iEMSs 2008: International Congress on Environmental Modelling and Software
Integrating Sciences and Information Technology for Environmental Assessment and Decision Making
4th Biennial Meeting of iEMSs, http://www.iemss.org/iemss2008/index.php?n=Main.Proceedings
M. Sànchez-Marrè, J. Béjar, J. Comas, A. Rizzoli and G. Guariso (Eds.)
International Environmental Modelling and Software Society (iEMSs), 2008

Primary Environmental Data Quality Model: Proposal of a Prototype of Model Concept

M. Hejč[a] and J. Hřebíček[b]

[a] *Faculty of Informatics, Masaryk University, Botanická 68a, 602 00 Brno, Czech Republic (hejc@ecomanag.cz)*

[b] *Institute of Biostatistics and Analyses, Masaryk University, Kamenice 126/3, 625 00 Brno, Czech Republic (hrebicek@iba.muni.cz)*

Abstract: There is the short overview of the terms data quality and environmental information, following up the model definition in the paper. It introduces one such prototype of model – easy to implement, covering all modelling niches in environmental informatics and promising easy model knowledge sharing – it will be suitable to make a basis for appropriate model library. There is also discussed the use of the model and its role in the process of measurement of data quality. The concept is illustrated by the case study of the South Moravian region waste management data evaluation (realized by authors) and compared with the approach of the Ministry of Environment of the Czech Republic. Short conclusion suggests the future exploitation of possible new ways of dealing with the primary data uncertainty.

Keywords: Environmental Data; Data Quality; Data Uncertainty; Data Quality Model; Waste Management; Data Validation.

1. INTRODUCTION

Primary environmental data are monitored and collected by different ways: technically (e.g. using sensors, monitoring devices, people, etc) and by organizations Eurostat, European Environment Agency (EEA) and U.S. Environmental Protection Agency (EPA). The quality of the data is also variable (depending on many factors) or the data can be incomplete [Eurostat1, 2002]. These primary data are processed and they form required environmental information. If we want to determine the reliability of such information, its necessary to measure the quality of the primary data and even make changes to them (add, change or delete values with poor quality).

Measurement of the quality of the primary data can be made in various manners [Eurostat1, 2002], Hejč et al. [2007], and Pipino et al. [2002]. Very often it is not made or used at all. Sometimes it is (or it can be) judged by more or less experienced administration authority, but this evaluation process costs time and money or other resources. Therefore, it is not suitable for mass data analysis – this is the reason, why it's not often accomplished. Another reason lies in the lack of proper knowledge about the primary data monitoring and collecting system. It is necessary to use some techniques, which will be suitable for automatic computer processing.

This paper proposes such new technique – the new model for describing and managing environmental data quality. It will allow better results of waste management evaluation done by national, regional and local governments in the Czech Republic.

2. ENVIRONMENTAL INFORMATION

A short excursion into the field of the term "Environmental Information" gives us the EEA definition: "Knowledge communicated or received concerning any aspect of the ecosystem, the natural resources within it or, more generally, the external factors surrounding and affecting human life."

The definition of environmental information is very broad and includes these types of information: the ***state of elements*** of the environment – such air, water, soil, land, landscape and natural sites, flora and fauna, including cattle, crops, genetically modified organisms, wildlife and biological diversity – and it includes any interaction between them; the ***state of human health and safety***, conditions of human life, the food chain, cultural sites and built structures, which are, or likely to be affected by the state of the elements of the environment and the interaction between them; any ***factor*** such as substances, energy, noise, radiation or waste, including radioactive waste, emissions, discharges and other releases affecting or likely to affect the state of the elements of environment or any interaction between them; ***measures*** and activities affecting or likely to affect, or intended to protect the state of the elements of the environment and the interaction between them. This includes administrative measures, policies, legislation, plans, programs and environmental agreements; ***emissions***, discharges and other releases into the environment; ***cost benefit*** and other economic analysis used in environmental decision making [EC2003, 2003].

Pick [2007] distinguished only state of environmental elements, factors, measures and effects.

Another current research defines Single Information Space for the Environment in Europe (SISE) specified in the Work Programme for ICT research in FP7 for 2007 and 2008 as the common platform of all kinds of environmental information [Schouppe, 2008]. This is also the common research topic of the research group, in which are the authors of this paper [Nagy, Legat, and Hrebicek, 2007].

2.1 Non-environmental Data

During the process of environmental information evaluation there are often used some non-environmental information (e.g. subject addresses, names and other mostly personal, society or business data) [Eurostat1, 2002]. Actors and their descriptive data are also playing important role in the process of environmental information evaluation.

