ISSN 1453 – 7303 **“HIDRAULICA” (No. 2/2019)**
Magazine *of Hydraulics, Pneumatics, Tribology, Ecology, Sensorics, Mechatronics*

From Human-Environment Interaction to Environmental Informatics (IV): Filling the Environmental Science gaps with Big Open-Access Data

Assoc. Prof. eng. **Mirela COMAN**[1], PhD stud. **Bogdan CIORUȚA**[2]

[1] Technical University of Cluj-Napoca - North University Centre of Baia Mare, Faculty of Engineering, str. Victor Babeș 62A, 430083, Baia Mare; comanmirela2000@yahoo.com

[2] Technical University of Cluj-Napoca - North University Centre of Baia Mare, Office of Information and Communication, str. Victor Babeș 62A, 430083, Baia Mare; bogdan.cioruta@staff.utcluj.ro

Abstract: *Nowadays, environmental data lifecycle management becomes more and more important in the context of local and regional communities sustainable development strategies. Enormous volume of environmental data is generated by an increasing number of devices across the world; correct and efficient surveillance of environment throughout data lifecycle management is essential for optimisation of the utility of those data pieces and minimisation of error potential.*

Starting from the idea of sustainable development in accordance to the latest sustainability diagram framework, using the tools provided by the Environmental Informatics Systems, we propose to show that, on the Human-Environment Interaction - Environment Informatics circuit, whatever the environment offers as primary data and what we offer back to the environment needs to represent the real state in the field, with which the environment copes. On the validity and viability of the respective situation depends, after all, our common future, based on harmony between society-economy-environment.

Keywords: *Human-Environment Interaction, environmental big data, data acquisition, Knowledge Society.*

1. Introduction

Nowadays, the era that we live can be described as the “Information Age”, sometimes even as the “Knowledge Age”. No matter what area of science and technology we look at, it is obvious that we are dealing with an ‘information overflow’ without precedent in the history of mankind [1, 2].

In this context, *Environmental Sciences* are no exception and recent advances in this field, also considering environmental monitoring and protection activities, would have been unthinkable, unmanageable and unattainable without the support offered by modern information technology [1], in the sense of *Environment Information Systems* (EISs) and *Environment Informatics* (EI) [4, 5], as part of the *Sustainable Informatics* (SI).

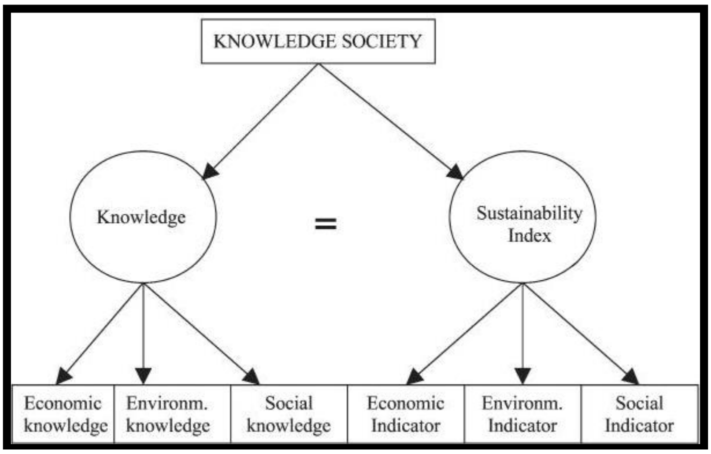


Fig. 1. Knowledge Society as a mixture between knowledge and sustainability index [2]

In its turn, Sustainable Informatics (SI) via *sustainability index* is integrated in Knowledge Society (as in Fig. 1) [3], where it play an important role, for environmental data-dependent actors, in decision-making [4, 5], problem solving, analyzing trends [6-8], understanding their customers, and doing research, being closely linked with environmental requirements in decades [9, 10].

2. From Environment Surveillance and Monitoring to Environmental Big Data evolution

More often than ever, we are talking about the "information explosion" or "information abundance", suggesting in fact through these expressions the existence of a huge data and/or information volume and, at the same time, the difficulty in accessing, selecting and evaluating the information resources [11]. In this sense, it is appreciated that the feeling of being overwhelmed by data and/or information is due to the inability to extract from a huge volume of data - big data (see a concept representation in Fig. 2) or information resources the really relevant elements.

Fig. 2. A mind-map example for the Big Data concept

The great problem of information remains the vast amount of data resources that are unaware of limitations under any form of society's activity. In order to have an image of the informational dimension of society as a whole, we must not overlook the fact that "flooded with techniques, some scientists complain that instead of experimenting, they lose more time trying to see whether was not done before someone else" [11]. In such context, it is extremely difficult to identify information items, especially those that are really important.

2.1 Environment surveillance and monitoring

Likewise, environment observations, as primitive forms of information about a change, tendency or development in industrial, economical, national or global environments, are key pieces of data, which together contribute to building and improving the understanding of interactive processes in the Earth's environment or reality and developing a more coherent image (as suggested in Fig. 3) of its sensitivities towards natural and human-induced changes.

