UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE Working paper
CONFERENCE OF EUROPEAN STATISTICIANS 27 July 2016

Workshop on Implementing Standards for Statistical Modernisation,
21 – 23 September 2016

National adaptation of GSBPM and its application in the development of a process metadata model.

D. Salgado and A.I. Sanchez-Luengo

Dept. Methodology and Development of Statistical Production Statistics Spain (INE) Paseo de la Castellana, 183 28046 Madrid (Spain) May 3, 2016

Abstract: Statistics Spain (INE) has recently developed and is currently implementing a standard for the documentation of all statistical production processes. This standard is based upon the Generic Statistical Business Process Model (GSBPM) and comprises a third level of subprocesses adapted to our needs. Each subprocess is documented by specifying its inputs, outputs, throughput, tools, documentation, and responsible unit(s). We borrow from computer science general principles such as modularity, abstraction, hierarchy, and layering to cope with the inherent complexity of a statistical production system. Here we offer a general description of the creation of this standard and of its on-going implementation. We include some reflections about the main difficulties towards a modern industrialised statistical production system.

1 Introduction

Statistics Spain has been immersed in the past years in the development and implementation of a system of process metadata in consonance with international standards and, in particular, with the GSBPM v5.0 (UNECE, 2013a). This project is pursuing simultaneously three main goals:

(i) To fulfill Statistics Spain's commitment made in the second round (2013-2015) of Peer Reviews within the European Statistical System (ESS) (Peer Reviewer, 2015).

(ii) To assure the institutional sustainability of the production of Statistics Spain by documenting the current statistical production processes executed by the organization.

(iii) To pave the way for a deep analysis of the current production system driving us to a standardised production model.

In the second round (2013-2015) of the ESS Peer Reviews Statistics Spain made the commitment of intensifying "its efforts to specify and start applying the Generic Statistical Business Process Model across the statistical production processes and introduce systematic standardisation for the different stages of the statistical production process" according to the European Statistics Code of Practice (Principle 4, indicators 7.2 and 12.1) (CoP, 2011).

To guarantee that the production is institutionally sustainable under a decreasing trend in staff number and budgetary limitations, it is necessary, among other things, to put in place standardised production tasks to be exchangeably executed by diverse personnel and automatised as much as possible. Every detail of all production processes must be documented according to internationally accepted standards so that the knowledge is made firmly resident in the organization and not person-dependent.

Finally the industrialisation of the production processes in the world of official statistics is an internationally recognized necessity since some years ago (HLG-MOS, 2011). Moreover, the digitalization of the economy, the monetization of data and their exploitation by private firms and different stakeholders with increasingly widespread data science skills put a high pressure upon this need for a change. In this sense, a detailed analysis of the current production system which allows us to develop a new standardized production model must be a clear objective of a metadata system.

At Statistics Spain we have undertaken this task by (i) launching a pilot experience with the original GSBPM v5.0, (ii) analysing the results, (iii) developing a third level of the GSBPM v5.0 adjusted to our national needs thus producing a standard for the implementation of our process metadata system and (iv) finally launching the fieldwork stage feeding this system.

Here we present the main milestones of this on-going process. In section 2 we present a brief motivating analysis for our choice of general principles in the development of the standard. In section 3 we describe the adopted standard. In section 4 we include a general description of the implementation of the standard and the activities currently on-going. Finally we close with some comments and conclusions in section 5.

This work has been carried out by an internal working group comprising survey conductors, domain experts on different subject matters, IT, sampling, data collection, statistical dissemination, and data quality under the coordination of the department of methodology and development of statistical production. We acknowledge their intense efforts without which this standard would have been impossible.

2 General principles

In the development of our process metadata system we rapidly identified the need for a set of guiding principles allowing us to achieve the goals posed in the introduction. In this sense, in consonance with international and other national initiatives we have clearly pinpointed the UNECE GSBPM v5.0 (UNECE, 2013a) as the framework to develop a more detailed standard.

