Data Quality Management in the Public Domain: A case study within the Dutch Justice System

Abstract

The need for anonymity preservation within the Justice domain requires the introduction of a Trusted Third Party as an intermediary, while integrating individual databases within its boundaries. After the Trusted Third Party encrypts records, it is no longer possible to perform checks on the data quality and correct data anomalies.
Therefore, this research examines the concepts of Data Quality, Data Integration, Record Linkage and Trusted Third Party and, then, combines these with four expert interviews in order to identify ways to assess and improve Data Quality while linking privacy-sensitive data. Next, the Trusted Data Linkage Framework (TDLF) is presented to aid Data Quality Management while combining citizens' privacy-sensitive data from different organisations. Finally, we evaluate the framework in a case study to demonstrate how the quality of structured judicial data can be managed prior to its encryption, while using multiple databases as sources of data and a different final recipient.

Key words: Data Quality; Record Linkage; Data Integration; Trusted Third Party.

1. Introduction

Knowledge transfer actions like data sharing and electronic collaboration are essential in the public domain, in order to perform their tasks effectively and deliver quality services (van den Braak, Choenni, Meijer and Zuiderwijk, 2012). However, the personal data of citizens must be treated accordingly and the related legal measures (acts of law) must be taken into serious consideration each time that a data sharing process takes place.
Because of the fact that data privacy must be ensured, the data have to be encrypted while trying to integrate them in a unified view. As Otjacques et al. (2007) admit, data privacy is always an issue when such data is examined, aggregated or processed and that is the reason why data privacy and sharing and identification are strongly interdependent subjects.
In this research, we examine the concepts of Data Quality (DQ), Data Integration, Record Linkage and Trusted Third Party (TTP) found in scientific papers. The scientific literature findings are combined with expert interviews, in order to identify ways to assess and improve Data Quality while linking privacy-sensitive data. The Trusted Data Linkage Framework (TDLF) is presented to aid Data Quality Management while combining citizens' privacy-sensitive data from different organizations.

1.1 Problem definition

The civilians expect from the government/organization to perform these actions following certain legal protocols, in order to preserve their privacy and protected from unauthorized access. Nevertheless, inter-organizational data sharing in this specific domain can be difficult since as van den Braak et. al (2012) describe, "*the collaborating public organizations may all have their own rights and obligations with respect to security and privacy*" and "*the civilians expect that public organizations follow rules and procedures carefully in order to protect their privacy*".
Data integration is an example of data sharing. It is considered as "*the problem of combining data residing at individual sources, and providing the user with a unified view of this data*" (Lenzerini, 2002). In order to achieve that, the individual records have to be matched. Record linkage is a very similar concept to this, since it talks about object identification, data cleaning and approximate matching of objects that belong to the same real-world object (Christen and Churches, 2006).
Data sharing functions and processes tend to lead to security issues and disclosure of privacy- sensitive information. A TTP should be utilized as an intermediary between the different sources of data in cases that sensitive data is transferred from one organization to another, in order to perform data integration and sharing without trespassing any laws and regulations over data privacy. As far as the term TTP is concerned, van de Braak et al. (2012) define it in their paper as "*an independent organization that acts as a liaison between two or more collaborating public organizations*".
Strong et al. (1997) define DQ problem as "*any difficulty encountered along one or more quality dimensions that renders data completely or largely unfit for use*". In order to manage the concerning DQ each time there are various methods that aid in that direction. It has to be noted here that there are three main common stages of DQ methods were identified that are described by Batini, Cappiello and Francalanci (2009), namely State reconstruction, DQ assessment and DQ Improvement.
Taking into consideration the latter, we were motivated to set the following research objectives:

- Link specific DQ dimensions to the Judicial domain, while sharing privacy-sensitive data
- Demonstrate the process and flow of the data when records from two individual sources are linked under TTP supervision and the results are given to an independent recipient
- Construct a framework to manage DQ prior to data encryption

1.2 Contribution

A fundamental/first outcome of this study will be the identification of DQ properties/dimensions that affect the data quality of DBs during a data integration process in the Judicial Domain, when trying to link privacy-sensitive data. One of the most important contributions of this research will be also the proposal of a situational DQ management framework that can be used when applying different DQ methods. Different parties in the Dutch Ministry of Justice and Security argue whether the use of a TTP is necessary, so it will be very useful to demonstrate in depth both the activities of the process and its deliverables. In that manner, the contribution and role of the TTP use in the process will be made clear. By implementing the aforementioned DQ management framework in an actual case study within the ministry, data anomalies that produce mismatches during linking the data will be identified.

1.3 Dutch Criminal Justice System

Within the spectrum of this research, the exchange of data within the Dutch Criminal Justice System will be studied. It must be mentioned here that the DB system of the specific DB system is very complex and e-collaboration, as well as data sharing, is being used extensively to assist the public services. Part of the Dutch criminal system is also the WODC (*Research and Documentation Centre*), which contributes in policy development and evaluation, in order for the services provided by the Ministry of Security and Justice to be enhanced.

