Within the CPSI project Work Package 3 concerns the development of the data warehouse. This document will describe the way the data warehouse is designed by the following topics:

- What is the position of the data warehouse within the CPSI project
- Which requirements should be met by the data warehouse
- How does the data warehouse implement this

INTRODUCTION

The present project, CPSI, addresses security, its determinants and measures designed to improve security. It provides end-users, in the form of for example, governments, law enforcement agencies and emergency services with a methodology to increase insight into the determinants of actual and perceived security and into which interventions are effective for increasing actual and perceived security.

The results of the project represent tools, which can be employed by end-users, to help formulate policy regarding security. The project is divided into four objectives:

1. develop a conceptual model to describe the relationship between actual and perceived security and the determinants that influence this relationship
2. develop a methodology to collect, quantify, organize, analyze and interpret data on the factors described in the conceptual model
3. develop a data warehouse to aggregate and store data from different sources
4. conduct a validation study to test the model, methodology and data warehouse in the field.

A data warehouse (abbreviation DWH) is necessary as a central environment for all the data needed to analyse the security domain, as we have defined it in CPSI. To conduct the analysis work the DWH must integrate all security-related data from the validation study and distribute relevant information to the authorized users. The important subjects are:

- Actual (objective) security information
- Perceived security (subjective) security information
- Media (public opinion) information
- Cultural and public opinion factors information
- Intervention information
- Demographical information

REQUIREMENTS

The final deliverable is designed to be implemented in different countries in the European Community. In anticipation of differences between countries, the DWH can handle country-specific issues.

The structure of the data warehouse must guarantee high flexibility in order to adapt to a country's local requirements. For this reason, the effort for restructuring the data warehouse and the related costs must be as low as possible. At the same time, the data warehouse must be reliable, stable and future proof.

The data warehouse can store all this information, augmented on a periodic basis as new data become available. As mentioned before the DWH supports (qualitative and quantitative) analysis. So the DWH must produce output in two ways:

1. Reporting: This can be tabular or graphical reports, but also reporting from a geographical point of view. The DWH supports focussing down to the lowest granularity of the data.
2. Analyzing: To support statistical analysis an extraction can be made to process the information in a specific tool like SPSS.

The CPSI DWH gets its data from several sources from which the data differ in quality and format. In order to be able to transform the data to a suitable shape and quality for loading in the DWH, specific preparations have to be done first.

For enabling comparisons among data with different structures, content and types of information, the data warehouse brings a degree of uniformity. For security and privacy reasons no individual data is stored nor is the determination of an individual's identity possible.

GLOBAL SOLUTION

The way to implement the DWH is called the Business Intelligence and Data Warehouse (BI & DW) environment. This architecture enables the business to make better and faster decisions based on qualified business data. Therefore the data will be separated into two parts. The first part is the data warehousing (DWH), the second part business intelligence (BI).

Architecture

Data warehousing concerns the integration and storage of data historically. This is a complex task because the data has to be gathered from several external sources, which are not directly accessible or which are implemented on different technical platforms.

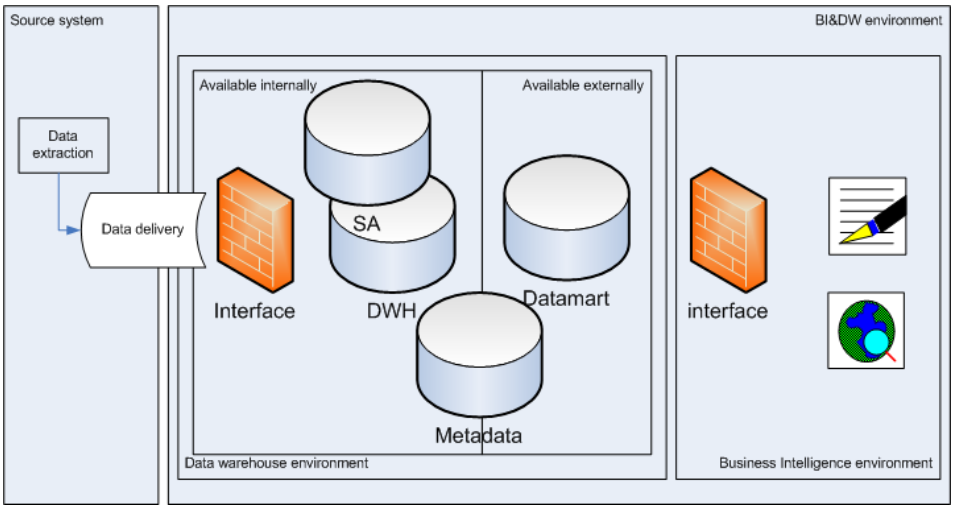


figure 1: Business intelligence & Data warehouse environment

Source environment: The DWH has no data of its own. The required data has to be delivered by external sources. These data are delivered by third parties in a way in which they fit into the DWH, which is specified in a well-defined and structured interface. For this project all data is delivered by structured files like excel, xml or comma delimited files.