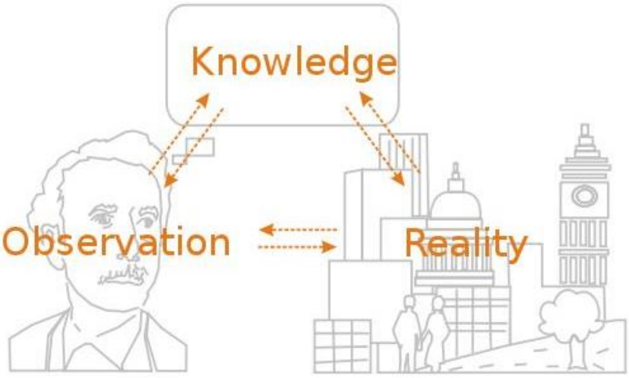


Fig. 3. The reality-observation-knowledge triangle as a primitive representation for environment surveillance

Modern environment surveillance and monitoring systems in collaboration with policies, strategies and governmental actions - transposed to national, regional and local levels, provide evidence regarding causes, tendencies and consequences of ecosystem changes on different scales.

As reflected by Fig. 4, there are four steps in connection with modern environment surveillance and monitoring, as follows: *data acquisition and aggregation* (step 1), *data processing, machine learning and pattern recognition* (step 2), *geospatial analytics and insights* (step 3), and, last but not least, *online decision support and data science* (step 4).

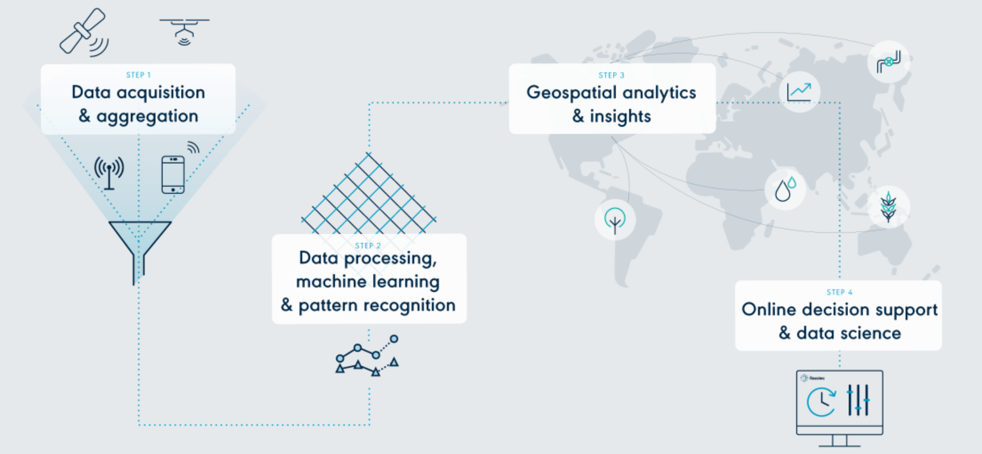


Fig. 4. Different aspects of the environment surveillance and monitoring

Environmental observations, since "the environment" gained its place in the public international agenda, have already been used to reduce the impact of natural hazards, which can have a really significant cost per year, and have the potential to also improve other areas like business analytics, agricultural production and urban planning, the sustainable development of the community [12-14].

In order to support the green economy, with highly accurate environmental data, by creating flexible long-term environmental surveillance systems, is more than just natural and necessary to consider the SI exploring approaches to measurement, analysis, up- and down-scaling and modeling of whole ecosystems. This aspect will guarantee that we, as a community with responsibility towards the environment, can easily detect, and interpret change and help decision-makers assess policy and resource management options by quantifying environmental and socio-economic impacts, via systematic and unsystematic environmental observations (see Table 1).

Table 1. Types of systematic and unsystematic environmental observations (adapted after [15-18])

Systematic environmental observations	**vs.**	**Unsystematic environmental observations**	**Scientific or unscientific methods of environmental observation**
direct	vs.	indirect	measurement taken in relation to behavior measured
global	vs.	specific	variety of behavior observed by researcher
noticed	vs.	unnoticed	participant awareness
obtrusive	vs.	non-obtrusive	influence on participants or environment
participant	vs.	non-participant	involvement with participants
reactive	vs.	non-reactive	participant reaction
structured observation	vs.	unstructured observation	with or without observation checklists
qualitative	vs.	quantitative	-
naturalistic	vs.	laboratory	-
human	vs.	non-human	human (classic) observations will be replace with non-human (innovative) observations of environment quality

According to the Table 1, an environmental observation can be sometimes *casual* in nature or sometimes it may act *scientifically*; an observation with a casual approach involves observing the right thing at the right place and also at the right time by a matter of chance or by luck, whereas a scientific observation involves the use of the tools of the measurement. A very important point to be kept in mind here is that all the environmental observations are not scientific in nature.

Natural observation involves observing the behavior in a normal setting, no efforts are made to bring any type of change in the behavior of the observed; improvement in the collection of information and improvement in the environment of making an observation can be done with the help of *non-natural observations* (laboratory, by excellence). With the help of the *direct method of observation*, the observer is physically present; *indirect method of observation* involves studies of the different data recordings or information formats about the state of the environment, as in Fig. 5.

