Списание „Математика, компютърни науки и образование“
том 1, брой 1, 2018

Journal “Mathematics, Computer Science and Education”
Vol. 1, No. 1, 2018

Expert Knowledge Content Database Design Suggestions from Validation Aspect

Éva Hajnal

Abstract: *Expert Knowledge Content Database (EKCD) contains small or medium amount of data but its data content is valuable, which can be determined by expert work. Expert work can be a complex measurement, estimation, or data integration. Database design and validation was analyzed in a case study with EKCD Eulakes – Phytoplankton database of Lake Stechlin. Different data quality metrics was calculated and evaluated, and the content based inconsistency was evaluated. In the conclusion a schema is given for helping the design and validation.*

Keywords: *database validation, expert database, content type inconsistency.*

INTRODUCTION

The fundamentals of database design are well proved and documented methods, which are used during the development among others of a relational database. While these databases are the representatives of a traditional model, they are worldwide spread till nowadays. In contrary in connection with them there are some questions remained, which are worth to investigate. In this paper the Expert Knowledge Content Database (EKCD) design and validation questions are discussed mainly from automation aspect.

What does EKCD mean? The main feature of this database is not the size, but the value of the data content. It cannot be compared with large industrial or scientific databases, where the large amount of data and their handling itself means hard problem. These databases contain data, which were determined by expert work. Expert work can be a complex measurement, estimation, or the result of human data integration process. Data originate frequently in a horizontally and/or vertically wide collecting area as spatially as temporarily. Furthermore not only the data collection, but database design and maintenance require expertise not only in databases but the concerning field of the database content.

What is the importance of these databases? They are occurring in scientific research, when not the primer [5, 261–266] measured data are used. Especially there is a large amount of literature about medical databases, where the physician as an expert stores own determined data into a database [10, 4–14], [3, 25–28], or gaining subjective data from the client of the medical praxis [14, 1–13]. These types of databases became frequently used in the modern research, and health care when automatic data extraction and evaluation [1, 43-50], [2, 56–61], are executed and results are stored. These databases generally contain secondary data, and because they originate from some data integration work they are similar to the OLAP systems in aims and some features in spite of that they are formally OLTP systems.

The predecessors of this work were the design and development of the longtime phytoplankton database of lake Balaton (ALMOBAL) [6, 1–4], [8, 227–237], which stores datasets of a 110 year long period, the online database Peridat [7, 1–5], which stores water quality and perifiton data of small streams of Pannonia, and a high education database which stores integrated demographic, financial and employment data of a 40 years long period [5, 261–266], [12, 22–26].

In this paper, as a case study the longtime phytoplankton database of Lake Stechlin (Germany GPS 53.15, 13.028) named Eulakes is investigated and analyzed by design and validation point of view. The database is formally consistent and according to the professional design, the careful data input and input data checks its data content is also have to be excellent quality. However this database inherently must contain content based inconsistency if a careful validation is not executed. These inconsistencies come from the human data integration work and from the following digitalization process. These features have to be investigated, explored and their effects eliminated [14, 1–13], if it is possible. Further question is, that how accurate and reliable information can be queried, and from EKCD with thorough design and validation how relevant data can be found for scientific conclusions. The aim is to automate the content base validity check and maintenance and to minimize the human expert work.

Further question is the connection between the formal data quality metrics and the content based inconsistency. The hypothesis is, that the content type inconsistencies can be indicated by formal metrics, if the database contains enough redundant information and so the maintenance of this database can be almost totally automated.

First the database structure is shown and the designed redundancy is explained, after that the database verification and validation will be analyzed including the internal and external validation methods and a suggestion will be given for helping the design and development process from the aspect of the maintenance automation.

DATABASE DESCRIPTION

Eulakes database contains phytoplankton data. These are estimated by water samples investigated by microscopy when phytoplankton species are determined, their number is calculated and with help of an estimated volume their biomass is estimated as well. Its data have large importance in water quality assessment and in the ecological research. The basis of the biomass estimation is the individual’s volume of the given species, which may encompasses five order of magnitude.

In this case the database structure is not too complex (figure 1), but its data content is very valuable due to the expert time consumption, and due to the sometimes expensive and/or unrepeatable measurements.

The Eulakes database (Table 1) contains 618 water samples where the sample collection and processing needs minimum 8 expert hours per sample, totally ~5000 expert work hours. This small database is an appropriate object for this case study.