2.2 Uncertainties

Uncertainties in the scientific sense are the component of all aspects of the environmental modelling process. They describe lack of knowledge about models, their parameters, constants, data, information and beliefs [Jolma and Bortin, 2005].

Data (or information) quality is the measure of the data (or information), which measures uncertainties. The quality of data (or information) is high when the present uncertainties are low and vice-versa.

We will not cover any of these terms in to much detail in this paper, as the detailed description can be found in [Hejč and Hřebíček, 2006], Hřebíček et al. [2006] and [Olson, 2003].

3. DATA QUALITY MODEL

The primary environmental data values are often simulated by a model, but there is no well known and respected standard of the model. Most of experts use their own models and their own concept of the data quality model Hřebíček et al. [2006], [Olson, 2003], Pipino et al. [2002].

The new prepared standard ISO 25012 is an attempt to give the framework of primary environmental data quality model, but it is not still accepted by all ISO countries. The purpose of this standard is to prompt creators of large and small scale databases of primary environmental data to observe predefined criteria which will enable them to evaluate and test the quality of data, set up integrated and interoperable databases, reduce ambiguity, avoid redundancy, promote ease of data maintenance and promote reliable, secure databases. We have taken into account this standard in the development of our model.

3.1 New Definitions

We have to define some new terms and assumptions, which we will use in the supposed model concept implementation. These additional definitions and assumptions form the model concept. The *primary data set* (database) is the real input of the model. This data can be characterized by the structure of data model (e.g. ERD, etc.). This data model contains all the dependencies of the attributes and tables. One "piece" of the *data set*, (we will call it an *item*), is formed by the attribute name, its value and its relation to the rest of database.

It is very important to distinguish two kinds of model. The *data quality model* is the main model and (as its name implies) it is used to enumerate the data quality. The *data value model* (as illustrated in Hejč et al. [2007]) is the model, used to enumerate the value of the data, when the value is not known from some reasons.

The *enhanced primary data set* is the *data set*, which contains *tags* for every its *item*. The proposed model makes the use of two new *tags*: the *first tag* represents the *probability of item value to be true* and the *second tag* the *probability of the item value to be useful*. These are two different *tags*, because the value of the data can be false, but we know it's close to the true value and on the opposite hand we can have the true value which is from some reason completely useless for us.

The *item* can be accompanied by more *primary tags*, because it has more interesting characteristics from the standard ISO 25012 (accessibility, accuracy, currency, completeness, credibility, etc) and even other characteristics, not fully covered in ISO 25012. The proposed model uses them (and stores them if they are available) but not for the reason of the final evaluation of the data. This decision is motivated by practice. These *primary tags* are mapped into 2 above-mentioned new *tags*.

The most important characteristics of environmental data are: time and the field of environment (elements, factors, etc.) being described. It is very difficult to identify the most important *primary tags* as the process of mapping of these *primary tags* into *new tags* is different for each evaluation procedure. Generally, the *anomaly* can be used, when the data are right, but from some reason they are useless (e.g. the production of waste in municipality affected by extraordinary flood is not suitable for the computation of the average waste production in municipalities). The standard ISO 25012 credibility and also the difference from *data value model* can be used as main *primary tags* to evaluate the data with high influence of uncertainty, which has been raised from the human factor (or similar factors, e.g. complete fail of measuring device, etc.), Hejč et al. [2007].

Table 1. Data quality model.

Enhanced Primary Data Set (for the given purpose)						
Primary data	Tags		Primary tags			DVM
	P(TRUE)	P(USEFUL)	*Anomaly*	*Credibility*	...	
Item 1 (key, attr., val.)	t1	u1	a1	c1	…	x1
Item 2 (key, attr., val.)	t2	u2	a2	c2	…	x2
Item 3 (key, attr., val.)	…	…	…	…	…	…
…	…	…	…	…	…	…

The whole concept and new terms are illustrated by the Table 1, where every item consists from the key (the relation to the rest of data in the given data model), the attribute name and the attribute value.

The tag P(TRUE) means the probability of item to be true, P(USEFUL) the probability of item to be useful and DVM means the data value model in Table 1. There are illustrated only few primary tags, but it is possible to present more of them.

The purpose of the enhanced primary data set is given by the evaluated case and reflected in different values of tags P(USEFUL).

3.2 Mapping

Mapping of *primary tags* into two *new tags* during the evaluation procedure is the key part of the *data quality model*.