Research performed by WODC concerns people that live in the Netherlands and have a criminal background. Because of the nature of the data, every party involved must follow a strict set of regulations and legal protocols in order to secure the privacy of the individuals under examination. It has been noticed that names and generally entries in different organizations may not be registered in the exact same way. So, it is essential to provide a DQ management framework during a preliminary DQ control that will aid in the direction of understanding the reason that mismatches occur and suggest ways to improve them.

1.4 Research Questions

All the above lead to the following research question that is divided in three sub-questions:

"*In what ways can data quality be managed prior to its encryption, when privacy-sensitive structured data is being shared among different organizations within the Dutch criminal justice system, taking in consideration TTPs as intermediaries?*"

- *Rq1* - Which DQ dimensions and methods are relevant when trying to link privacy-sensitive data among individual DBs?
- *Rq2* - What would be the typical structure of a record linkage process between two individual databases with the use of a TTP, when having a different final recipient?

- *Rq3* - In what ways can the DQ concept be linked with Record Linkage of individual DBs, when containing structured and privacy-sensitive data?

1.5 Research Approach

The general steps of Design Science Research Methodology (Peffers et al., 2007) inspired us to proceed with the construction of the research approach. After identifying the problem and setting the objectives, a literature research is conducted concerning the concepts of Data Quality, Data Integration, Record Linkage and Trusted Third Party. This is combined with interviews with domain experts in order to produce the Trusted Data Linkage Framework (TDLF) that will provide ways to solve the problem. In order to evaluate TDLF, a case study is performed concerning a process through which medical data are linked within the Dutch Justice system. Finally, the communication step is addressed when reaching the conclusions, limitations and further research points.

2. Data Quality

A generic definition is the one that describes DQ as "*fitness of data for use*" (Mendes et al., 2012; Wang and Strong 1996; Huang and Stvilia, 2012). Furthermore, DQ is expressed in term of DQ dimensions. As a DQ dimension, we will consider any component of the DQ concept (Stvilia et al., 2007). The most important DQ dimensions will be deliberately explained. Finally, some of the most important DQ methods will be mentioned, in order for the reader to understand the different ways that DQ can be managed.

2.1. Data Quality Dimensions

When it comes to DQ Dimensions, there are various definitions of each one of them, depending on the situation at hand and there are also many ways to classify them. In addition to that, Bobrowski, Marré, and Yankelevich (1999) state that "*achieving a high score in one of the dimensions does not mean necessarily that the DQ will be of the desired level*". Nevertheless, it is suggested from Batini et al. (2011) that the quality of structured and semi-structured data to be linked is being measured most of the times by the means of Accuracy, Completeness, Consistency and Currency (Timeliness).
As far as Accuracy is concerned, Naumann et al. (1999) refer to it as the percentage of objects without errors. Scannapieco, Missier and Batini (2005) state in their paper, that there are two main types of *Accuracy*, namely *Syntactic* and *Semantic*. The *Syntactic Accuracy* means that the syntax of a string is not correct and semantic accuracy refers mainly to the cases that the syntax of one string is correct, but has a different value than the one it should. An example of Timeliness definition can be found in the paper of Bobrowski et al. (1999), where Timeliness is expressed as "*the state of date to be up-to-date as needed*". Completeness is essentially the presence of NULL values in relation the universe of real-world objects. There are many ways to classify them, like in the papers of Pipino et al. (2002) and Scannapieco et al. (2005). Finally, Consistency concerns subjects like format errors and integrity constraints violation. A definition from the ones found is Zhu and Wang (2010), who define it as "the extent to which data is in the same format".

2.2 Data Quality Methods

Furthermore, depending on the situation, different methods exist that aid in the direction of measuring or assessing the concerning DQ that have the basic steps mentioned in the