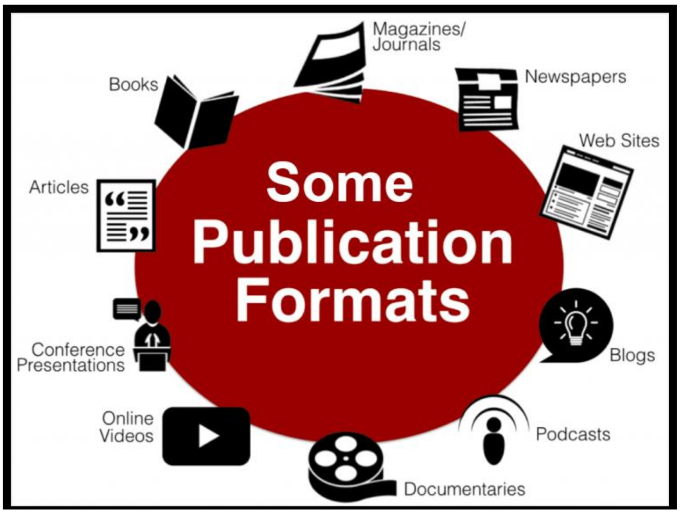


Fig. 5. Different publication formats for environmental data or information

At a closer look to common formats for environmental data and/or information, we could mention that the Knowledge Society manifest significant interest in:

- *books* - usually a substantial amount of information, published at one time and requiring great effort on the part of the author and a publisher;
- *magazines* or *journals* - published frequently, containing lots of articles related to some general or specific professional research interest;
- *newspapers* - each is a daily publication of events of social, political and lifestyle interest;
- *web sites* - digital items, each consisting of multiple pages produced by someone with technical skills or the ability to pay someone with technical skills;
- *articles* - distinct, short, written pieces that might contain photos and are generally timely;
- *conference papers* - written form of papers delivered at a professional or research-related conference. Authors are generally practicing professionals or scholars in the field;
- *blogs* - frequently updated websites that do not necessarily require extensive technical skills and can be published by virtually anyone for no cost to themselves other than the time they devote to content creation. Usually marked by postings that indicate the date when each was written, and *documentaries* - works, such as a film or television program, presenting political, social, or historical subject matter in a factual and informative manner and often consisting of actual news films or interviews accompanied by narration;
- *online videos* - short videos produced by anybody, with a lot of money or a little money, about anything for the world to see, and *podcasts* - digital audio files, produced by anyone and about anything, that are available for downloading, often by subscription.

Structured environmental observation works according to a plan, involves specific information of the units that are to be observed and also about the data that is to be recorded - the operations that are to be observed and the various features that are to be noted or recorded are decided well in advance; in the case of the *unstructured environmental observation*, its basics are diametrically against the structured environmental observation - the observer has the freedom to note down what he/she feels is correct and relevant to the point of study and also this approach of observation is very suitable in the case of exploratory research.

Controlled environmental observations are the observations made under the influence of some of the external forces and such observations rarely lead to improvement in the precision of the research results. *Non-controlled environmental observations* are made in the natural environment, and reverse to the controlled observation these observations involve no influence or guidance of any type of external force. Observations made while using the scientific method can be *quantitative* or *qualitative*; observations are quantitative if they return numerical data, and qualitative research is when observations are recorded without capturing numeric data values - this type of observation is more subjective and relies on the researcher's interpretations.

2.2 Defining the Environmental Big Data

Big Data was the buzz phrase of the recent years, but in truth, the concept has been around far longer than that; we know what data is - it is the raw information collected from any study, but particularly in sciences, so we expose some dictionary definitions, as follows [19, 20]:

- environmental data means any measurements or information that describe environmental processes, location or conditions, ecological or health effects and consequences, or the performance of environmental technology;
- environmental data include information collected directly from measurements, produced from models and compiled from other sources such as databases or literature;
- environmental data means any parameters or pieces of information collected or produced from measurements, analyses, or models of environmental processes, conditions, and effects of pollutants on human health and the environment, including results from laboratory analyses or from experimental systems representing such processes and conditions.

Environmental Big Data takes environmental data concept one step further - it is an environmental data set of such complexity that it would be impossible to select, preprocess, examine, manipulate, present and evaluate, as in Fig. 6, using traditional and dedicated methods.

The intended results are often so complex that it's difficult to process even using tried and tested methods. It's important to note that the term does not necessarily denote the size of the data set (although a large volume of data is unavoidable), merely its complexity.

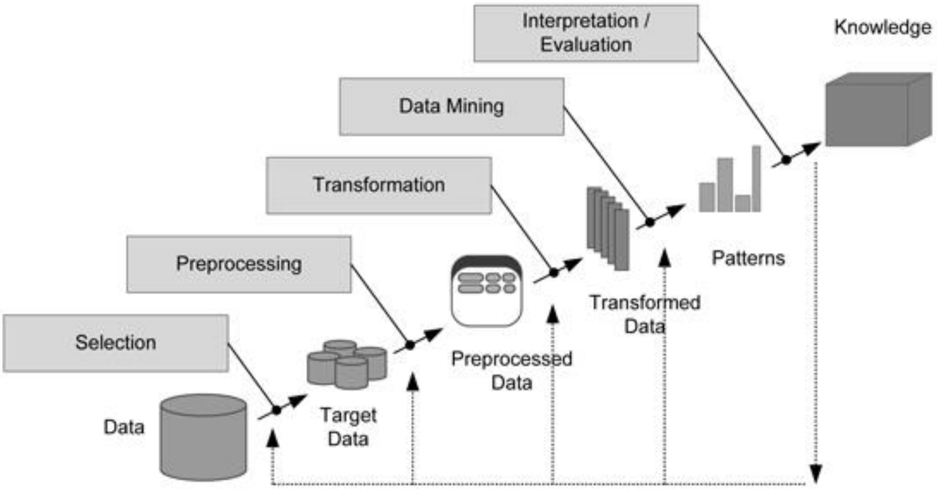


Fig. 6. An example of environmental data transformation process

Environmental Big Data is a term that describes the large volume of data - both structured and unstructured - that inundates a domain of interest or area of activity on a day-to-day basis, but it's not the amount of data that's important, it's what community do with the data that matters.