DWH environment: In the DWH environment the delivered data is processed and stored. A staging area (SA) and data warehouse will be in place to integrate, consolidate and transform the data, giving it a suitable shape and quality. Finally the data is loaded into the datamart for the final purpose of analysis and reporting. The DWH itself is not directly accessible, the datamart is.

More in detail, the data is moved through the several environments. This process is called the transformation process (abbreviated ETL). All the environments fit together using well-defined interfaces.

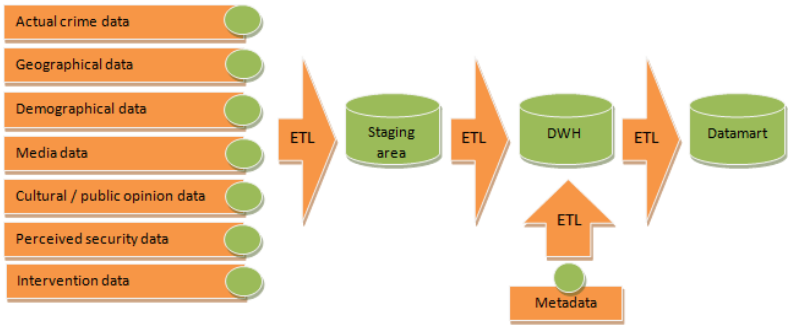


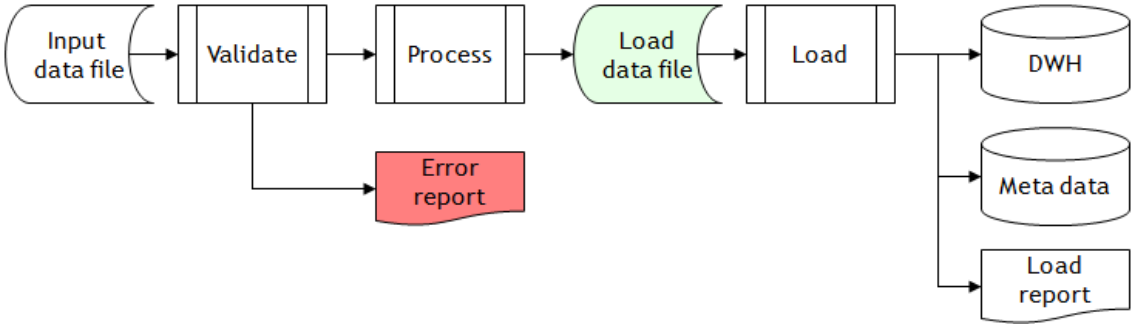
figure 2: Data movements in CPSI environment

The metadata describe all data elements which have been stored in the data warehouse. This concerns information about the data like name, definition and type, but also process data like when the data have been stored.

BI environment: This environment discloses the information required for reporting and analysis purposes. The interface shown is called the information model. Both tabular, graphical and geographic reporting as well as data exporting for statistical analysis (SPSS) are facilitated.

Transformation

Each process in the transformation is constructed in the following way:



First the data are validated. For instance, are the format and structure consistent with the specifications. If they are not, the errors are reported. If they are, the data will be processed and loaded into the DWH. Each data load will be monitored in the metadata database.

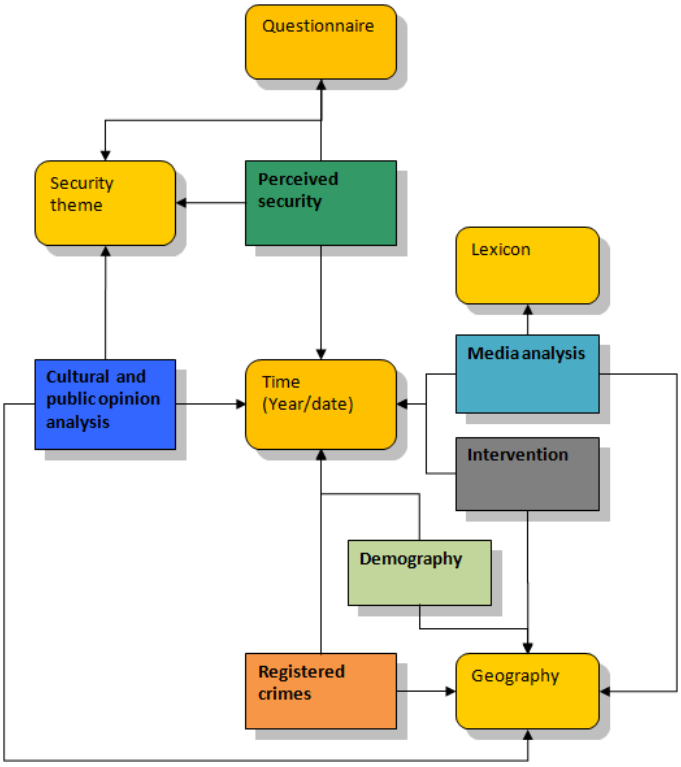
Data modelling

For each environment the security related data is modelled into a data model to meet its specific requirements.