introduction. The first stage "*State reconstruction*" presents the collection of contextual information on organizational processes in order to define DQ. The second stage (DQ assessment) includes the actual measurement of DQ, concerning the identified DQ dimensions of the first stage (State reconstruction). In the end, taking as input the results of the previous stages, actions are applied to improve the overall DQ (DQ Improvement). Such methods can be distinguished depending on different characteristics (Carey, Ceri and Berstein, 2007). Whereas *Process-Driven* methods aim mainly at redesign processes in order to remove the root causes of the DQ problem, *Data-Driven* ones refer specifically to Data sources in order to fix the actual problems identified amongst the data. Furthermore, the sum of DQ methods have main target the measurement of DQ, and some of them have additional stages that try to improve the DQ in a specific established way. They can also be classified as general purpose where they can be applied in a wide spectrum of domains and activities, or on a specific data and application domain for a specific activity (e.g. record linkage). Last but not least, the DQ methods can refer to a specific organization (Intra-organizational) or to a group of organizations and the interaction among them.
By conducting a literature research, the methods identified were categorized by the above classification. Thus, we conclude to *Table 1* (see Appendix).
First of all, the redesigned process at hand requires the identification of the mistakes in the data itself, in order to proceed with the improvement of the DBs as information sources. Thus, we are talking about *Data-Driven* DQ methods that have the study of the actual data as their main concern. More than that, the actual *measurement* and reasoning of the occurring mistakes worked as a main trigger for this research, but on the other hand *DQ improvement* is considered to be an essential part of the process. Nevertheless, the process should be able to adapt to the specific activity of performing a preliminary DQ control prior to the merging of two individual DBs using a TTP within the Dutch Justice domain. Thus, HDQM (Heterogeneous DQ management) from Batini et al. (2011) is found to be the most eligible from the ones found in literature and will be used during the case study.

3. Data Integration

In this chapter, an effort will be made to explain the process of integrating individual DBs, while using a TTP as an intermediary. First the term Data Warehousing will be explained, then the Record Linkage process will be shown, as well as the problems of data that cause linkage problems. In the end, the TTP concept will be analysed by showing its role in the integration process.

3.1 Data Warehousing

While searching through the literature, it was found that record linkage and *Data Integration* terms belong to the broader domain of *Data Warehousing*. When referring to this term, we mean "*a repository into which all the data relevant to the management of the organization and from which emerge the information and knowledge of needed to effectively manage the organization*" (Watson, 2001). March and Hevner (2007) essentially categorize the term *Data Warehouse* in 4 layers. It must be noticed that for each upper layer is dependent on the lower layers. All stages contribute in the acquisition of information and knowledge by the Organizational decision makers (Top Management) and all involved parties that can be used as insight in order to proceed with changes that will contribute in the enhancement of the DB environment and finally their Business goals.

The essence of Data Warehousing projects is the integration of data into one coherent repository of information (March and Hevner, 2007). In DB Management, Data Integration is used In order to achieve successful Content Management Process, and specific ETL (Extract, Transfer, Load) processes are utilized depending on the project's requirements. As far as ETL is concerned, it consists of three main stages (Vassiliadis and Simitsis, 2009), namely Extract (identifying the sources of data), Transform (series of functions to improve DQ) and Load (loading of the transformed data into a final table).

3.2 Record Linkage

Record linkage is considered to be an ETL process. Data from individual sources, even if they are part of the same organization, have to be matched and aggregated in order to meet the business goals. Record linkage can be used to improve DQ and integrity (Christen, 2004), in order to allow re-use of existing data sources for new studies.
There are two types of Record Linkage (Deterministic and Probabilistic) as Clark and Hahn (1995) discuss in their paper. Throughout this paper, only deterministic approach is examined. The latter is used in order to decide if two entries refer to the same real-world object, by taking into consideration a combination of attributes as an identifier.
Record Linkage is achieved by performing three main stages prior to matching (Christen, 2004), namely *Standardization*, *Cleansing* and *Blocking*. During the first stage a check is performed if the data adheres to specific format and appropriate data transformations are performed. Cleansing is about finding and eliminating the most common mistakes among the data. Last but not least, Blocking is performed in order to reduce the search space by classifying the records in categories. As far as blocking is concerned, four methods were found in the paper of Baxter et al. (2003). From these, Standard Blocking is applied when following deterministic approach and Sorted Neighbourhood, Bigram Indexing and Canopy Clustering to the probabilistic approach.
When it comes to problems that affect the record linkage process, there are two types of them as categorized by Rahm and Do (2000). Firstly, Single- Source Problems occur when studying individually a specific DB. Secondly, the Multi-source problems occur when trying to aggregate/integrate more than one DB and this can lead to serious problems for the matching process.
Nevertheless, when studying a specific DB Muller and Freytag (2005) classify them in three categories. The first one talks about Syntactical anomalies, which are further divided in lexical errors, domain format errors and irregularities. Secondly, the category of semantic anomalies is divided in Integrity constraint violation, contradictions of dependencies, duplicates and invalid tuples. Last but not least, coverage anomalies refer to the presence of NULL values within the DB.

3.3 Trusted Third Party

As stated in the introduction, TTP is used as an independent liaison when trying to merge individual DBs from different sources. Pseudonymization is a very closely connected term to the TTP concept, since it talks about encrypting the data in order to preserve data privacy. The main property that a TTP should have is the Public reputation, since trust is essential part when sharing information (Katsikas et al., 2005). Braak et al. (2012) distinguish other two properties that the TTP should have. Firstly, it has to be completely trusted by all parties involved in a process, promoting a feeling of mutual trust among the organizations that collaborate/ share data. Furthermore, the TTP has the sole responsibility for integrating and distributing data from and to public organizations.