Environmental Big Data can be analyzed for insights that lead to better decisions and strategic eco-friendly moves in the context of a social-ecological system (SES), from 3 Vs models up to 15 Vs models, having the main 10 characteristics exposed in Fig. 7.

In actual context, the environmental data lifecycle is the sequence of stages that a particular unit of data goes through from its initial generation (or acquisition) to its eventual archival and/or deletion at the end of its useful life, or information dissemination [17].

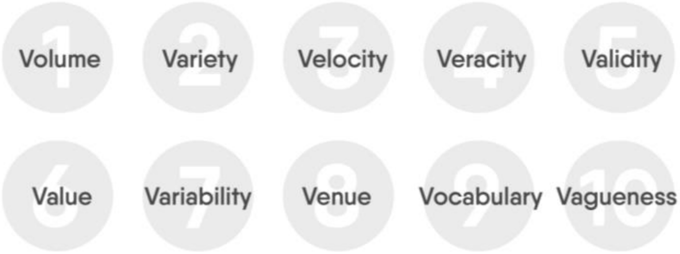


Fig. 7. The 10 main characteristics of Environmental Big Open-Access Data Vs models

Although specifics vary, data management experts often identify five or more stages in the data life cycle. Once we know what data we are looking for, we should follow the next steps [17]:

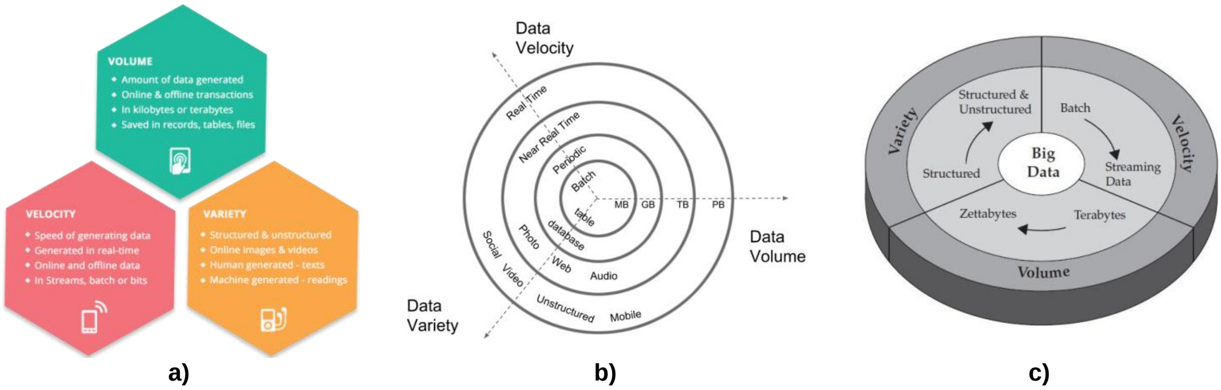
- *data acquisition* refers to data collection from different sources;
- *data pre-processing* refers to data transformation from streams of bytes into the proper formats that will have a meaning to Big Data tools and technologies;
- *data clean up* is a critical step, though in many occasions it is forgotten or carried out leniently, the gross data acquired in the first phase may include bias, may have Data Deserts and/or may be subjected to Data Mirage effects, but during data clean up, we create a new data set that represents a better image of we are interested in analyzing;
- *data analysis* is the following step. Once we have the data in the appropriate format and structure and cleaned up, we start asking questions such as: Do we see any trend? Is there any kind of consolidated information this data shows? Can we infer some patterns?
- *data visualization*, sometimes carried out in parallel with data visualization, there are a great number and variety of data visualization tools.
- *data interpretation*. Together with the Data Clean up step in the Big Data processing cycle, this is also a critical step. There are many instances in which Data Interpretation has not been carried out adequately, leading to the wrong conclusions.
- *data intervention*. Once the conclusions have been achieved, this will lead to the next step which may involve using the data to take decisions, train Machine Learning algorithms, or rethinking the data gathering for the future.

