A formal language for the specification of matching algorithms as a general framework for pseudonymization

M. Wagner, K. Pommerening

April 10, 2003

1 Introduction

Projects in clinical research are often distributed over several places. Information on the subjects must be well coordinated, because different locations of data entry means different standards of data quality and data sureness.

Since 2000, a pseudonymization service is developed in the course of several data security projects at the University of Mainz [4, 2]. The main objective is the coordination of medical professionals when dealing with sensitive patient data, which may be unsure or incomplete. Besides that there should be means to adapt the system to almost any environment with other data and other requirements.

The adaptation requires a special layer that interfaces the thoughts and ideas at the conceptual level with the concrete procedures at the algorithmic level. This layer should provide some kind of abstraction that hides the details of database queries and other programming issues.

2 Methods

To enable a precise adaptation a formal language was developed that allows the specification of a complete pseudonymization scheme in a single configuration file. Such a scheme should describe any conceptual detail that characterizes a local installation of the system.

A pseudonymization scheme consists of three parts. The first section defines the different fields a patient record is made of. This includes a type, length restrictions as well as an equality statement, which defines precisely, under which conditions two data items are considered to be equal.

The second section defines a set of results the matching process may produce. Each result includes a symbolic name to be referenced later in the specification, a pseudonym retrieval mode, which decides if an existing pseudonym is retrieved or if a new one is generated, as well as an update mode, which causes an existing record to be completed upon a request.

The third section defines the matching algorithm itself, formally as a finite state system. A set of single database queries represents a number of tests, which build the points of decision within the procedure. The tests are connected by links, which are attached to their entry and exit points. These links will build the sequence of tests performed on a request, depending on the result each test produces.

With the specification language we introduced the so-called KSXO notation, which specifies a database query in a special, problem-oriented manner. Four parameters describe, how the query will be constructed. The key restriction (K) may

demand a matching key field, e. g. an insurance company code. The sureness restriction (S) selects database records marked as sure, or unsure, respectively. The exactitude restriction specifies, if fields are compared by strict equality or in terms of similarity, using different phonetic codes. The optionality restriction (O) defines, if optional input fields are used for comparison.

The decision process begins with a special test, which incorporates the starting state of the matching procedure. The corresponding test will be executed and a numeric result will be produced. This result will select exactly one of the test's exits, which leads either to another test or to a result. In this way, the process continues until a final result is reached.

3 Results

The schema specification language has proven to be very close to the algorithmic problem itself. Issues that were brought in at a conceptual level could be implemented with the language very quickly, in a precise and natural manner. Besides that, the operational management of different installations became relatively simple, because any details of a pseudonymization service at a specific site are completely encapsulated within a single file.

4 Discussion

The process of patient data pseudonymization involves many different tasks at different levels of data processing, beginning at the syntactic level, where normalization is done and delimiters are recognized, and ending at the semantic level, where decisions are made on the equality of records. However, the existing work on these topics focusses on specialities, e. g. the use of phonetic transformation [1], the construction of cipher codes [3] or the procedural matching based on stochastical methods. In contrast, our language provides a unique specification method, which includes all of the details mentioned above. In fact, many attributes of data transformation are not analyzed by the parser, but directly passed to the corresponding module. Thus, we have a well-structured specification method, which is easily extensible and independent of special transformation techniques.

References

[1] J. Michael. Doppelgänger gesucht - Ein Programm für kontextsensitive phonetische Textumwandlung. *c't*, 25:252–261, 1999.

[2] K. Pommerening and M. Wagner. Ein Pseudonymisierungdienst für medizinische Forschungsnetze. In *Informatik, Biometrie und Epidemiologie in Medizin und Biologie*, Köln, Germany, 2001. Urban & Fischer.

[3] I. Schmidtmann, H.-J. Appelrath, J. Michaelis, and W. Thoben. Empfehlungen an die Bundesländer zur technischen Umsetzung der Verfahrensweisen gemäß Gesetz über Krebsregister (KRG). 27:101–110, 1996.

[4] M. Wagner and K. Pommerening. A reusable pseudonymization interface for epidemiologic research. In *46th Annual Conference of the GMDS (GMDS 2001)*, Köln, Germany, September 2001. Poster.