THE IMPORTANCE OF DATABASE MANAGEMENT SYSTEMS IN PSYCHOLOGICAL AND EDUCATIONAL RESEARCH - A SOLUTION FOR THE FUTURE

René Bochmann, Matthias Beggiato, Conrad Koczielski, Heiner Rindermann

Chemnitz University of Technology (GERMANY)

Abstract

The classical way to store and analyze empirical research data in social science and education research is program application. Excel and SPSS are typical examples. Data is stored in generic formats and analyzed with specific program scripts. Binary or text-oriented memory formats are commonly used, but their procedure is suitable if only small quantities of data have to be analyzed and little branched measurement have to be treated by one or few researchers. This system is problematic when (1) many people work simultaneously on data sets (uploading, editing and analyzing), (2) data have widely branched dependencies, (3) complex questions, complex query and analysis methods are required, and (4) Excel and SPSS are overwhelmed with the storage and display of large data sets. Therefore, it requires a change in the classical analysis path, for which relational database management systems (DBMS) can offer solutions.

A research project on teaching quality of university instructors (“SoKonBe”: Consulation of university instructors for improving teaching quality focusing social-cognitive conflicts) is used to show the applicability of the database management systems. In this contribution, we present a new way for a better, easier, faster and more valid storage and analysis of research data. In addition, we present a self-test to check the need for the use of databases in different research areas.

In the research project, we have actually stored 30 million records in approximately 300 variables. This is required because of the branched dependencies we previously addressed. The total sample contains data from 44 lectures by 43 instructors (male = 55.8%; female = 44.2%). The number of participating students per lecturer varies between \( N = 10 \) and \( N = 124 \) students. The total sample of the student participants contains 1,714 individual details. In the pretest of the project, the lecture of every instructor was filmed with three cameras on an average of 70 minutes. Every student filled in multiple questionnaires about the lecture, the instructor and the attractiveness of the lecturer. After this, up to five student research assistants analyzed the videos based on 70 categories. With a DBMS, it was possible to store, manage and aggregate the data.

Considering the steadily growing amount and heterogeneity of data (questionnaires, video, sensor data, interviews…), a database management system offers innovative, efficient and future-proof data analysis opportunities in the whole domain of psychological and educational research. Minimal data redundancy due to separate tables and relations, improved data consistency and data accessibility, multiple users as well as program-data independence are possible.

Keywords: database management system, research, big data, psychology, education.

1 INTRODUCTION

Every one of us produces massive amounts of data in many aspects of our daily lives. Communicating, reading news, listening to music, searching for information, shopping, banking, expressing our opinions, being mobile leave digital traces of behavior. This “datafication” [1] of human life aspects affects both social science and education research. The growing amount of technical opportunities to assess behavior creates a demand and, at the same time, offers new possibilities in these research domains. However, dealing with these large datasets containing millions of records of every individual poses new challenges for researchers.

The classical way to store and analyze empirical research data in social science and education research is program application. Excel and SPSS are typical examples. Data is stored in generic formats and analyzed with specific program scripts. Binary or text-oriented memory formats are commonly used, but their procedure is suitable if only small quantities of data have to be analyzed and little branched measurement have to be treated by one or few researchers. This system is problematic whenever (1) many people work simultaneously on data sets (uploading, editing and analyzing), (2)
data have widely branched dependencies, (3) complex questions, complex query and analysis methods are required, and (4) Excel and SPSS are overwhelmed with the storage and display of large data sets. This requires a change in the classical analysis path, where relational databases can offer solutions.

A database is a structured collection of datasets hosted by a database management system (DBMS). DBMS provides functions such as permanent and secure data storage, transaction control, maintenance of data consistency, user management and user-rights systems, backup capability and a data query language for accessing the data [2]. Many DBMSs are designed as client-server systems and allow the simultaneous access by multiple users. The tasks are done by programs that are divided into clients and servers. There are several types of database models. The relational model is frequently utilized in several database management systems, such as Microsoft SQL Server, PostgreSQL and MySQL. In the relational model, data is sorted in tables and the information stored in each row is called a record. The conceptual modeling of such a database is usually done with entity-relationship diagrams [2]. These diagrams support the developer in the definition and representation of the database architecture, i.e. the different elements (entities) and the links between them (relationships). Entities are independent sets of information that are included in the system. They have certain properties that correspond to the groups of recorded information about them. Relationships connect entities together and describe how the entities and associated information are related. Relationships can involve two or more entities. To establish these relationships, primary keys must be defined. The primary key is a unique identifier for any record in a relational database. Each relation is usually described as a table, which is organized into records (rows of a table) and attributes (columns of a table).