In the scientific literature there are a few Vs models for Environmental Big Open-Access Data that integrate the following aspects:

- 3 Vs of Environmental Big Open-Access Data (*volume* - data galore, is self-explanatory; *variety* - complexity, types and formats; and *velocity* - actual speed of data);
- 4 Vs of Environmental Big Open-Access Data (3 Vs + *veracity* - uncertain or imprecise data; or 3 Vs + *value* - characterizing the potential of data to transform plans in actions);
- 5 Vs of Environmental Big Open-Access Data (4 Vs + *visualization* or 4 Vs + *value*);
- 6 Vs of Environmental Big Open-Access Data (5 Vs + *variability* - this refers to dynamic, evolving data, time series, seasonal, and any other type of non-static behavior in data sources; or 5 Vs + *validity* - clean data);
- 7 Vs of Environmental Big Open-Access Data (*volume*, *variety*, *velocity*, *veracity*, *value*, *visualization*, *variability*);
- 8 Vs of Environmental Big Open-Access Data (*volume*, *variety*, *velocity*, *veracity*, *value*, *visualization*, *viscosity*, *virality*);
- 9 Vs of Environmental Big Open-Access Data (8 Vs + *venue* - where the data comes from, multiple platforms; or 8 Vs + *vocabulary* - semantics and other context-based metadata that describe the data‘s structure, syntax, content and provenance);
- 10 Vs of Environmental Big Open-Access Data (9 Vs + *vagueness* - confusion over the meaning of big data);
- 11 Vs of Environmental Big Open-Access Data (10 Vs + *volatility*);

- 12 ... 14 Vs of Environmental Big Open-Access Data (different combinations of Vs);
- 15 Vs of Environmental Big Open-Access Data (*volume*, *variety*, *velocity*, *veracity*, *value*, *variability*, *viability*, *visualization*, *virality*, *viscosity*, *volatility*, *validity*, *vocabulary*, *venue* and *vagueness*).

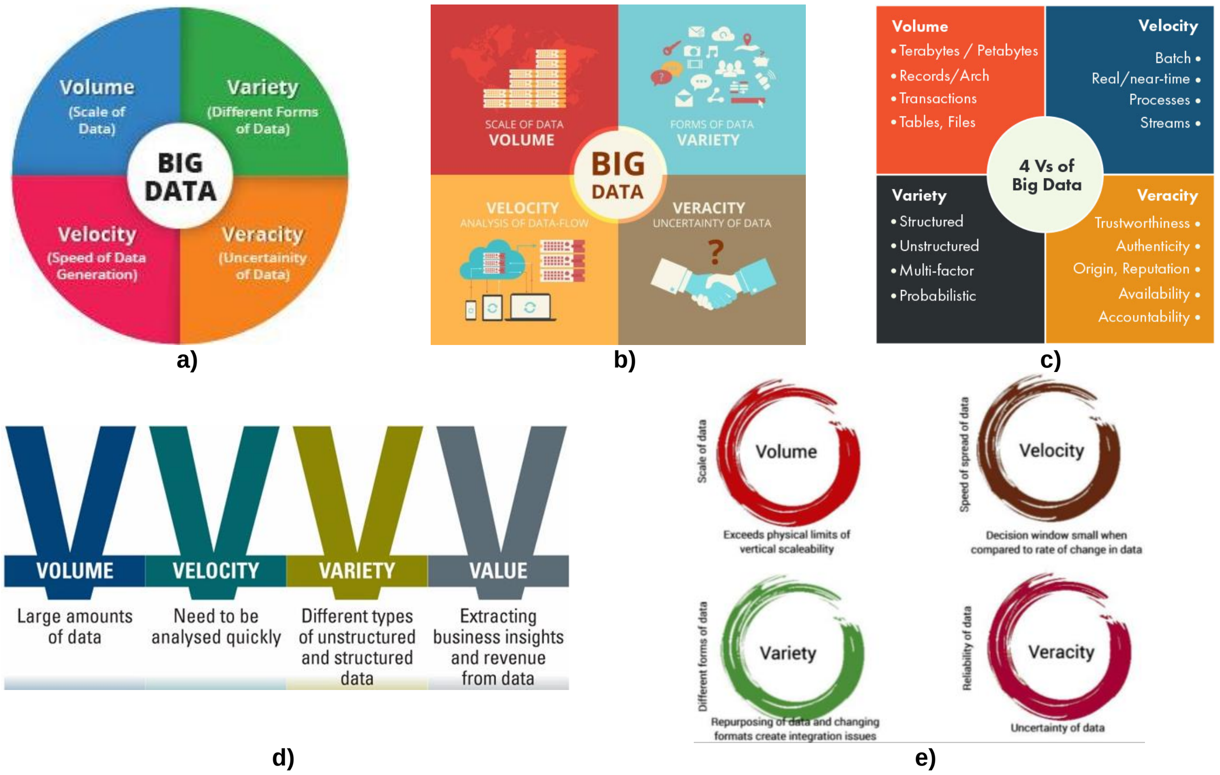
Environmental Big Open-Access Data model/concept-diagram is determined using a few metrics, closely related to the number of parameters, starting with the simplest approach - 3 Vs (as presented in Fig. 8) and ending with the most recent - with no less than 15 parameters.



a) b) c)

Fig. 8. Examples of the 3 Vs Environmental Big Open-Access Data models
a) detailed hexagonal model; b) detailed target distributed model; c) detailed segmented cycle model.

In the same context, to be relevant, Environmental Big Open-Access Data must be able to cope with the speed at which data is generated in order to store it and retain the most up-to-date and relevant information. This is useful in most areas related to environment protection, but vital in early warning systems ahead of natural disasters or to detect inappropriate activities which interferes with the environmental protection. The fourth dimension of Environmental Big Open-Access Data is veracity - handling data in doubt (see a few 4 Vs models in Fig. 9).