We suppose to react only on some kinds of data uncertainty (as mentioned above). There are many sources of uncertainty, including: uncertainty in scientific constants, observation error, implementation uncertainty, etc., see Hřebíček et al [2006], but we suppose they can be solved separately by other models or by EEA, Eurostat or EPA procedures and they can be later incorporated into new model (or vice-versa).

Our model is defined as a function of several parameters. Often a very complex function (with a lot of exclusions), but not always – sometimes can be simple. Function value represents *new tag's* value. Input parameters of the function include various knowledge about the item (*primary tags*), the value of the item itself and the value of the *data value model*. Different (in the sense of *primary tags* used) *data quality models* can be easily combined as functions do the same.

When we get the data, we have to fill, look for or compute the values of all *primary tags*. It can be done very simply (by setting some default values) or by application of some rules (e.g. all the data from some sources are more suspicious of being wrong – that means setting their credibility lower then the others). Application of rules may be cumulative and this implies the need for some arithmetic to compute the final value of the *primary tags*.

The last application of the rules would be the comparison with the *data value model*. If the value of the item is not far from the value suggested by the *data value model*, the probability of the item value to be true is high, similar rules apply for usefulness.

Finally the mapping of *primary tags* into *new tags* (probability and usefulness) is done for all items and for given purpose (type of evaluation of the data). This is the new approach.

In rare cases we can employ some optimization function which recognizes the information quality by some independent (this means different than the application of the above mentioned *data quality model*) method. This gives us the possibility of feedback for the correctness of the *data quality model* (and thus possibility of automatic model shaping through mapping changes). The only way to demonstrate the correctness of the *data quality model* is in the other cases the independent study, made by some expert in the field.

3.3 Data Changes

We will get the data with some new attributes and we can decide what to do further – we can define some rules. Either we can replace item values with low probability by the *data value model* values or we can exploit some values with high usefulness.

Other possibilities lie in the filling of the gaps of data. First we have to identify them (by the *data value model*) and then we need only to deliver appropriate *data value model* values into data set. Again the set of rules would be useful in the process of data quality evaluation.

Finally we have the data set with some quality evaluation and enhancement and we can use it for information retrieval. In the same time we are aware of the information quality, as it is tightly bound to the data quality evaluated before.

4. USE OF THE MODEL

The use of new *data quality model* has been tested during annual evaluation of waste management indicators in the South Moravia Region since 2004. The evaluation of waste management indicators in the Czech Republic is usually done by the different approach of the Ministry of Environment (MoE), Hejč et al. [2007]. So-called null-variant approach is used in some other regions than South Moravian. This null-variant means simply the direct evaluation of primary data without any pre-processing treatment. Other types of approaches (hypothetical full variant by EPA and the compromise approach of statistic offices), which are described in [Eurostat1, 2002], Hejč et al. [2007], will not be compared with the above approaches.

4.1 Comparison

The differences between MoE and our approach are shown by the specific example. We choose as the example an evaluation of the household waste production at municipalities. In Visegrad countries, the household waste is collected and separated into waste containers depending on the collection system of the given municipality. The amount of the household waste production and disposal of the given municipality is announced / reported in compliance with the national legislation of the Czech Republic to MoE through local state administration bodies and the Centre of Waste Management (http://ceho.vuv.cz/). All available annual reports are evaluated and the overall production of municipal household waste is aggregated into the final environmental reports of the Czech Republic to EEA and Eurostat.

There is the common part for both approaches – the primary data about the municipal household waste production are collected and evaluated. However, there are differences in the types of data collection and their processing. When the null-variant comes into play, annual reports of municipalities are just collected and the plain summary is processed and evaluated. Sometimes some most flashy cases of errors are filtered (by means of interval arithmetic). The approach of MoE is closer to null-variant than to any of the others. Interval arithmetic is the only strong tool. A lot of knowledge is not used in the evaluation process.

But we know more about the nature of these data. The municipal household waste production strongly depends on the number of inhabitants and the standard of living. Then there are some other dependencies on the size of the community, the type of housing, unemployment rate, time series of waste production, etc. All these dependencies can be incorporated into the *data value model* of these primary data. Such model forms the knowledge and can be used for the verification of the data or to replace the gaps of the data (as statistics approach does).

We describe formally a simple model of waste production as the function of appropriate variables and bellow is presented as an example of *data value model*:

$$P = F(\#inh, spec, std, sz, unemp, hsg, heat),$$

where are defined: P is the amount of the waste production per year; *#inh* is the number of inhabitants; *spec* is the specific waste production coefficient (reference values of other coefficients), measured in kg; *std* is the standard of living coefficient; *sz* is the size of the community coefficient; *unemp* is the unemployment rate coefficient; *hsg* is the type of housing (recreation, blocks of flats, empty houses…) coefficient and *heat* is the type of heating coefficient.