The function of the Data Integration process with use of a TTP is described in the paper of Christen and Churches (2006), in the case that two organizations function as data sources. During the first step, organizations 1 and 2 agree on a predefined secret key that will be used to encrypt the data. Next (step 2), organizations 1 and 2 send their data to the TTP that performs the record linkage without having a view on the actual data. This becomes feasible while the TTP uses individual keys for communicating with each one of the organizations. During step 3, the TTP is responsible to perform the record linkage and send information about the matched records to each one of them individually.

4. Interviews

In order to proceed with TDLF development, four interviews were performed with domain experts, one round with two WODC researchers (identification of DQ dimensions) and two rounds with two consultants at ZorgTTP, an established TTP with a specialization in medical data.

4.1 WODC interviews

To begin with, semi-structured interviews were performed with two researchers in the WODC (information consumers). After the interviewees were properly introduced to the DQ concept, questions were set on the DQ of the data located in the DBs of the Ministry. Each one of them gave his own definition on DQ, and they were asked to divide 100 points to the DQ dimensions (Chapter 2) while taking into consideration the DQ dimension Security. Interviewee 1 divided the points equally to Completeness and Accuracy (50 points to each one). On the other hand, interviewee 2 gave Completeness 50 points, 40 points to Accuracy and 10 to Consistency.

4.2 ZorgTTP interviews

In addition to the previous, two rounds of unstructured interviews were conducted in ZorgTTP. The first one had as purpose the description of the supervised data integration process and the second one included a deeper look in the way that a TTP pseudonymizes the data but also the checks and transformation that occurs on it prior to pseudonymization.

The main outcomes from the first interview were two. First, information was derived about the role of a TTP, based on the experience of the interviewees. When asked about the approach to match the data, it was concluded that deterministic approach is preferred to probabilistic approach, where a custom algorithm that is aligned with the business needs and the specific rules is used to normalize the different categories of the name. Last but not least, it was mentioned that prior to merging the individual DBs, double hashing is performed to the identifier that is going to be used for linking the records, and finally the data is encrypted and sent to the recipient(s). By hashing, it is meant that the one string is transformed into a different one, using algorithms that are developed for that cause (e.g. MD5 and AES). The hashing function is used to ensure that a privacy-sensitive value will not be revealed in processes of data integration and record linkage.

As far as the second interview is concerned, a closer look was performed on the pseudonymization process and what kinds of checks are performed on the data prior to hashing. It was found that various checks on Null values, the conformance to a specified format and the existence of invalid characters are performed on the identifiers to be used for merging the individual DBs.

Outcomes

Firstly, by combining the WODC interviews with the second ZorgTTP, it was derived that three dimensions are relevant when trying to merge individual DBs, namely Accuracy, Completeness and Consistency. By examining the definitions of DQ set in the WODC interviews, a proposed definition for the specific case would be that "*DQ is the accurate merging of records from different organizations that belong to the same real-world object, while preserving data privacy*". Furthermore, as far as ZorgTTP interviews are concerned, when using privacy-sensitive identifiers, double hashing is added to encoding in order to ensure data privacy.

5. Trusted Data Linkage Framework

TDLF is operationalized for the case that structured privacy-sensitive data of two individual DBs within the same DB system have to be integrated and sent to a different recipient. It is developed by taking into consideration the generic stages of a DQ method as described in Batini et. al (2009) and by adjusting the different approaches found in literature and by combining the results of the conducted interviews for each generic stage of it. It is used for a preliminary DQ control and it consists of three stages, as found in Chapter 2.2. Figure 1 shows an overview of the latter (see Appendix).
In order to perform the stages of TDLF, a DQ team has to be operationalized. This team has to be embodied from individuals that belong to the organizations that own the DBs to be merged. More than that, they must have prior knowledge on the data itself, in order to perform the assessment step in a meaningful and effective way. Since the nature of the data is privacy-sensitive, they must be authorized to have access to the specific DBs in order to implement the first two stages of TDLF.

5.1 Stage 1: State Reconstruction

In this framework, the supervised record linkage between two individual sources will be examined. This involves structured data that are to be merged with the intervention of a TTP, while having a separate party as the final recipient. The result of combining literature research with ZorgTTP interviews can be found in Figure 2 (see Appendix). First of all, P1 and P2 are the outcomes of pre-pseudonymization quality control. During this control, the DQ dimensions are Accuracy, Completeness and Consistency, as derived from the results of the interviews and the DQ definition proposed is used. Two rounds of hashing are used (P1',P2' and P1", P2" are the resulting datasets of each individual hashing round) and one encryption round (resulting in P") for delivering it to the final recipient.