The first course of action was to launch a pilot experience with 7 statistical operations to collect their process metadata of phases 4 to 7 in this model asking for a description of the tasks described at the second level of the GSBPM. Apart from issues bound to the perception of these initiatives among different survey conductors and domain experts (see section 4), we found the result clearly unsatisfactory, since (i) no detailed information about the different production processes could be attained not even being minimally useful for the stated purposes, (ii) GSBPM level-2 processes were documented to an extremely diverse degree of detail (from very limited to highly condensed), and (iii) it was impossible to have comparable GSBPM level-2 processes among different statistical operations.

We took the decision to develop a third level of the GSBPM adapted to the current production system at Statistics Spain. Notice that this implies that up to the second level our process metadata system strictly follows the GSBPM and that the third level is in agreement with the general principles posed in this international standard.

In the construction of this third level we sought for complementary guiding principles, which we found in the following observation: **a statistical production system is a complex system**.

Although a definitive scientific definition of complexity is extremely difficult, we can easily recognize in statistical production systems several features defining a complex system (Saltzer and Kaashoek, 2009):

- Large number of components: as an illustration just consider the 44 level-2 subprocesses identified in the GSBPM, which in turn can be further decomposed into finer production tasks, and this must be multiplied by the number of statistical operations under production in a statistical office.

- Large number of interconnections: needless to say, most of the different production tasks are intricately interconnected in such a way that a variation in a given task can have unexpected *waterbed* effects in another tasks (**?**).

- Many irregularities: as survey conductors and domain experts rightfully underlined when referring to their own statistical operations, each survey portraits specific characteristics somehow singling them out from the rest. No clear-cut regularity allowing production designers to pose universal rules can be cross-sectionally identified among all statistical operations.

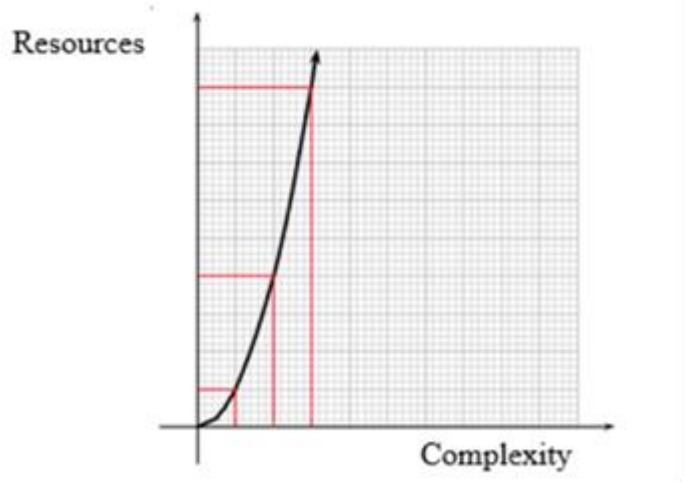

Figure 1: Square law of computation in terms of complexity and resources

- A long description: as a further complication to the preceding feature, protocols, rules, guidelines, or instructions to accomplish the diverse production tasks cannot be described in a homogeneous fashion. A lot of exceptions are indeed the rule.

- A team of designers, implementers, or maintainers: not only is it that different professional profiles ranging from IT experts to statisticians are needed to produce official statistics, but also do they need coordination and communication.

Based on these features, different simple models (Weinberg, 2011) can be used to justify the so-called *square law of computation*, which can be adequately represented by figure 1. The bottom line of this figure is the fact that for each increasing unit of complexity (in arbitrary units; e.g. a new breakdown of estimates, a change of normative regulations, etc.) requires an increasing amount of resources. Indeed, should the production model be kept under these premises, the increasing demand of information upon statistical offices will eventually collapse the production. Notice, that the quadratic behavior is somehow arbitrary for our argument and any increasing convex curve deduced from those features will support our claim. Complexity is the ultimate reason why resources hypothetically enough to accomplish a given task are not sufficient when that task is part of the production system.

Our approach bears on the reflection that, *since the fabric of statistical production is information, we should use some principles of computer system design to cope with the complexity of statistical production*. In particular, we claim that *functional modularity* (modularity + abstraction) together with *hierarchy* and *layering* (Saltzer and Kaashoek, 2009) arise as useful principles to structure statistical production in a way to cope with its inherent complexity.