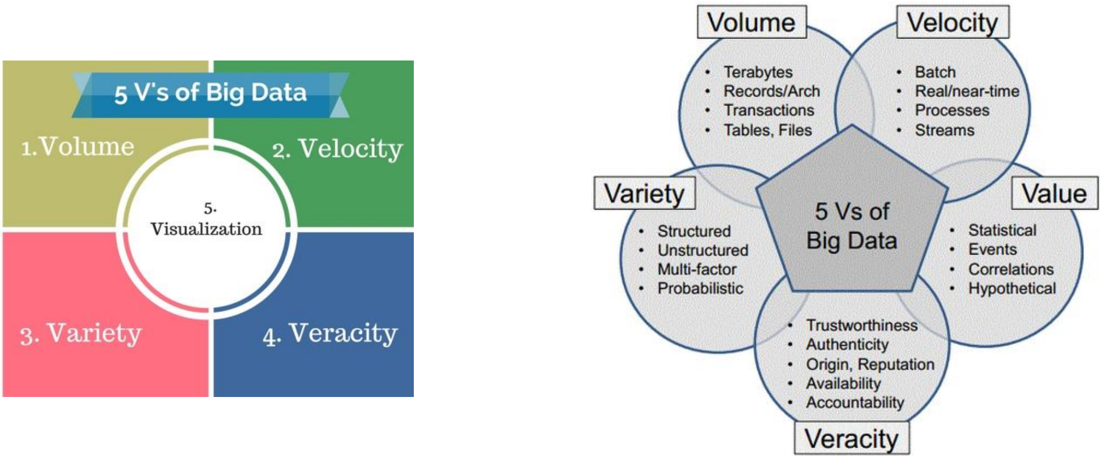


a) b) c)

d) e)

Fig. 9. Examples of the 4 Vs Environmental Big Open-Access Data models
a) basic segmented cycle model; b) basic segmented square model; c) detailed segmented square model; d) V-shaped detailed block model; e) O-shaped detailed quadrant model.

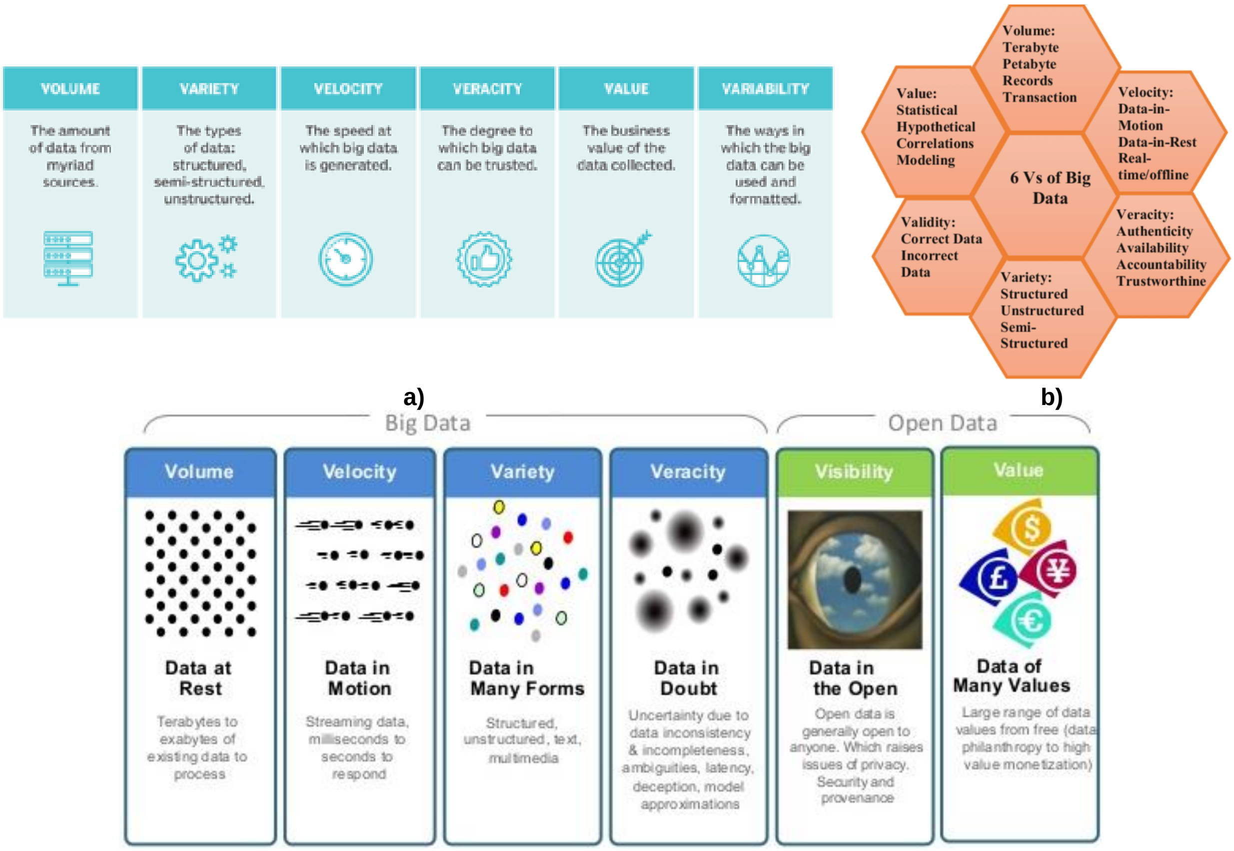
Arguably the most important, but surprisingly a new addition, from 4 to 5 Vs models, in environmental sciences, is the need for verifiable environmental data - veracity, uncertainly of environmental data (see Fig. 10).



a) **b)**

Fig. 10. Examples of the Environmental Big Open-Access Data models with 5 parameters
a) basic segmented cycle model; b) detailed cycle matrix model

In order to determine a data set's accuracy and integrity, not just of the data, but also the sources that generate it, we need to underline the connection between Big Data and Open-Data models, as in Fig. 11; if there is no trust in the environmental data source, the data itself is virtually useless.