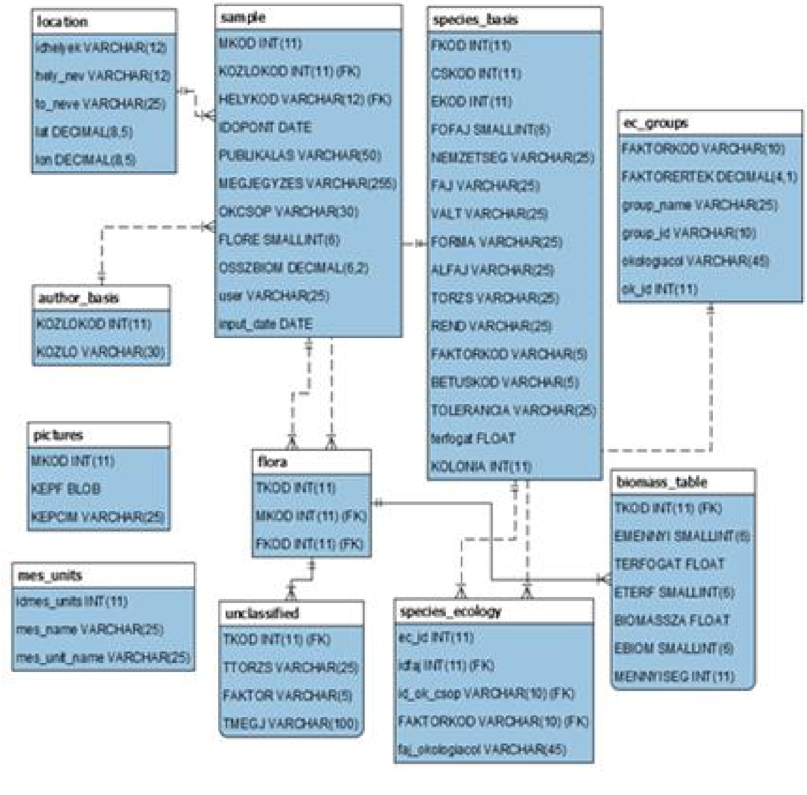

Fig. 1. *Database schema of Eulakes*

Table 1.
Data content

Year	Number of samples/year	Number of estimated biomass/year	Number of appearing new species/year
1994	48	1647	121
1995	43	1251	16
1996	37	1389	23
1997	34	1329	21
1998	36	1216	15
1999	39	1609	14
2000	43	1829	15
2001	47	1942	18
2002	18	724	9
2003	14	726	5
2004	16	600	2
2005	20	847	7
2006	22	914	11
2007	25	1026	9
2008	26	1023	3
2009	21	767	2
2010	23	753	7
2011	21	618	4
2012	1	40	1
2013	19	630	6
2014	22	842	5
2015	22	746	4
2016	21	713	2
Sum:	**618**	**23181**	**320**

The database contains data from 1994 till 2016 that means 23 years long data collection period. In some other cases not only the period is long but the number of data provider experts [10, 4–14] could also be large [5, 261–266], [12, 22–26]. Frequently the data input itself needs expert work, and additionally data provider and data processing experts are different. Sometimes the data collection and processing date is totally different (100 years difference in ALMOBAL [6, 1–4], [8, 227–237], [9, 149–150]) that means large difference in the expert knowledge, and methodology. This phenomenon itself causes inconsistency in the data content.

DESIGN

The reason of the database creation is that the traditional sheet based data storage is not maintainable above a critical data mass (Table 1), because the validation and the appropriate data queries become very hard. The database was developed by the considered traditional relational data model with a quite simple structure (Fig 1.), so it could be – as a model – a good example for answering the questions of this paper. It consists of 11 tables (Fig 1.) after normalization. The main expectation is to fit the Atomicity Consistency Isolation Durability (ACID) criteria. Accepting the modern CAP [11, 1–26] theorem (**C**onsistency, **A**vailability, **P**artition Tolerance) the main needs are at the CA side, so the cloud based NoSQL system usage is not suggested. The size and the way of usage also do not suggest the cloud based solution, and in the other side the author's grants, the database value and scientific usage raise safety and security aspects which suggest for the customer to remain at the traditional relational model based solution.

The disadvantage of the relational database schema usage is that the data input into the normalized data tables need laborious data preparation work, and it needs special software solutions. Furthermore the scientific work use generally specific queries, which are determined by the actual research question. The implementations of these queries cause difficulties in these databases. So other alternative data model implementation possibilities are ongoing questions.

Formal inconsistency can be avoided by careful relational database design and appropriate validation rules, but the previously described features induce inevitable content based inconsistency. The measurement and compensation of it will be possible with addition of supplementary information and extra redundancy to the database.