The GSBPM already articulates statistical production in a modular way, being its fundamental modules the level-2 subprocesses. Furthermore, this modularity is accomplished by dividing the production process following to a high degree more or less natural or effective boundaries (although in practice we have found some tensions between some statistical methodology natural boundaries and the current modules, e.g. data editing strategies are necessarily split up across several modules, in detriment of functional modularity). Thus abstraction is also envisaged in the GSBPM.

Regarding hierarchy and layering, these have been used as assisting principles in the construction of the metadata system. In particular, so far the project has not undertaken the development of term-value pairs like in the Dublin Core Metadata Initiative (DCMI, 2016) describing the process, since the production is still far from that level of standardisation. Thus a lot of textual descriptions by the different experts taking part in the design, development, execution, and monitoring of the production process are the main building blocks in the metadata system. Hierarchy and layering are used to structure these texts so that a future promotion to a DCMI-like system could be undertaken more easily.

Finally we are aware of the high degree of interrelation between the GSBPM and other international metadata standards, in particular, the Generic Statistical Information Model (GSIM) (UNECE, 2013b). Statistics Spain has not yet adopted the GSIM as a production standard, but, as we shall see in the next section, the process metadata system already includes some elements preparing a future transition to describe information objects according to this model.

3 The process metadata standard at Statistics Spain

The documentation of the process metadata standard can be found at Statistics Spain's web page (StatSpain Standard, 2015). By and large, the main two features are (i) the development of a third phase of the GSBPM adapted to Statistics Spain production and (ii) the adoption of the modelling language Business Process Model and Notation (BPMN) 2.0 (OMG, 2016) as the tool to model and document the different business workflows. This has been complemented with extensive user-oriented documentation.

The development of a third phase of the GSBPM has been undertaken following the same philosophy of this model with a view on the future adoption of the GSIM. The identification of the new production tasks conforms to the following guideline. Every task is identified with the syntactic construction *verbname*. Each task belongs to either phase 1 up to 7, each with a clear meaning and role within the whole process, role which we express through the choice of *verb*: *identify* for phase 1, *design* for phase 2, *develop* (*component of the information system for*), *execute* for phases 4 to 6, *disseminate* for phase 7, and *monitor* for phase 8. The *name* is chosen to identify the element of production upon which the action (identify, design, develop, execute, monitor) is exerted. This construction finds its limits in the use of natural language itself so that some combinations of those verbs with the different elements are not appropriate and alternative verbs must be used with the same general meaning. As clear advantages we claim that the adoption of an information model standard (e.g. based on the GSIM) will be somehow more effortless and that the analysis of the evolution of elements of production across the production chain will also be more straightforward.

As a first obstacle in this approach it was necessary to carefully choose the terms to be used in the standard as well as their precise meaning. The normalization of language came up against the general usage of diversified jargon in each statistical domain, so an extra effort was needed. In other words, the need for a *controlled vocabulary* (Controlled Vocabulary, 2016) was put in place and those terms referring to elements of production entering tasks designation names constitute the first step towards such a controlled vocabulary.

Let us consider an example. We have identified as an element of production the so-called *population aggregates*, defined as functions of object and/or auxiliary variables of all population units or of a domain thereof. Usually these are the quantities to be estimated and whose knowledge is pursued with the statistical operation. The choice of the term *aggregate* is intentionally made to underline its distinction from the concept and term of *variable*, defined as an either quantitative or qualitative characteristic of a single population unit. In the standard, this element of production can be found in different phase-3 tasks:

- 1.3.1 Identify population aggregates;
- 2.2.3 Operationalise population aggregates;
- 2.5.8 Design estimators of population aggregates;
- 2.5.9 Design correction of estimators of population aggregates;
- 3.2.11 Program estimators of population aggregates;
- 3.2.12 Program correction of estimators of population aggregates;
- 5.7.1 Compute (corrected) estimators of population aggregates.

Notice how the element of production *population aggregate* can be straightforwardly traced along the production chain[1].