5.2 Stage 2: DQ Assessment

During this stage, by examining the definitions of DQ dimensions (Chapter 2) and Data anomalies (Chapter 3), and by combining them with the DQ assessment approaches for each dimension found in the paper of Naumann and Rolker (2000), the Table 2 is constructed (see Appendix). As far as the DQ assessment approaches are concerned, Sampling refers to dividing the DB into smaller representative fragments, Cleansing techniques refer to methods used for finding and eliminating the most common errors in the data and parsing to finding a specific value among the data.
Accuracy was connected with Irregularities (differentiated from the rest of the rest of Syntactical anomalies), invalid tuples (mostly referring to misspellings since no specific rule exists to assess them) and duplicates. On the other hand, Consistency is related to Integrity constraints violation, contradictions of dependencies and Syntactical errors

because they refer to format errors and finally coverage anomalies to the Completeness since they are anomalies that address to the NULL values.

5.3 Stage 3: DQ Improvement

At this stage, the generic stages of unsupervised record linkage that were mentioned in Chapter 3.1 were used (Figure 3). Note that the initial circle denotes the initiation of the process to be performed. In addition to that, blocking is done individually and the results of this process are sent to the TTP together with the findings during the blocking. Individual data transformation and integrity constraint enforcement actions are being taken in order to correct the most common mistakes that occur amongst the data and finally blocking is used to aid in reducing search space and deriving useful information that can be used as input to be implemented as preliminary checks prior to hashing. By taking as input the first round of ZorgTTP interviews, it was concluded that Standard Blocking is the appropriate method to be used in this stage.

6 Case Study

The tangible goal of the following case study is to evaluate the developed framework and check the ways that this can be applied in a specific case within the scope of the Ministry of Security and Justice of the Netherlands, by embedding a specific method in its stages. Taking into consideration table 1, HDQM will be implemented in a way that it fits the case, its applicability will be tested by checking at what level it can be adjusted to it.

6.1 Stage 1: State Reconstruction

During this stage, the process is described and the DQ is defined by embedding the HDQM (Batini et al., 2011) into the TDLF. Figure 4 (see Appendix) depicts the introduced process by adapting Figure 2 to the case:
In this case, records from the Forensic Clinics department (FC) are merged with records from the JustID department (holds all the personal data of citizens that have interaction with Ministry of Justice) in order to be sent to WODC for further research. Firstly, WODC sends the specifications of the sample that wants to study to the FC department. After that occurs, P1 and P2 are the results of the preliminary DQ control. These are sent to the TTP, which performs two individual rounds of hashing on the identifiers that are used to link the datasets and merges them in one final dataset P". Finally, P" is encrypted and sent to WODC for further study. Two Resources (RS; namely FC and JustID), four Organizational Units (OU) (namely FC, JustID, TTP, and WODC) and four Conceptual Entities (groups of employees of the concerning OUs) were found to participate in the specific process.
As far as DQ dimensions are concerned, the two WODC researchers that were interviewed were asked for the level of each identified DQ dimension, by setting relevant questions and using a Likert scale (1 to 10). As presented in Table 3 (see Appendix), the results showed that Consistency (9.5), Accuracy (8.67) and Completeness (8.25) are indeed the only dimensions that are not of perfect level.

6.2 Stage 2: DQ Assessment

Based on the identified DQ dimensions of the previous stage, the actual DQ of the data was measured using Table 2. Due to privacy issues, the Juvenile Delinquency (DJI) DB was used instead of the FC in a joint table with the relevant JustID entries. Resources in this case were found to be of equal ranking. The common attributes that existed in the

tables were First Name, Infix, Last Name, Birth Date, Birth Place and Birth Country region code. The results of this assessment are summarized in Table 4 (see Appendix).

Consistency

To assess this dimension, the first action that was taken was the definition of a specific format that the records should follow. By specific format, it is considered mainly that two records from individual DBs, which refer to the same real-world entity, should have the same number of tokens (i.e. field name of entity A contains two strings and so should entity B). More than that, it should be checked if the fields name and surname contain the correct string or if they are inverted. Furthermore, the fields must not contain spaces and must every string should be expressed in capital letters. When parsing towards this format, it was found that the main errors were different number of tokens and name/surname inversion, since capitalization and trimming is a standard correction when merging data.

Accuracy

During this assessment, the generic cleaning approach was used, in order to identify the most common mistakes that occur, taking into consideration Table 2. Thus, we conclude that mainly irregularities and invalid tuples were found to affect the data. As far as irregularities are concerned, it was found that certain characters were not used in both relevant records. The characters that were found were comma, full-stop, dash and quote and implied existence of initials and double names and surnames.

Completeness

For this case, parsing towards NULL attributes was used, the result of which is presented in table 5 (see Appendix). After performing that, it was found that attributes First Name, Last Name and Birth Date have always a certain value in DJI DB, whereas the relevant entries in JustID reach a level of incompleteness up to 5.6%. The only attribute that is more complete in JustID than DJI DB is the one of Birth place, with 96.28% to 94.3% respectively.