Some examples:

A. The species stored in hierarchical structure together with their taxonomical classification. This hierarchy is mapped to a relational schema which gives opportunity to reorganize data and somehow validate them with objective instrumental measurements.

B. The redundant species storage (species multiplication on synonyms) in species base table makes possible to follow the changes of taxonomy and the alternatives in depth of the species determination.

C. The storage of both volume of the species and the individuals of a sample makes possible the validation of volumetric data. There are primary and secondary data which coexist in EKCDs, so an extra qualifier data field also has to be stored.

VERIFICATION AND VALIDATION

A. Verification

First step is the data check, which can be executed by clearly explained and reported methods.

Larger problem is exploring content errors which are occurring during the data collection period, which frequently contains a preprocessing phase e.g. data digitalization from printed sheets. Their types, importance, handling methods are collected in Table 2.

These potential error sources should be collected during the development and the relevant error sources which error is at least partially checkable should be supplemented with supplementary information which can help us in the automation of maintenance. In Table 2 there are the evaluations of each error types with three values according to their importance, the chance of exploration and the chance of the improvement (*not* – cannot be repaired, *compensated* – with estimation the order of magnitude of the correct result can be calculated, *improvable* – the exact value is reachable).

The most important error is caused by incorrect administration of water samples (error in sampling date, place, and data record lost from sample details). By the consequence of the batch data input these serious errors can be traceable in plausibility metric of data quality, their effect can be compensated afterwards.

Table 2.

Error types

Error type	Importance	Exploration/ Improvement
Individual number estimation error	Not relevant	Not/ Not
Species identification error	Not relevant	Not/ Not
Species administration error (duplicate names, synonyms, changes in taxonomy with species merge or share)	Relevant	Checkable Partially compensated
Biomass estimation error caused by incorrect volume	Relevant, systematic error	Partially checkable Compensated
Sample administration error	Relevant	Checkable Compensated

The biomass estimated with systematic incorrect volume is a serious problem, but this effect can be compensated afterwards, if the volumetric data are corrected [15, 243–257]. Species identification error and measurement errors cause random aberrations, their effect is lost during the further statistical evaluations.

B. Internal Validation

The second step is the database validation. The formal exactness and data quality can be characterized by data quality metrics. There are several metrics for the internal validation of a database. In industrial databases and generally in transaction tracking systems (TTR) these validation process is well documented, furthermore there are special frameworks for support these processes. In this paper some data quality validation metrics – which were developed originally for reliability databases – were used. These metrics seem to be the best appropriate for this purpose.

Database was evaluated by the next data quality metrics: [4, 1–7] completeness, free-of-error, inconsistency, plausibility, richness of information. All metrics were calculated to all data fields, and the minimum values can be seen in Tables 3–4.

The completeness metrics reflects the percentage of properties which are not empty. This metric was generally 100% except three attributes (Table 4).

$$C=\frac{\sum_{i=0}^{n} c(x)_i}{n} \quad \left\{ \begin{matrix} c_i = 0, x_i = null \\ c_i = 1, x_i \neq null \end{matrix} \right\}, \tag{1}$$

C is the completeness, x_i is the i th data

Beside the almost total formal completeness we should estimate the functional completeness. Previous metrics can measure the formal completeness, but sometimes the database contains a lot of indirect, estimated information. The percentage of this information is the functional completeness. The direct or indirect feature of the data should be stored in the database.

The free-of-error metric is the percentage without of units or events that violate a particular rule. Types of it: logical, set membership and syntax errors. It can be easily avoided in this case, the usage of this metrics is not relevant. Errors should be avoided in EKCD with clear rules and correct data import processes.

$$F=\frac{\sum_{i=0}^{n} f(x)_i}{n} \quad \left\{ \begin{matrix} f_i = 0, x_i\text{: } violate\ rule \\ f_i = 1, x_i\text{:} not\ violate\ rule \end{matrix} \right\}, \tag{2}$$

F is the free-of-error, x_i is the ith data.

Inconsistency can count the number of elements which conform to a specific pattern, e.g. a predefined structure for serial numbers. Formal inconsistency is avoided in this database because of the data structure.

$$Y=\frac{\sum_{i=0}^{n} y(x)_i}{n} \quad \left\{ \begin{matrix} y_i = 0, x_i\text{:} conform\ to\ specific\ pattern \\ y_i = 1, x_i\text{: } does\ not\ conform\ to\ specific\ pattern \end{matrix} \right\}, \tag{3}$$

Y is the inconsistency, x_i is the ith data.