Additionally, the standard also tries to reflect the aggregation of elements of production into more complex elements. Let us consider the following example. The concept of *variable* starts off

[1] By *population aggregate* we agree to designate *a function of object and/or auxiliary variables of every element of the population or of a domain of interest therein*. Note the subtle contrast with the notion of *variable*, agreed to designate *a characteristic of a population unit taking either qualitative or quantitative values*

by its identification (1.4.2 Identify variables) to be later operationalised (2.2.2 Operationalise variables). The set of operationalised variables are to be collected in the form
of survey questions conforming the *questionnaire* (2.3.3 Design questionnaire). Notice that *questionnaire* is understood in abstract terms as the set of survey questions driving us to the values of the operationalised variables. This abstract *questionnaire* begins to take physical form when considering the collection instrument (2.3.6 Design collection instrument), i.e. paper, telephone, web… Usually most of current electronic collection instruments incorporate data collection metadata giving rise to paradata increasingly used in the streamlining of the production process (2.3.8 Design paradata). All these are ingredients for the design of a data collection management system (2.3.9 Design data collection management system), which embraces all the preceding elements. Notice the accretion of elements constituting the complex element *data collection management system*.

Each individual task in the standard is documented according to international guidelines (ITFMF, 2013). In particular, each task is specified through its inputs, outputs, throughput (or process), documentation, tools and responsible unit(s), which will presumably pave the way for a milder adoption of the GSIM standard (UNECE, 2013b).

The inputs are specified as an itemized list where each item refers to a concrete element of production for the statistical operation at stake identified with its name (if any) and a very brief description (if necessary). For example:

- File E30103.FF- V1.MM[*mmyyyy*].D-1 with the data set of final edited microdata for reference period *mmyyyy* (output of task m.n.p).
- Design of coding (output of task m.n.p).
- Error detection rules (output of task m.n.p).

Notice that for later ease of the documentation of business workflows through the BPMN language, we also include the task producing these elements as outputs.

The outputs follow similar lines, except for the specification of the tasks producing the elements (*output of m.n.p*). Notice that exact coincidence between inputs and out-puts of different tasks is enforced. Furthermore, only elements of production used in other tasks or being a final result of the process (e.g. the press release) are to be included in the outputs. In this sense, inputs and outputs are indeed the interface among different tasks (modules) thus achieving functional modularity: the execution of each task in independent of one another, being the inputs and outputs the only point of interaction.

The throughput (or process) is a detailed description of the steps to produce the outputs out of the specified inputs. A difficulty arises because of the lack of a term-value structure in the standard: how much detail is to be included in the documentation and how are the different process steps to be documented?

A delicate trade-off between the work load of staff feeding the metadata system (the different survey conductors and domain experts themselves) and the fulfillment of the former objectives was put in place. Two criteria were adopted: each task must be minimally documented with enough details as to (i) allow a novel staff member to replicate the task and (ii) allow the metadata unit to analyse the current production system to propose a standardised model to converge

towards. In this sense the description must descend gradually from the process to the procedure to fulfill these requirements.

Besides, to pursue functional modularity, hierarchy, and layering in the documentation of the processes we strongly recommended to provide parameterised descriptions as much as possible. For example, if a task involves an activity consisting of checking the values of several variables against validation intervals to produce editing flags for further error treatment, a good practice is to assign this activity a name and to describe it with input parameters *VariableName*, *ValidIntervalSet*: activity *CheckVarbyValidInt*(VariableName, ValidIntervalSet). Thus, whenever this activity is executed, only a reference to its name with the running parameters is actually needed. The activity is described only once (DRY principle (Hunt and Thomas, 2000)).

As another guideline, the use of technical terms is favored to produce precise descriptions. For example, referring to a simple stratified sampling design with Neyman allocation using a given auxiliary variable as a covariate (Särndal *et al.*, 1992) is so precise a language that no further explanation is needed.

As a final comment regarding the throughput, the sustainability of the metadata system over time has also been taken into account, pursuing ease of maintainability and updating.