6.3 Stage 3: DQ Improvement

During this stage, specific actions were taken in order to improve the overall DQ of the process, by following the guidelines of Figure 3. Sample data transformations were performed that included mainly the removal of illegal characters (Accuracy), name/surname inversion (Consistency) and different number of tokens (Consistency). As an effect, there was an overall increase of 5% during Standardization and 10% during Cleansing step while using the full name.

As far as blocking is concerned, Standard Blocking method was used. The selection of this method was made because a combination of attributes is used each time as identifier for matching records during WODC research. Furthermore, this method can be operationalized to derive interesting results and statistics for each attribute used. By choosing Birth Country region code as an attribute (8 regions in total) and examining the attributes First and Last name, interesting results were reached, the results of which are found in the table 6 (see Appendix).

In addition to the latter, the exclusion of categories of certain regions was tested in order to see how it affects the overall rate of mismatches and biases the results. As a result, Table 7 (see Appendix) is constructed.

By setting a significance level at 1%, it was configured that the only category that affects the mismatch rate is records originating from the Netherlands. All the other categories have a maximum of 0.64% effect on the results and thus they do not bias the results.

In this chapter, TDLF was evaluated and the applicability of HDQM was checked. Certain categories of mistakes were found to be among the data and mostly Accuracy and Completeness were found to affect mostly the mismatches' rate. During the improvement stage, sample transformations were performed and Blocking was used in order to extract useful information on the data that can be used as a step for implementation by a judicial TTP.

7. Conclusions and Further Research

During this research we examined mainly four concepts, the Data Quality, Record linkage, Data Integration and TTP. This was achieved by performing extensive literature research, interviews with researchers within Dutch Justice Domain and domain experts from an already established TTP. By taking into consideration the conclusions reached in the previous sections, recommendations are made for the specific case study, but also limitations and points for further research are presented in the following sub-chapters.

7.1. Revisiting Research Questions

After performing the case study, the following recommendations are made to the Ministry:

- A DQ team should be established that will be responsible of performing the steps of the TDLF, each time that a certain privacy protocol advices so. This team is not only required to have knowledge on Data Sources' structure to be integrated, but also on the data itself.
- By blocking the DB based on the country of origin, it was found that the number of entries affects the overall matching rate. Considering that difference in mismatch rate less than 1% when excluding a specific category is insignificant, only exclusion of Dutch entries can bias the results.
- The use of Birth country region code as a blocking key made feasible to derive interesting facts for the data and the frequency that specific anomalies occur.
- In order to avoid the use of privacy-sensitive fields like first and last name, the use of a universal identifier should be used universally in all the DBs within the Ministry's DB system.

7.2. Recommendations and Further Research

Limitations were present while conducting this research. First of all, privacy protocols within the Ministry forbid access to the actual FC DB, and a legacy DB of JDI was used instead with a limited number of privacy-sensitive attributes. Interviewing more people in that manner could possibly enhance even more the findings of this research and further support the results, since the number of the interviewees in this case is small and could be considered insufficient. As far as time limitations are concerned, the main aim of this research was to identify why mismatches occur among the data and to assess the level for each anomaly. Taking into consideration the duration of the internship, normalization rules for each category of names could not be performed during the Improvement step, and only sample transformations were performed to demonstrate how this step could be implemented.

To conclude, the TDLF framework could be used as an initial step for creating a method that can be used when a judicial TTP is introduced to the process of integrating privacy-sensitive data. In order to establish it as a method, it must not be only performed to other similar projects in the specific Ministry, but also in general cases that privacy-sensitive data is shared with supervised data integration. In this manner, additional DQ dimensions

could be identified that play a role while managing DQ prior to encrypted data integration.
In this research the DQ of integrating individual DBs with the intervention of a TTP was researched. However, it was considered that the encryption of incorrect names will always lead to data reduction. Nevertheless, developments in encryption methods could potentially allow some kind of quality control. It is therefore worthwhile to investigate whether certain hashing and encryption methods are more eligible to be used by a judicial TTP. In this manner, each method will be examined and its influences on the data can be demonstrated.

References

Batini, C., Barone, D., Cabitza, F., & Grega, S., 2011. A Data Quality Methodology for Heterogeneous Data. *International Journal of Database Management Systems*, 3(1), pp.60–79.

Batini, C. et al., 2009. Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 41(3), pp.1–52.

Baxter, R., Christen, P. & Churches, T., 2003. A comparison of fast blocking methods for record linkage. *ACM SIGKDD*.

Braak, S. van den & Choenni, S., 2012. Trusted third parties for secure and privacy-preserving data integration and sharing in the public sector. *Proceedings of the 13th Annual International Conference on Digital Government Research*, pp.135-144.