Much more interesting is the content based inconsistency (Table 4.), but this feature can be traceable in this database by the plausibility metrics via the redundant supplementary data.

$$P=1-\frac{\sum_{i=0}^{n} y(x)_i}{n} \quad \left\{ \begin{matrix} y_i = 0, x_i\ value\ is\ likely \\ y_i = 1, x_i\ value\ is\ unlikely \end{matrix} \right\}, \tag{4}$$

P is the plausibility, x_i is the ith data which has or has not unlikely value.

Plausibility is 100 minus the possibility of the identified unlikely values. It draws attention onto some errors. In this case study plausibility helped us to find a sample administration error, which was improved afterwards.

Table 3.
Data quality metrics and their relavance in EKCD

Data quality metrics	Usage	Data field	Value
Completeness	relevant	Biomass data	100 %
		Species volume	88.8 %
		Number of individuals	69.74%
Free-of-error	not relevant	all	100%
Inconsistency	not relevant	all	0%
Plausibility	relevant	sample date is not monotone order	99.83%
Richness of information	relevant	Unclassified individuals (etc. other record Cyanobacteria)	91.78%

Richness of information is a metric which would measure the percentage of units that are as accurate as the units of a property would suggest (The percentage of data without „other” category).

$$R=\frac{\sum_{i=0}^{n} y(x)_i}{n} \qquad \left\{\begin{matrix} y_i = 0, x_i\text{: accuracy is not appropriate to the measurement unit} \\ y_i = 1, x_i\text{: accuracy is appropriate to the measurement unit} \end{matrix}\right\}, \qquad (5)$$

R is the richness of information, x_i is the i^th^ data.

C. External validation

Data quality metrics have directed the attention onto the content based inconsistency, which can be explored if the database design makes it possible, or external validation methods can explore them, evaluate their effect, to compensation or repair.

Table 4 collects the content based inconsistency types, reasons and their importance in our case study Eulakes database.

Table 4.
Content based inconsistency types

Consistency type	Estimated importance	Improvement
Volume estimation	Remarkable error	Compensated afterwards
Changing microscope and other instruments	Minor error	Improvement not possible
Changing methodology	Medium importance error	Improvement not possible, but this effect can be estimated
Several expert, Different expert knowledge	Minor error	Effect can be estimated, compensated with statistical methods
Changing taxonomy	Medium importance error	Compensated afterwards

The most important problem is the volume estimation of the phytoplankton species, because its value can differ five order of magnitude. Additionally the volume is hardly estimated because its value is changing at each individual, and we can calculate with an average value. The volume measurement is not possible with microscope, because the measurements are executed only by two dimension and not by the third (depth) dimension. Lots of species have complex shapes which causes further difficulty. There are papers about volume estimations with help of plastic models [15, 243–257] or any other volumetric integration methods. If the database design makes it possible and we gain more exact volumetric data, the database can be recalculated more accurately afterwards.

The measurement methodology can cause minor and systematic error. These effects were investigated [13, 33–48]. The changes of the expert person and changes of the instrument also resulted in differences in species identification. Frequently it appears in richness of information in the aspect of identification.

It is a further plan to create metric for the quantification of this error, it means that the same item is identified as different species in different water samples. This error can be compensated by data loss. The general technics for avoiding these effects is to create statistical groups (functional groups, morphotypes and traits) for data evaluations or expanding the database with synonyms. The estimation of this effect can be executed by experimental work.

In consequences of the changes in taxonomy also appears as synonym taxon names in the database. This effect can be compensated. Sometimes due to this change two or more species are merged into one, or one species is divided. These effects can be compensated only by grouping methodology afterwards.

CONCLUSION

In this paper the Expert Knowledge Content Databases EKCD were investigated by the Eulakes database example. Database design was analyzed from the aspect of content type consistency maintenance and quality metrics were calculated. However according the careful design and input, the values of data quality metrics proved the hypothesized inherent content based inconsistency. The types of inconsistencies, their importance, compensations or repair were analyzed. It was supposed, that a content based inconsistency can be explored by not only external validation. It was shown that the completeness, plausibility and richness of information metrics can explore the content problems. If supplementary data were added to the database these problems can be noticed by these formal internal validation methods. We can conclude, that the main inconsistencies are recognizable and with the appropriate database design can avoid them. It is suggested, that during the database development phase the error possibilities have to be collected and analyzed. The importance and maintainability of each error have to be evaluated. Supplementary information should be searched and add to the database, and the validation rules have to be collected to the further automation of the maintenance.

