Population based Cancer Registry Bavaria - Registration office

Bevölkerungsbezogenes Krebsregister Bayern - Registerstelle, Carl-Thiersch-Straße 7, D-91052 Erlangen, Germany
http://www.ekr.med.uni-erlangen.de e-mail: krebsregister@ekr.med.uni-erlangen.de

Automatic Validity and Consistency Check of Cancer Incidence Data using an SQL-based Dictionary

Martin Meyer, Martina Franzkowiak de Rodriguez, Matthias Land

Source of cancer records

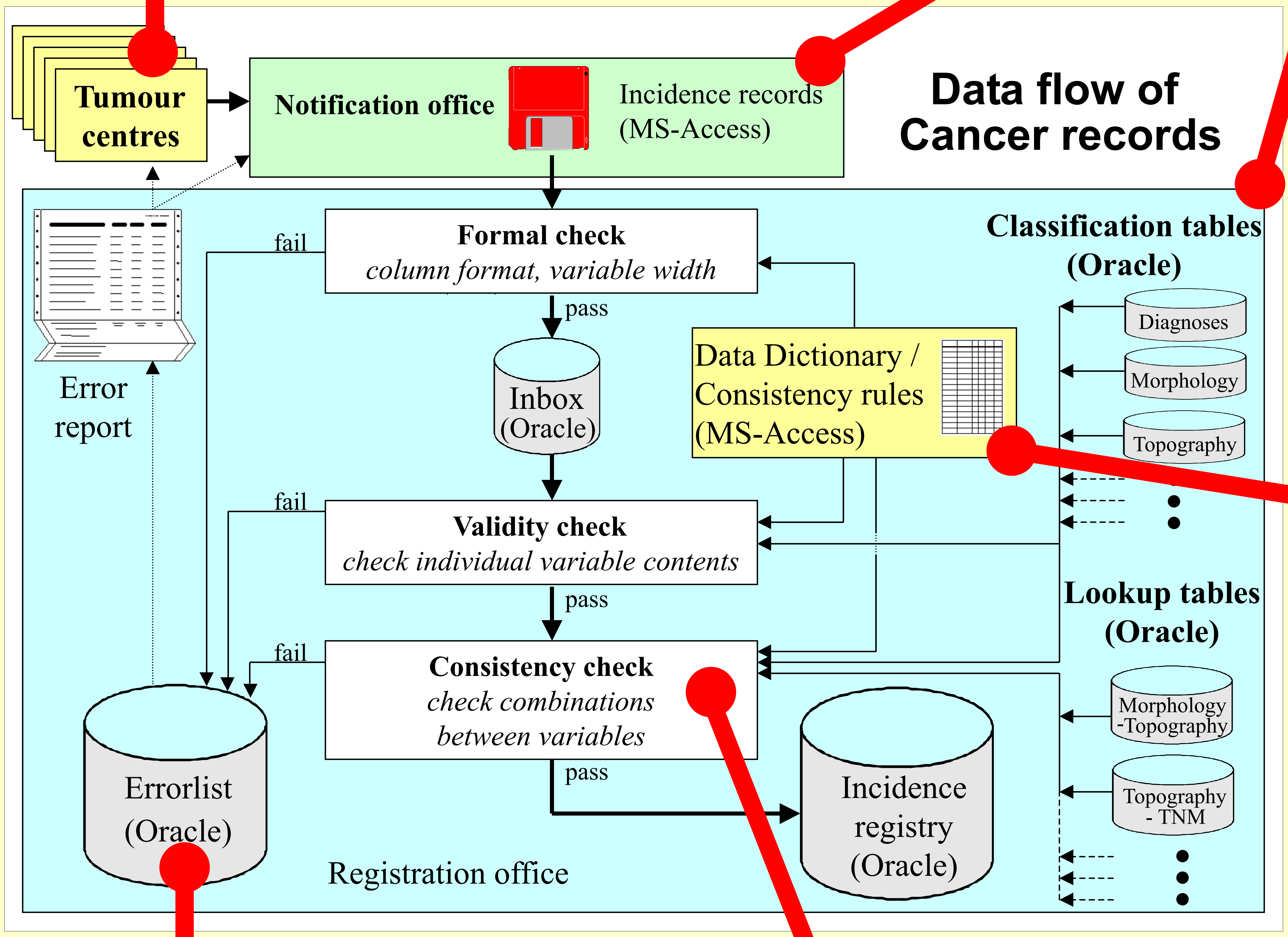
Five regional Bavarian tumour centres collect cancer records and enter the data into their clinical database.
The epidemiological part of the cancer records is sent on floppy disks to the cancer registry's notification office. A detailed transfer file description has been passed to the tumour centres in order to guarantee homogeneity of file and variable formats.