Bobrowski, M., Marré, M., & Yankelevich, D. (1999). A Homogeneous Framework to Measure Data Quality. mitiq.mit.edu.

Carey, M., Ceri, S. & Bernstein, P., 2003. *Data-Centric Systems and Applications.*

Christen, P., 2004. A Two-Step Classification Approach to Unsupervised Record Linkage. (Clarke), pp.107–116.

Christen, P., Churches, T. & Hegland, M., 2004. Febrl – A parallel open source data linkage system Data cleaning and standardisation. In *Advances in Knowledge Discovery and Data Mining*. pp. 638–647.

Christen, P., & Churches, T. (2006, October). Secure health data linkage and geocoding: current approaches and research directions. In National e-Health Privacy and Security Symposium.

Clark, D.E. & Hahn, D.R., 1995. Comparison of probabilistic and deterministic record linkage in the development of a statewide trauma registry. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pp.397–401.

Huang, H. & Stvilia, B., 2012. Prioritization of data quality dimensions and skills requirements in Genome annotation work. *Journal of the American Society for Information Science and Technology*, (850), pp.1–28.

Katsikas, S., Lopez, J. & Pernul, G., 2005. Trust, privacy and security in e-business: Requirements and solutions. *Advances in Informatics*.

Lenzerini, M., 2002. Data integration: A theoretical perspective. *Proceedings of the twenty-first ACM SIGMOD-SIGACT symposium on Principles of database systems*.

March, S.T. & Hevner, A.R., 2007. Integrated decision support systems: A data warehousing perspective. *Decision Support Systems*, 43(3), pp.1031–1043.

Mendes, P., Mühleisen, H. & Bizer, C., 2012. Sieve: linked data quality assessment and fusion. *Proceedings of the 2012 Joint EDBT*.

Müller, H. & Freytag, J., 2005. *Problems, methods, and challenges in comprehensive data cleansing*. Technicsal Report.

Naumann, F., Leser, U. & Freytag, J., 1999. *Quality-driven integration of heterogeneous information systems.*

Naumann, F. & Rolker, C., 2000. Assessment methods for information quality criteria. *IQ*.

Otjacques, B., Hitzelberger, P. & Feltz, F., 2007. Interoperability of e-government information systems: Issues of identification and data sharing. *Journal of Management Information Systems*, 23(4), pp.29–51.

Peffers, K. et al., 2007. A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, 24(3), pp.45–77.

Pipino, L.L., Lee, Y.W. & Wang, R.Y., 2002. Data quality assessment. *Communications of the ACM*, 45(4), p.211.

Rahm, E. & Do, H., 2000. Data cleaning: Problems and current approaches. *IEEE Data Eng. Bull.*, pp.1–11.

Scannapieco, M., Missier, P. & Batini, C., 2005. Data Quality at a Glance. *Datenbank-Spektrum*, pp.1–23.

Stvilia, B. & Gasser, L., 2007. A framework for information quality assessment. *Journal of the American Society for Information Science and Technology*, 58(12), pp.1720–1733.

Vassiliadis, P., & Simitsis, A. (2009). Extraction, Transformation, and Loading. Encyclopedia of Database Systems, 32.

Wang, R., Strong, D. & Guarascio, L., 1996. Beyond accuracy: What data quality means to data consumers. *J. of Management Information Systems.*

Watson, H. J. (2001). Recent developments in data warehousing. *Communications of the Association for Information Systems (Volume 8, 2001)*,*1*(25), 25.

Zhu, H., & Wang, R. Y. (2010). Information quality framework for verifiable intelligence products. In *Data Engineering* (pp. 315-333). Springer US.

Appendix

Tables

DQ method	Process-driven	Data-driven	Assess	Improve	Special purpose	Generic purpose	Intra-org	Inter-org
AIMQ (Lee et. al, 2002)	X		X			X	X	
TDQM (Wang, 1998)	X		X	X		X	X	
GQM (Bobrowski et. al, 1999)		X	X		X		X	
HDQM (Batini et. al, 2011)	X	X	X	X		X		X

Table 1: Identified DQ methods classification

DQ Dimension	Data anomalies	DQ assessment approaches
Accuracy	- Irregularities - Invalid tuples - Duplicates	- Sampling - Cleansing techniques
Consistency	- Integrity constraints violation - Contradictions of dependencies - Syntactical errors	- Parsing
Completeness	- Coverage anomalies	- Parsing - Sampling

Table 2: DQ anomalies and assessment approaches

Person	Completeness	Timeliness	Accuracy	Consistency	Security
DJI	8.33	10	8.5	9.5	10
Just ID	9	10	8	9.5	10
Average	8.67	10	8.25	9.5	10