ACKNOWLEDGMENTS

We acknowledge the financial support of this work by the Hungarian State and the European Community under the TÁMOP-4.2.2.B-15/1/KONV-2015-0010 project.

REFERENCES

[1] **Balan, Rege, 2017.** Balan, S., Rege, J. Mining for Social Media/ : Usage Patterns of Small Businesses. *Business Systems Research*. 8(1), 43-50. DOI: https://doi.org/10.1515/bsrj-2017-0004.

[2] **Cervenka, et al. 2016.** Cervenka, P. et al. Using cognitive systems in marketing analysis. *Economic Annals-XXI*. 160, 56–61. DOI: https://doi.org/10.21003/ea.V160-11.

[3] **Fine, et al. 2003.** Fine, L.G. et al. Information in practice: How to evaluate and improve the quality and credibility of an outcomes database: validation and feedback study on the UK Cardiac Surgery Experience. *BMJ*. 326 (January 2003), 25–28. DOI: https://doi.org/10.1136/bmj.326.7379.25.

[4] **Gitzel, et al. 2016.** Gitzel, R. et al. A Data Quality Metrics Hierarchy for Reliability Data. *The 9th IMA International Conference on Modelling in Industrial Maintenance and Reliability*.

[5] **Hajnal, et al. 2016.** Hajnal, E. et al. Intelligent estimation method for demographic effect onto the research and development sector in Hungary. In *Proceedings of the 11th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI 2016),* Timisoara, Romania, 261–266.

[6] **Hajnal, Padisák. 2006.** Hajnal, É., Padisák, J. Elaboration Phytoplankton Database (ALMOBAL) of Lake Balaton for Monitoring Water Quality. In *Proceedings of the Kandó Conference*, Budapest, Hungary, 1–4.

[7] **Hajnal, et al. 2011.** Hajnal, E.Software Model for the Abundance Distribution of Real Diatom Samples. In *Proceedings of the International Symposium on Applied Informatics and Related Areas*. Óbudai Egyetem.

[8] **Hajnal, Padisák.2008.** Hajnal, É., Padisák, J.Analysis of long-term ecological status of Lake Balaton based on the ALMOBAL phytoplankton database. *Hydrobiologia*. 599(1), 227–237. DOI: https://doi.org/10.1007/s10750-007-9207-x.

[9] **Hajnal, Padisák. 2006.** Hajnal, É., Padisák, J. Balatoni fitoplankton adatbázis (ALMOBAL) létrehozása és alkalmazhatósága vízminõségi monitorozásra. *HIDROLÓGIAI KÖZLÖNY*, 86(6), 149–150.

[10] **Herrett, et al. 2010.** E. Herrett, et al.Validation and validity of diagnoses in the General Practice Research Database/ : a systematic review. *British Journal of Clinical Pharmacology*, 69(1), 4–14. DOI:https://doi.org/10.1111/j.1365-2125.2009.03537.x.

[11] **Lourenço. 2015.** Lourenço, J. R. Choosing the right NoSQL database for the job: a quality attribute evaluation. *Journal of Big Data*. 2(18), 1–26. DOI:https://doi.org/10.1186/s40537-015-0025-0.

[12] **Nagy, et al. 2015.** Nagy, B. et al. The Use of Self-organizing Maps (SOM) in Demographic Analysis. In *Proceedings of the 10th International Symposium on Applied Informatics and Related Areas (AIS 2015)*, Óbudai Egyetem, 22–26.

[13] **Nőges, et al. 2010.** Nőges, P. et al.Analysis of changes over 44 years in the phytoplankton of Lake Vőrtsjärv (Estonia): the effect of nutrients, climate and the investigator on phytoplankton-based water quality indices. *HYDROBIOLOGIA*. 646, 33–48. DOI: https://doi.org/10.1007/s10750-010-0178-y.

[14] **Van Nooten, et al. 2018.** van Nooten, F. E. et al. Development and content validation of a patient-reported endometriosis pain daily diary. *Health and Quality of Life Outcomes*, 16(3), 1–13. DOI:https://doi.org/10.1186/s12955-017-0819-1.

[15] **Padisák, Soróczki-Pintér. 2003.** Padisák, J., Soróczki-Pintér, É. Sinking properties of some phytoplankton shapes and the relation of form resistance to morphological diversity of plankton – An experimental study. *HYDROBIOLOGIA*. 500 (January 2003), 243–257. DOI:https://doi.org/10.1023/A.