for personal data processors: personal data protection act. Ministry of Justice, The Hague
16 Tan P, Steinbach M, Kumar V (2005) Introduction to data mining. Addison Wesley, London