For beginners, the theoretical concept of the relations is not easy to understand. So it is useful to demonstrate this by an example: To store a dataset of participants of an experiment with their names and telephone numbers using a program like Excel, one column for each variable is necessary. That is also the common way. However, sometimes it is useful if some columns are separated and not accessible for multiple users. In this case, the column of the variable names must be separated from the columns of the telephone numbers. Two Excel-Files are produced and the information which name is linked to which telephone number is lost. To avoid this, a linking variable has to be created. In the easiest case it is simply a counting variable. So in both Excel-files (telephone number as well as name) the first variable has the same value in the linking-variable (e.g. “1”). Thereby the old relation between the variables names and telephone numbers is restored. DBMS stores different information in different tables, linked by a primary key (relation variable). For instance, the relation described in the example above is called a 1:1 relation. If one entity in the first table is related to more than one entities in the other table, the dataset becomes more complex (e.g., one lecturer is related to more than one student). Then it is called a 1:n relation. Only with a program like Excel this data relation is still manageable but not very clear. In the study of the relationship between instructors and students at different universities (SokonBe), a lot of n:m relations are produced. Because one instructor is related to different students, but the same student can be related further to different instructors. That is the most complex case. With this type of relation, Excel is overwhelmed.

In comparison to classical analysis tools in social science such as SPSS, Excel SAS, R or STATA, an intermediate step by storing data in relational databases offers some advantages ([2], [3], [4]):

- Minimal data redundancy due to separate tables and relations between tables
- Improved data consistency due to the guarantied database-principle of reliable “transactions”, founded on the ACID-properties (Atomicity, Consistency, Isolation, Durability)
- Improved data accessibility due to the client-server-principle, permitting parallel access by multiple users and sophisticated rights management
- Program-data independence, allowing access using various software products
- Amount of data only limited by hardware resources
- Only one server needs extended hardware resources for executing queries; clients send requests and receive the results
- All data is stored in one place, which simplifies backup and avoids confusion regarding most recent data

These database features are especially useful when dealing with large, heterogeneous and related datasets, such as sensor data with different frequency, questionnaires, video and interview data
collected in one study. Despite of all advantages listed above, databases do not replace classical analysis tools but extend them. In the data analysis process, databases cover the storage and retrieval parts (Fig. 1), whereas statistical analysis is still performed by common tools. The main operations inside the database concern structuring and aggregating data, which are indeed the main analysis step for large and heterogeneous datasets. The use of a database is especially recommended when dealing with large datasets (i.e. more than 1 million records), with relational data, when data sharing or parallel access is required, when aiming for persistent storage or when complex retrieval procedures are needed.

### Figure 1. Use cases and positioning of databases in the research process.

We continue to show the practical applicability and necessity of database management systems in psychological and educational research with the example of the research project SoKonBe (Consultation of university instructors for improving teaching quality focusing social-cognitive conflicts), which aims at improving teaching quality of university instructors. Numerous studies on the effectiveness of students’ evaluations of teaching quality have shown that it is not enough to provide the instructors with the evaluation results in order to improve teaching quality (for an overview, see [5]; [6]). In fact, the instructors need professional and external educational consultation for noticeable and stable improvements of their teaching ([7]; [8]). In the project “SoKonBe”, the instructors themselves, their students as well as external instructors evaluate the lectures – the external instructors by means of video sequences of the lectures (cf., [5], [9]; [10]; [11]). In the project, the teaching quality of the lecture is assessed by means of the Heidelberg Inventory for Evaluation of Teaching (HILVE II; [6]).

In this contribution, we present a new way for a better, easier, faster and more valid storage and analysis of research data. In addition we present a self-test to check the need for the use of databases in the own research field.