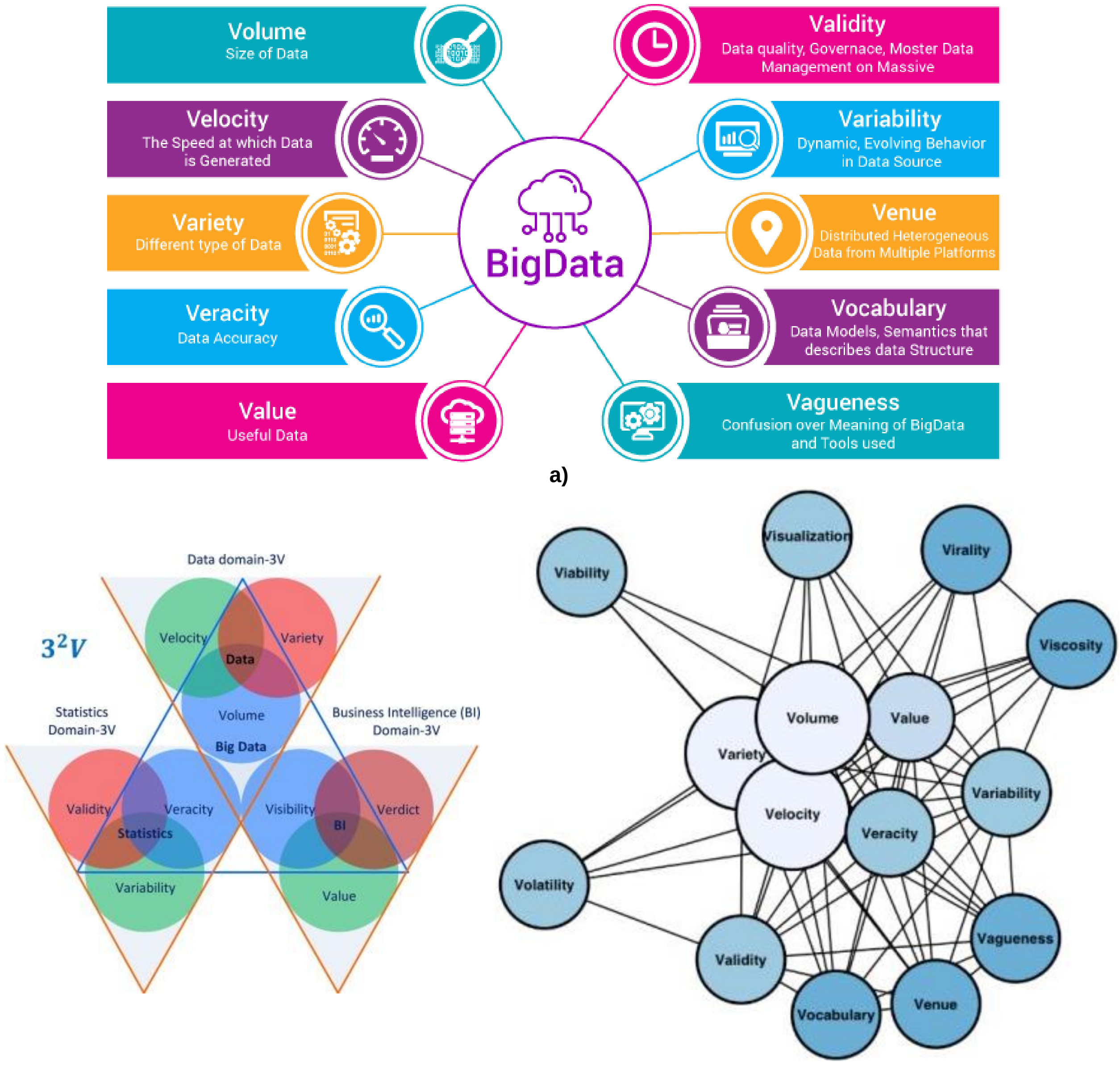


a) **b)**

c)

Fig. 11. Examples of the Environmental Big Open-Access Data models with 6 parameters
a) detailed block list model; b) detailed hexagonal-block model;
c) the integration between Big Data and Open-Data models

The complexity of Environmental Big Open-Access Data models with more than 6 parameters is defined, in the development-context of EISs and EI, as any data that cannot be captured, managed and/or processed using traditional data management components and techniques. To express this kind of complexity we refer to it in Fig. 12.



a)

b)

c)

Fig. 12. The ecomplexity of the Environmental Big Open-Access Data models
a) detailed block list model; b) detailed multi-triangle-block model; c) the graph/network model

3. Conclusions

Environmental data lifecycle management is becoming increasingly important, especially in the context of the attempts of sustainable development of communities, and since the explosion of environmental big data and the ongoing development of the Internet of Things (IoT).
Enormous volumes of environmental data or environmental big open-access data are being generated by an ever-increasing number of devices all over the world. Proper oversight of this data throughout its lifecycle is essential to optimize the decision-making, problem solving, analyzing trends and so on, to generate a better future for the next generations.