The documentation section is intended to include any piece of documentation providing further details of the execution of the process, whereas the tools section is to embrace any element (document, software tool, protocol…) necessary for the execution of the process. All these elements are referenced by a single name to be included in a *resource catalog* linking this name to an informal internal URI, a description, and a list of tasks where they are used.

The responsible unit is the unit responsible both of the execution of the task and of its updated documentation. Functional units effectively executing the tasks have been chosen as appropriate instead of organic or nominal teams as directorates, subdirectorates, etc. (nonetheless, most of them are coincident, although not always).

As final agreements regarding the tasks, firstly the standard will be applied to each single statistical operation present in the Statistical Operations Inventory of the Spanish National Statistical Plan (StatSpain IOE, 2016). Thus, each task will be documented in the context of the statistical operation in which it is executed. Secondly, those tasks within a give operation jointly executed, i.e. integrated in the identification, design, development, execution, or monitoring, will also be documented in an integrated way indicating in one of the involved tasks that the process is documented together with the accompanying task. Thirdly, tasks extending across several statistical operations (as e.g. cross-sectional tools development) will be documented apart from each statistical operations. Cross-sectional elements of production present in several statistical operations will be referenced by their common name. As a prominent example, the Spanish automatic data editing and imputation tool system called DIA (Villán-Criado, 1992) used by different surveys is just referred to in the documentation by DIA. No further reference is included.

Finally, the second main feature of the standard is the adoption of the BPMN language as the tool to model and document the different business workflows within the production system. The BPMN language standard (OMG, 2016) presents four levels of conformance, namely process modeling, process execution, business process execution language (BPEL) process execution and choreography modeling. Only the first one has been adopted so far and not fully. There are three conformance subclasses (descriptive, analytic, and common executable) within the process

modeling conformance. We have adopted the first subclass in our standard possibly with elements from the analytic subclass when more expressivity is needed.

4 The implementation

Once the standard was approved by Statistics Spain's Executive Board on April 7, 2015, the implementation phase began. The work is currently in progress.

The work was planned in two different stages. Firstly, five statistical operations were chosen based on their diverse characteristics: short-term and structural business statistics, a satellite national account, an administrative data-based statistics, and a household survey. The motivation was to span a wide range of characteristics to analyze how the adopted standard would deal with them. The process metadata of these statistical operations were elaborated by the group in full detail (except for the BPMN diagrams). Complementarily, these five statistics metadata would serve as guiding examples in the construction of the process metadata for the rest of statistical operations in the office.

Secondly, detailed documentation of this process was prepared. Internal briefings with the involved units were organised and an internal software application was developed to collect the metadata. The goal is to have a fully-fledged and feeded process metadata system as soon as possible. In this sense another 30 operations are already on their way, and as of this writing, another 30 will very soon start their work.

Due to the current high work load across the institute, we have so far focused upon the documentation of each production task according to the standard, leaving the construction of BPMN diagrams of the different processes as well as the documentation of cross-sectional processes (such as software application development) for a later phase.

5 Some comments and conclusions

Now we comment relevant issues standing as a hindrance for the deployment of an industrialising metadata system. We find this interesting regarding the urging need of modernisation and industrialisation of official statistics. The main obstacle we have detected is the persistent cultural resistance to change and the pressing lack of human resources.

In the same lines presented in section 2 we firmly believe that official statistics production is gaining in complexity due, among other things, to the increasing demands of information from society through public administrations, governments, social and scientific institutions, and different stakeholders. To cope with this complexity adequate tools must be used, many of which, we claim, can be naturally found in computer science.

The dipolar relationship between statistics and computer science is in our view especially pernicious in official statistics. This is probably a reflection of the historical academic background of the diverse staff at statistical offices. On the one hand, you may find little sensibility towards statistical methods (e.g. indifferently using either design-based or model-based or some defective treatment of non-sampling errors). On the other hand, simplifying identifications of computer science with conceptions such as "just a matter of programming a formula" can be heard from time to time. All these attitudes drive us to this dangerous dipolar situation.