## 2 METHODOLOGY

### 2.1 Sample

In the research project, we actually store 30 million records of approximately 300 variables. This is required because of the branched dependencies we previously addressed. The total sample contains data from 44 lectures of 43 instructors (men = 55.8%; women = 44.2%). The number of participating students per lecturer varies between $N = 10$ and $N = 124$ students. The total sample of the student participants contains 1714 individual details.
2.2 Procedure

All of the instructors of the project held a teaching unit of their topic in front of their students, which was recorded on video. After that, the instructors themselves and their students evaluated the lectures by a questionnaire (HILVE II). The external instructors evaluated the teaching quality by means of video sequences of a web portal (also with the HILVE II). In the next step of the project “SoKonBe”, the video sequences of the lectures are analyzed by means of 70 different categories of the instructors’ behavior and compared with the results of the evaluation. The video analysis was carried out using the software "Eudico Linguistic Annotator" (Elan; [12]). This program allowed video or audio recordings to be marked with so called annotations (markings). Thus, the duration of specific behavior of the lecturers during the entire lecture period (in milliseconds) can be characterized. As studies have also shown, merely providing instructors with students and external instructors’ evaluations is not enough to improve teaching quality. Instructors need additional educational consultation ([6]; [7]; [8]). Therefore, the instructors will receive different forms of external consultation aimed at improving their teaching quality. These variants of consultation will be compared in this project in order to identify the most effective form. Then, the lectures will be evaluated once again, and the possible changes in teaching quality will be assessed.

3 DATABASE APPLICATION

As mentioned in the introduction, the use of databases is especially useful for large, heterogeneous and related datasets. However, for concrete research projects, some practical questions remain such as "How large is large?" or "How do I know if datasets are heterogeneous and/or related?" To tackle these practical issues (independently of the research domain), the self-test questionnaire presented in Figure 2 can be helpful. The more questions are answered with "yes", the higher the recommendation to use a database.

<table>
<thead>
<tr>
<th>Question</th>
<th>Recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Do you enter the same information in several tables?</td>
<td>Yes</td>
</tr>
<tr>
<td>2. If you make changes in one table, are you forced to make the same changes in other tables?</td>
<td>Yes</td>
</tr>
<tr>
<td>3. Do you work with amounts of data that are increasingly difficult to manage and process? Do you work with more than 1,000,000 records?</td>
<td>Yes</td>
</tr>
<tr>
<td>4. Does it take a long time to select data or to perform calculations?</td>
<td>Yes</td>
</tr>
<tr>
<td>5. Do you select data from a table and therefore you have to use information from other tables?</td>
<td>Yes</td>
</tr>
<tr>
<td>6. Do several people have to work with the same data at the same time?</td>
<td>Yes</td>
</tr>
<tr>
<td>7. Are you tracking related information in several tables, such as separate tables for every person with the same structure and content?</td>
<td>Yes</td>
</tr>
<tr>
<td>8. Are several versions of a dataset stored at different locations and it is unclear what is current?</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Figure 2. Self-test to check the need for the use of a DBMS in different research areas.

PostgreSQL (https://www.postgresql.org/) is a stable and fast DBMS for executing queries on large datasets. It is free and easy to install on various operating systems; documentation is comprehensive and the included geospatial database extension PostGIS offers excellent opportunities for analyzing geographical data such as GPS (Global Positioning System) data. In addition, recent PostgreSQL versions starting from 9.0 include powerful extended functionalities, such as common table expressions and window functions [13] which are summarized under the abbreviation “OLAP” (Online Analytical Processing).

For our Project, we use PostgreSQL DBMS 9.1. For the data collection, we linked our DBMS with several Web-Based client applications (Fig. 3). For the data upload of our paper and pencil based questionnaires, we use some LabView-based programs. LabView is a system-design platform and development environment for a visual programming language. All the client software are self-developed with the help of psychologists, technicians and programmers. The implemented mathematical functions (e.g. MEAN, SD etc.) are useful for data aggregation. Higher functions which were not implemented in the base package are developed in a client program (e.g. pgAdmin) and are available for all project member.
Media files such as video, pictures or audio are not stored entirely inside the database. These files are located on a separate fileserver and the database tables only contain a reference, for example, the filename and frame number for each video. This structure allows keeping the database size smaller and therefore it operates faster, whereas the link to the media files is always present (Fig. 3). In the project SoKonBe, all of the 43 instructors of the project held a teaching unit (between 40 and 60 minutes) of their topic in front of their students, which was recorded on video (two cameras with a fixed frame rate of 30 fps; mp4 format). The video analysis was carried out using the software Elan ([12]). In the next step, four files (two video files, one audio track and one file with 70 categories for Elan) were produced. In order to maximize the objectivity of the results, the video sequences are reviewed and annotated by five trained students from the department of psychology. After this the data of each annotation file were exported to a raw text format and uploaded to our database. One example of the structure of a variable is listed for better understanding (Fig. 4).
Every instructor and every student have an identification number (id-code) and every dataset has an additional counting number (primary key of three variables). Unfortunately, this structure is not very suitable for a deeper analysis because the categories of the annotation process are not database categories. Therefore, a SQL-Statement has to temporarily convert every data row (record) in a column (category). Unfortunately this is an extremely time consuming process. To avoid this problem, it was necessary to carry out this conversion process simply before the analysis. For this task, a specific LabView program was designed. This step takes about two days and converts the approximately 180,000 data rows of the table above into approximately 60,000,000 data rows. The structure of these expanded data is shown in Figure 5.