References

[1] Cioruţa, B., M. Coman, and A. Cioruţa. (2012) “GeoGebra software – a new possibility for studying the environmental problematics?” *Journal of Environmental Research and Protection (Ecoterra®)*, no. 30: 40-47, www.ecoterra-online.ro/files/1339069750.pdf.
[2] ***. https://blog.epa.gov/2016/10/21/filling-the-gaps-in-environmental-science-with-big-data.
[3] Afgan, N.H., and M. Carvalho. ”The Knowledge Society: A Sustainability Paradigm.” *Cadmus Journal®* 1, no. 1 (2010): 28-41, www.cadmusjournal.org/.../The Knowledge Society A Sustainability Paradigm.pdf.
[4] Cioruţa, B., and M. Coman. “The evolution, definition and role of Environmental Information Systems in the development of environmental protection strategies”/”Evoluţia, definiţia şi rolul SIM în dezvoltarea strategiilor pentru protecţia mediului.” *Journal of Environmental Research and Protection (Ecoterra®*, no. 27 (2011): 11-14, www.ecoterra-online.ro/files/1321371401.pdf.
[5] Cioruţa, B., and M. Coman. “A forey in modern scientific research of the environment. From EISs to EI”/”Incursiune în cercetarea ştiinţifică modernă a mediului înconjurător. De la Sistemele Informatice de Mediu la Informatica Mediului.” *Journal of Environmental Research and Protection (Ecoterra®)*, no. 29 (2011): 17-20, www.ecoterra-online.ro/files/1330955124.pdf.
[6] Avouris, N.M., and B. Page. *Environmental Informatics: Methodology and Applications of Environmental Information Processing*. Dordrecht, Boston, Kluwer Academic, 1995.
[7] Günter, O. *Environmental Information Systems*. Berlin, Springer, 1998.
[8] Hilty, L.M., B. Page, F.J. Radermacher, and W.F. Riekert. *Environmental Informatics as a new Discipline of Applied Computer Science*. In Avouris, N., and B. Page, (Eds.). *Environmental Informatics – Methodology and Applications of Environmental Information Processing*. Dordrecht, 1995.
[9] Page, B., and L.M. Hilty. *Trends in Environmental Information Processing*. In Brunnstein, K., and E. Raubold (Eds.). *Applications and Impacts. IFIP Transactions A-52*. Amsterdam, 1995.
[10] Page, B. *Environmental Informatics - Towards a new Discipline in Applied Computer Science for Environmental Protection and Research*. In: Denzer, R., G. Schimak, and D. Russel (Eds). *Environmental Software Systems*. Proceedings of the International Symposium on Environmental Software Systems, May 1996, Pennsylvania State University, Malvern, London.
[11] Cioruţa, B., A. Cioruţa, and M. Coman. “Pleading for an environmental informatic culture forming need (Pledoarie pentru necesitatea formării unei culturi informaţionale environmentale.” *Journal of Environmental Research and Protection (Ecoterra®)*, no. 30 (2012): 31-39, www.ecoterra-online.ro/files/1339069625.pdf.
[12] Cioruța, B., and M. Coman. “Environmental Informatics - solutions and emerging challenges in environmental protection.”, *Studia Universitatis Babeș-Bolyai AMBIENTUM* 57, no. 2: 17-30. Paper presented at International Conference „Environmental Legislation, Safety Engineering and Disaster Management”(ELSEDIMA®), Oct. 25-27 2012, Cluj-Napoca http://studia.ubbcluj.ro/download/pdf/903.pdf.
[13] Cioruţa, B., M. Coman, and V. Mateșan. “Environmental Information Systems: solutions and emerging challenges for modern strategic development of Romanian local communities.” Paper presented at International Conference "Scientific Research and Education in the Air Force" (AFASES®), May 23-25, 2013, Braşov, www.afahc.ro/.../Cioruta_Coman_Matesan.pdf.
[14] Cioruța, B., M. Coman, A.A. Cioruța, and A Lauran. “From Human-Environment Interaction to Environmental Informatics (I): Theoretical and Practical Implications of Knowledge-based Computing.” *Magazine of Hydraulics, Pneumatics, Tribology, Ecology, Sensorics, Mechatronics (Hidraulica®)* no. 1 (March 2018): 71-82, hidraulica.fluidas.ro/2018/nr1/71-82.pdf.
[15] ***. https://digitalfullpotential.com/big-data-processing-cycle-pitfalls.
[16] ***. www.mbaofficial.com/mba-courses/research-methodology/what-are-the-types-of-observation.
[17] ***. http://analytics.rsystems.com/data-engineering.
[18] ***. https://peda.net/kenya/ass/subjects2/computer-studies/form-3/data-processing/dpc2.
[19] ***. www.chesapeakebay.net/discover/glossary#E.
[20] ***. www.popstoolkit.com/_glossary.aspx?l=E.