The staging area is a one to one copy of all files that will be delivered. This makes the processing towards the central data warehouse easier.

The figure to the right shows the conceptual data model of the data warehouse.

The data warehouse is technically modeled consistent with the datavault technique. This technique efficiently enables the historical data storage, and enables flexibility and scalability for functional as well as technical extensions.



This model contains all information like the basic elements (facts) and the references.

<u>The facts are:</u>

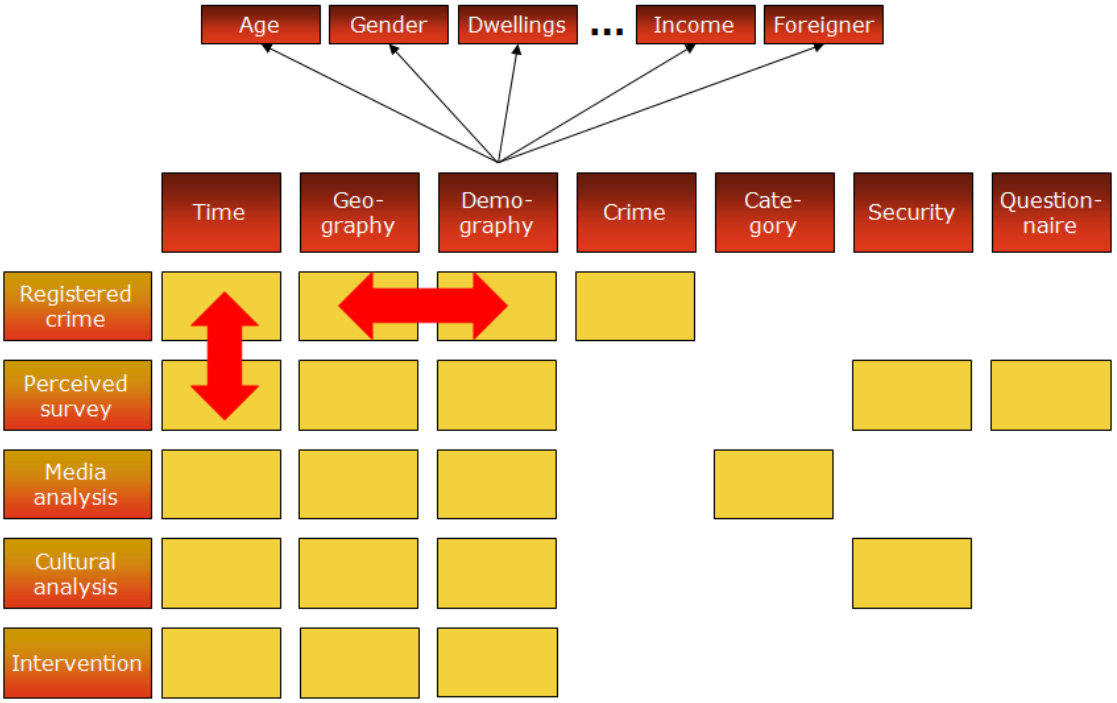
Registered crime	The number of actual committed crimes, which, when and where.
Perceived survey	The number of answers related to security themes, in time and where.
Media analysis	The number of hits on which media sources, category, when and where.
Cultural / public opinion analysis	Which security themes are influenced by which cultural and public opinion aspects in time and where.
Demography	Several demographic characteristics like age, gender, population density and income.

<u>The references are:</u>

Intervention	Which interventions were taken where and when.
Time	When: date, month, quarter, year.
Geography	Where: neighborhood, municipal, region, country.
Crime	What: offence and crime.
Lexicon	What: several media source characteristics like medium (TV, Radio, Newspaper, magazine), locality and type (news, opinion, glossy). What: the keywords within the categories for which the media sources were searched.
Security	What: security theme.
Questionnaire	What: question, statement, paragraph.

Information modelling

The business intelligence environment is modelled conform the dimensional model technique. This technique enables all questions to be answered easily and quickly.



The model's capability is the analyzability between the combinations of facts (vertically) and the dimensions (horizontally). A key position in the information model is the demography dimension because of its many demographical characteristics related to the geography dimension.

For instance, the following question can be answered:

- How does gender and age relate to the way security is perceived?
- How does this differ on a municipal level?

Analysis

The main objective of the data warehouse is to support the analysis phase in the CPSI project. Therefore a generic reporting system is constructed. This reporting system offers a framework in which a user can formulate his own questions and choose the way the output should be reported.