Table 3: DQ assessment results

DQ dimension	Data anomalies	Assessment approach
Consistency	Syntax errors (Different number of tokens, untrimmed fields, capitalization, name/surname inversion)	Parsing
Accuracy	- Irregularities (use of illegal characters) - Invalid tuples (misspellings)	Generic cleaning approach
Completeness	Attribute incompleteness	Parsing

Table 4: DQ assessment results

DB	Full name	First name	Last name	Birth Date	Birth place
DJI	100%	100%	100%	99.97%	94.3%
Just ID	96.29%	96.1%	96.29%	94.4%	96.28%

Table 5: Attributes completeness comparison between DJI and Just ID DBs

Attribute	**Region (Worst)**	**Region (Best)**
First Name (Initial)	Antilles	Other Western Countries
Last Name (Initial)	Turkey	Antilles
First Name (After Transformation)	Turkey	Other Western Countries
Last Name (After transformation)	Turkey	Antilles

Table 6: Blocking results

Category excluded	**Prior**	**Post**
Total	0.00%	0.00%
Unspecified region	0.14%	0.06%
Netherlands	-3.56%	-4.01%
Morocco	0.13%	0.17%
Antilles	0.20%	0.26%
Surinam	0.03%	0.12%
Turkey	0.06%	0.19%
Other Western Countries	-0.22%	0.01%
Other non-western countries	0.64%	0.26%

Table 7: Region Exclusion Results

Figures

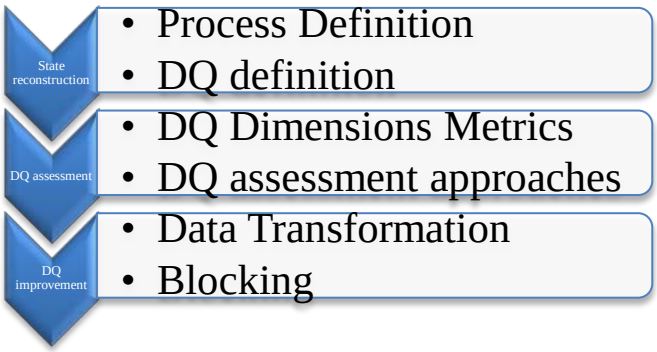

Figure 1: TDLF stages

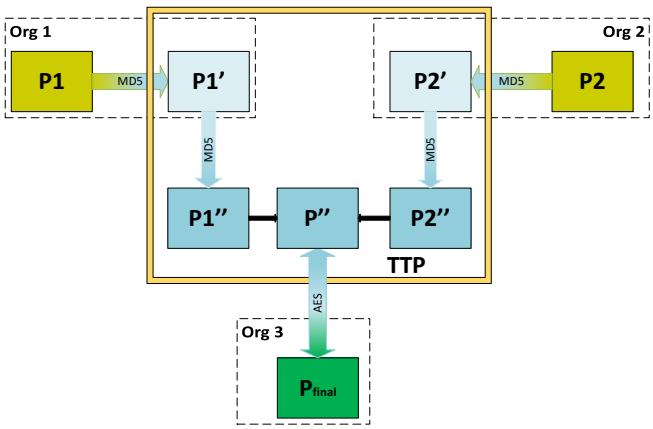

Figure 2: Data Flow Diagram

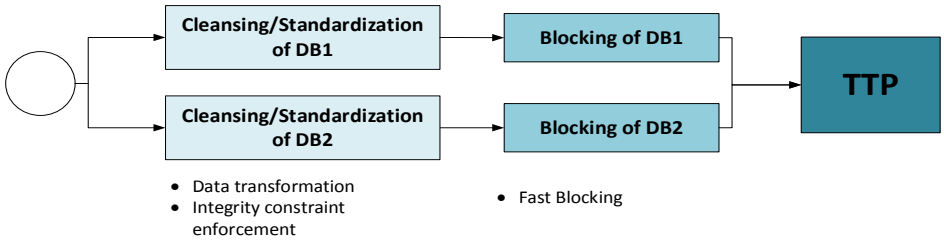

Figure 3: DQ improvement actions

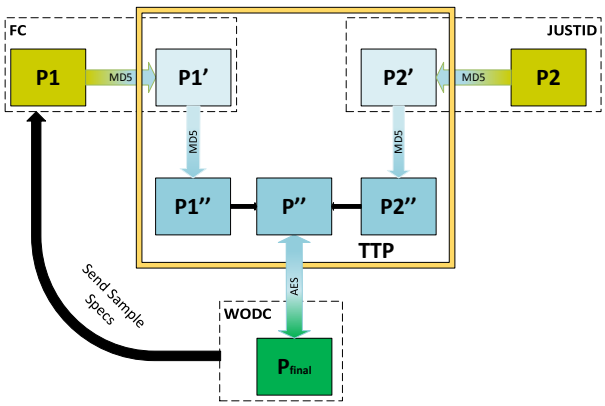

Figure 4: Case Study Diagram