![Figure 5. Data structure for Elan after conversion.](image)

Every picture of the video file has an additional code (frame/picture rate; 30 frames per second) and leads to this great expansion of the data. But only now it is possible to see differences in rating between multiple students (<category_n> is rated to true or false).

We mention this example because it shows how DBMS deal with a large number of data. If a user faces the task to compare two binary formatted variables with a logical function (like &&, or, xor etc.) without the help of a DBMS, he will have to compare 11,334,000 records (70 categories x 90min x 60 seconds x 30 frames per second) by hand. If the user needed only 0,5 seconds for each comparison, the work would be done in approximately 1,575 hours. But in this case the user would have analyzed only the difference between two students on one video sequence. In the project SoKonBe, the video sequences are reviewed and annotated by five students.

4 CONCLUSIONS

Considering the steadily growing amount and heterogeneity of data (questionnaires, video, sensor data, interviews…), DBMS offers innovative, efficient and future-proof data analysis opportunities in the whole domain of psychological and educational research. Minimal data redundancy due to separate tables and relations, improved data consistency and data accessibility, multiple users as well as program-data independence are possible.

Alongside the benefits of a database, new challenges arise for social scientists. Firstly, the complexity of the data structure increases due to several related tables. However, this fact is mainly related to the inherent complexity of the studies, e.g., using different assessment methods such as video, interviews and multiple sensors. Databases help to deal with data complexity and do not create it. Secondly, installation and management of the DBMS is required. Due to the client-server principle, the DBMS runs on one server and it is not installed on every researcher’s computer. Installation and management including rights management, backup strategy, failsafe hardware configuration… etc. should preferably be carried out by one specialized person, i.e. a computer scientist. However, as a third challenge, researchers need to acquire knowledge about the database, especially about querying the database using SQL (Structured Query Language). This task should preferably not be outsourced to computer scientists, because

1. researchers should have direct and immediate access to the data,
knowing the extended analysis potential of databases (e.g., OLAP-functions) offers new opportunities for data analysis,

researchers know their data best and should be able to explore, check and “play” with the data without restrictions and

due to the constant and irreversible growth of measurement data in research (e.g., physiological data, eye-tracking data, video, GPS, web-logfiles... etc.), social scientists should acquire analysis competence for this type of data to be ready for future research challenges.

To address the specific DBMS training needs of social scientists, a four daylong seminar has been developed and carried out several times at the Institute of Psychology at TU Chemnitz. The seminar includes all necessary steps from deciding if a database is needed to the installation procedure, structure design, data import, simple to complex queries and exporting/accessing aggregated data from other analysis tools. All tasks are immediately trained by working with the real database. This seminar determined the basis for other researchers and students to perform secondary data analysis of already collected data with new research questions. Beside education research in the project “SoKonBe”, databases are widely used for psychological research concerning human factors issues in transportation at TU Chemnitz. Examples of recent projects using database analyses are ADAPTATION [14, 15], DriveMe [16] or UR:BAN [17]. Moreover, the use of databases as common data handling tool has strongly enhanced interdisciplinary collaboration at TU Chemnitz, e.g. with physics [18], ergonomics [17] and communication engineering [17].

ACKNOWLEDGEMENTS

We thank all instructors and students for their cooperation and participation. This project is funded by the German Federal Ministry of Education and Research (BMBF), in the context of the funding priority “science and university research”, under the funding guideline “accompanying research of the pact for quality of teaching”. The project runs from 10/2014 till 09/2018, Grant # 01PB14002.

REFERENCES