In this case the function *F(#inh, spec, std, sz, unemp, hsg, heat)* can be defined, see Fig. 1, and we can write:

$$P = \#inh \cdot spec \cdot std \cdot sz \cdot unemp \cdot hsg \cdot heat / 1000 \text{ [t]}.$$

Further, we can compute some coefficients x of function F, x belongs to {*#inh, spec, std, sz, unemp, hsg, heat*}, by the expression:

$$x = (act / ref)^{cx},$$

where *ref* means a reference value; *act* an actual value and *cx* is the compensator (given by optimization process) of the considered coefficient x.

The model of the standard of living value (as one example of numerous sub-models) is used to compute the actual and the reference value of the considered coefficient:

$$stdV = Rinc \cdot Rsz,$$

where *stdV* is the standard of living value; *Rinc* the average income in the given region and *Rsz* the size of the community in region coefficient.

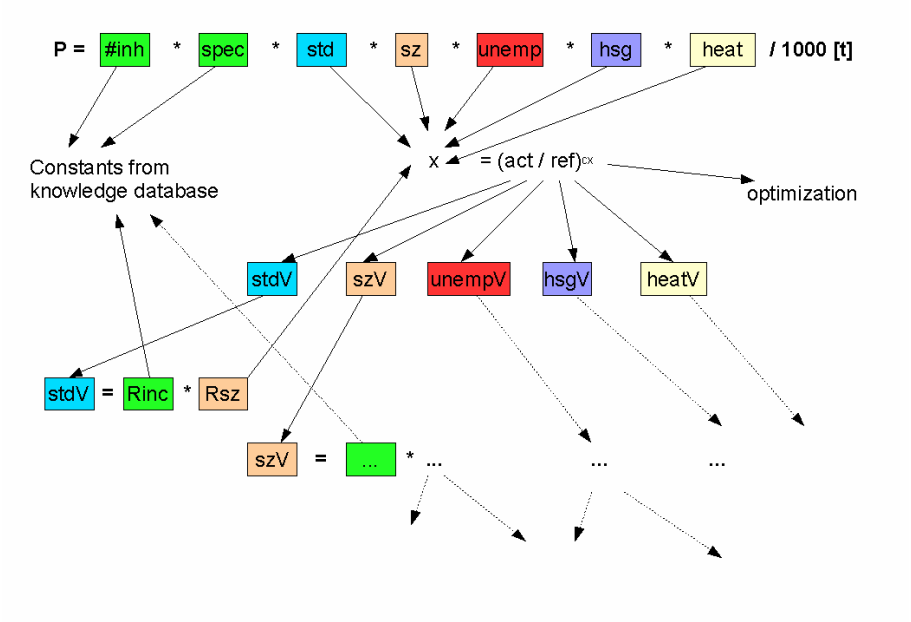

Figure 1. Model of waste production in communities (Hejč and Hřebíček, 2007).

Different colours of cells of Figure 1 are used to distinguish main areas of interest – number of inhabitants and specific waste production (green), standard of living (light blue), size of the municipality (orange), unemployment (red), housing (purple) and heating (yellow).

4.2 Details of Use

All data are enhanced by all available *primary tags*. For example: the data source district credibility; the data source subject credibility (both taken from time series); the value from *data value model*; cross-reference (by reason that subject and its partner are always reported), etc.

Other (non-environmental) data are collected (mostly for the purpose of the *data value model*) and they are enhanced by appropriate *primary tags*. For example demographic data, addresses, economic data, etc.

The appropriate mapping of *primary tags* on 2 *new tags* (see above) is used (assigned) for each type of evaluation. This mapping is fine-tuned by an optimization process in some

rare cases. The example of possible case is the overall waste production when the comparison of waste production volume and waste treatment volume is available.

The second phase of data processing uses only the primary data (without *primary tags*) and 2 *new tags* mapped by appropriate mapping which conforms to the purpose of evaluation (a specific indicator). This approach makes possible to attain better results in the process of final evaluation of the indicator. It also gives possible records (warning) for stakeholders – they are warned of problems in the data by easily understandable way.

4.3 Practical Experiences

The practical experiences with the use of the *data quality model* have been already very promising. Authors used the presented concept of the *data quality model* for the evaluation of the waste production data of the South Moravian Region in the Czech Republic. It was used in the years 2004, 2005 and 2006. Some basic experiences have been obtain also from the evaluations of the waste production data in the years 1999, 2000, 2001 and 2002, but these evaluations have not used the later proposed the *data quality model.* However, the experiences from these first years have been very useful to develop it. Table 2 illustrates some interesting statistics from the processes of evaluation of the waste production data in the years 2004, 2005 and 2006. The data of the year 2007 will be evaluated in the summer of the year 2008.