Incoming records

The registry's notification office merges the files received from the tumour centres, encrypts names and other identity variables and submits the result to the registration office. Both offices agreed on a high level file format (MS Access) for this data transfer to avoid any common problems associated with low level text formats like linefeed/carriage-return definition, special characters or ASCII/ANSI coding.

Documentation standards

Classification of medical items (diagnoses, morphology, topography and others) is done by the tumour centres. International standards are used if available (for example ICD-10, ICD-O-2, TNM). The registry provides tables of these medical coding systems to check the validity of incoming codes and to display full text information if necessary.



Data Dictionary

Four worksheets in an external file contain the definitions of all formal, validity and consistency checking rules. Choosing the way of a dictionary file independently from the database leads to a very flexible and user-friendly data import tool. Adding new checking rules or editing existing rules can easily be done without recompiling the database application.

Worksheet	Type of Contents	Contents
Column definition	Data definition	Import column name Oracle column name Description / comments
	Format checks	Variable type (numeric/text/date) Column width
	Validity checks	Obligate / optional fields Missing values Ranges (minimum/maximum)
Label list	Validity checks	Labels
Lookup table list	Consistency checks	SQL commands to access lookup tables
Consistency rules	Consistency checks	SQL commands to execute consistency rules

Lookup tables

For some variables it is possible to create a table of all possible combinations (e.g. topography-morphology or topography-sex). Each entry of such a lookup table can be marked as a permitted or forbidden combination of values for use in a cancer record. It is a fast method to query an indexed lookup table for the imported values. Lookup queries are performed for each imported data record.

Example:
```
select count(PK_CODE) RESULT from T_TOPOGRAPHY
where PK_CODE=:IMPORTCODE and  (SEX='0' or SEX=:IMPORTSEX);
```
This rule will pass (RESULT>0), if the topography code intended to be imported (:IMPORTCODE) is found in the Oracle table T_TOPOGRAPHY and if this topography code is marked as not gender specific (SEX='0') or if it is marked as specific to the imported gender (:IMPORTSEX).
Many consistency checks can be handled with lookup tables, nearly all of them just need simple SQL commands similar to the given example.

Consistency rules

For some consistency checks a lookup table cannot be created (e.g. quantitative relations like "birthday must be before incidence date"). Moreover in some instances it is not efficient to create lookup tables for each possible combination of variables. All error conditions not handled with lookup tables are written to the table *consistency rules* of the data dictionary. Consistency queries are performed only once on all imported data records.
The common SQL scheme for an error condition is:
```
   select RECORD_ID, SENDER_ID from T_INBOX
       where Errorcondition = TRUE;
```
Only the error condition varies between rules. Therefore, it is sufficient to enter the condition expression into the data dictionary. Any logical, relational or arithmetic SQL operator and SQL function may be used.

Example:
```
   substr(MORPHOLOGY,1,3)>='959'
   and substr(MORPHOLOGY,1,3)<'972' and CELL_TYPE is null
```

Error handling

Any violation of a format, validity or consistency checking rule is stored in the relational database table *Errorlist*. The error records are linked to the original data records via the record ID. An error record consists of record ID, sender ID, error class, error number and a full text error message listing the related field values. Depending on the sorting order of a errorlist query different error reports can be created: sender specific, case specific or error type specific. The crosstabulation of sender ID and error class enables detection of systematic errors.

Advantage of SQL-based rules

- no relevant complexity limits
- maximum flexibility
  if stored in an external worksheet
- rule portability
- no need to implement a rule interpreter
- high system performance
  by client/server architecture

Conclusions

Separating data definition from program code is a standard technique of software engineering. We expanded this principle to rule definition and execution of any complex validity and consistency rules.
The complete import process containing record handling, validity checks for 105 variables, check of 83 consistency rules and creation of detailed error messages could be implemented with a rather small C++ module (<4000 lines of source code).
We were able to show that this approach is a flexible and efficient tool to collect external data, when it is applied to real data records.

Acknowledgement

This work was partially supported by the Deutsche Krebshilfe e.V.

References

[1]D.M. Parkin, V.W. Chen, J. Ferlay, J. Galceran, H.H. Storm and S.L. Whelan, Comparability and Quality Control in Cancer Registration. IARC Technical Report, Lyon, 1994.
[2]L.H. Sobin, Ch. Wittekind (eds), TNM Classification of Malignant Tumours, fifth edition, Wiley & Sons, New York, 1997
[3]J. Dudeck, G. Wagner, E. Grundmann, P. Hermanek (eds.), Basisdokumentation für Tumorkranke: Prinzipien und Verschlüsselungsanweisungen für Klinik und Praxis, 5. rev. Auflage, Zuckschwerdt München, Bern, Wien, New York, 1999