Table 2. Statistics from the processes of evaluation of the waste production data.

	Database			**Errors**				
Year	Items	Plants	Subjects	found	suspected	Hit rate	estimated	Hit rate
2004	145 068	22 428	16 783	75	228	33%	1015	7%
2005	166 501	28 815	21 413	63	130	48%	749	8%
2006	176 676	31 439	22 551	44	429	10%	530	8%

It is clear from Table 2, that the amount of data in databases is growing and the number of estimated error is decreasing. The lower hit rate of the number of found errors vs. the number of suspected errors in 2006 is due the short time for the confirmation of suspects, while the same hit rate in 2005 is higher due the short time for the preparation of suspects (with the strategy of finding only flashy ones). The 2007 year promises a good increase of hit rate of found errors vs. estimated errors, because there will be devoted more time by the local government for the whole evaluation process of the waste production data.

5. CONCLUSIONS

We have presented short overview of the terms in data quality area and we also enhanced their capabilities by defining some new ones. We defined the new model of data quality and introduced it on the example of the waste production data of the South Moravian region of the Czech Republic. The main advantage of the new model prototype lies in the representation of the data and easy implementation and sharing of the modelling results. The further property of the model is its ability to locate the data uncertainty when any other ways of uncertainty measurement are not present [Hejč and Hřebíček, 2006]. Future research will be trend towards refining of the model and also incorporating it in the framework of broader research interest of authors – environmental information space [Schouppe, 2008].

ACKNOWLEDGEMENTS

The authors wish to thank the Ministry of Education of the Czech Republic for the financial support of the project No MSM 0021622412 INCHEMBIOL and the Ministry of

Environment of the Czech Republic for the financial support of the project No SPII2/f1/30/07.

REFERENCES

EC2003, Directive 2003/4/EC of the European Parliament and of the Council of 28 January 2003 on public access to environmental information. Official Journal of the European Union. L 41/26. 2003.

EEA, European Environment Agency, <http://www.eea.europa.eu/>, 2008/04/30

EPA, U.S. Environmental Protection Agency, http://www.epa.gov/>, 2008/04/30

Eurostat, <http://ec.europa.eu/eurostat>, 2008/04/30

Eurostat1, Quality in the European statistical system – the way forward. 2002 Edition. Office for Official Publications of the European Communities, Luxembourg, 2002.

Hejč, M., and Hřebíček, J., Solving Waste Management Data Uncertainties. Case Study of South Moravian Region. 20th International Conference on Informatics for Environmental Protection. Managing Environmental Knowledge. Shaker Verlag, Aachen, 405-409, 2006.

Hejč, M., Hlaváček, M., and Hřebíček, J., eGovernment Services in Environment - Automate Data Quality Assessment - Czech Republic Approach, In EnviroInfo 2007. 21st International Conference on Informatics for Environmental Protection. Environmental Informatics and System Research. Volume 2. Workshop and application papers. Warsaw, Poland: Shaker Verlag, Aachen, 159-166, 2007.

Hřebíček, J., Hejč, M., and Holoubek, I., Current Trends in Environmental Modelling with Uncertainty, Proceedings of the iEMSs Third Biennial Meeting "Summit on Environmental Modelling&Software". International Environmental Modelling and Software Society, Burlington, 342-347, 2006.

Jolma A., and J., Bortin, Methods of uncertainty treatment in environmental models, *Environmental Modelling & Software*, 20(8), 979-980, 2005.

Nagy, M., Legat, R., and Hrebicek, J., Electronic Access to Environmental Information – an Important Fundament for E-Democracy and Environmental Protection, CAHDE (2007) 23 E. <http://www.coe.int/t/e/integrated_projects/democracy/02_activities/002_e-democracy/>, 2008/04/30

Olson, J. E., Data Quality: the accuracy dimension. Morgan Kaufmann Publishers, San Francisco. 2003.

Pick, T., From Aarhus to INSPIRE: Putting Environmental Information on the Map. In EnviroInfo 2007. 21st International Conference on Informatics for Environmental Protection. Environmental Informatics and System Research. Volume 2. Workshop and application papers. Warsaw, Poland: Shaker Verlag, 239-246, 2007.

Pipino, L., Lee, Y., and Wang, R., Data Quality Assessment. *Communications of the ACM,* 45(4), 2002.