As illustrative examples of consequences of potential misconceptions we have observed that many misunderstand the idea of standardising the statistical production system with that of standardising the computer applications. Having standard computer applications does not amount to having a standardised production process (although it certainly aids in this purpose). Also, the simplification of computer applications development to friendly menu-navigating and button-clicking graphical user interfaces without really structuring the statistical processes below them clearly introduces inefficiencies in the production system.

Many computer science principles such as modularity (Baldwin and Clark, 2000) are extraordinarily useful to assist in the design of an industrialised production process. Both the strong convictions of some survey conductors and domain experts about the singularity of their surveys and the paralyzing designs of some applications not properly dealing with minor details of the daily production could be overcome by putting an end to this dipolar situation. An official statistician must be inexcusably aware of both the most adequate statistical methods and computer science principles necessary to efficiently implement those in an industrialised production system.

We detect as an important ingredient in the cultural resistance to change regarding the industrialisation and standardisation of official statistics the limited conception of metadata as a sheer documentation system of already executed processes. This contributes to the perception of metadata as an extra burden somehow alien to production tasks themselves.

In this sense we find it strategic to integrate the use of metadata in production tasks not only for newly created statistical operations but also for on-going surveys (e.g. to streamline and normalise different aspects). The benefits of metadata standards must be made as manifest as possible during design, development, execution, dissemination, and monitoring tasks. A metadata system must be considered as an input in a production system, not just as a documenting result.

Another relevant obstacle in the implementation of a metadata system is the work load put on production staff, especially after the budgetary restrictions brought in by the international financial crisis. Not only does staff have to face limited human resources but do they also have to implement this necessary change in the production model. Recognizably this is a complication needing special attention from the top management.

To sum up, official statistics production is a complex system which needs adequate tools to cope with this complexity in order to achieve both sustainability and efficiency. Among these tools we recognize as vital an adequate metadata system fostering interoperability of production processes not only among different statistical operations within a statistical office but also across diverse official statistics producers.

In this sense, at Statistics Spain we have followed the GSBPM as a general frame-work to develop a description of our production processes. This framework has been complemented with a third level in this model further adapted to our needs. This has been accomplished borrowing some general principles from the design of computer systems such as modularity, abstraction, hierarchy, and layering. This is intended to cope with the inherent complexity of official statistics production.

Each production task is documented through the specification of its inputs, outputs, throughput (process), documentation, tools, and responsible unit(s). Each of these elements are described as open-text items following some recommendations to homogenise the diverse descriptions. Gradually this will hopefully drive us to a DCMI-like value-term scheme. The analyses coming out of this metadata system will eventually allow us to take the next step in its development.

References

Baldwin, C.Y. and Clark, K.B (2000). Design Rules (vol.1): The power of modularity. MIT Press.

European Statistics Code of Practice (revised edition, 2011). Available at http://ec.europa.eu/eurostat/web/products-manuals-and-guide-lines/-/KS-32-11-955 (accessed on April 30, 2016).

Wikipedia entry on *Controlled vocabulary*. Available at https://en.wikipedia.org/wiki/Contr-olled vocabulary (accessed on April 30, 2016).

Dublin Core Metadata Initiative (2016). Available at http://www.dublincore.org (accessed on April 30, 2016).

High-Level Group for the Modernisation of Official Statistics (2011). Strategic vision of the High-Level Group for strategic developments in business architecture in Statistics. Conference of European Statisticians. Geneva, June 14-16.

Hunt, A. and Thomas, D. (2000). The pragmatic programmer: from journeyman to master. Addison-Wesley.

Informal Task Force on Metadata Flows (2013). Metadata flows in the GSBPM. WP 22. UNECE Work Session on Statistical Metadata. Geneva, May 6-8.

Object Management Group (2011). Business Process Model and Notation. Available at http://www.omg.org/spec/BPMN/2.0/PDF/ (accessed on April 30, 2016).

Peer Review Team (2015). Peer reviewers recommedations and INE of Spain improvement actions in responses to the recommendations. Available at http://ec.europa.eu/eurostat/documents/64157/4372828/2015-ES-improve-ment-actions/17c9399c-2801-4696-9b2b-b2c212a9dcf5 (accessed on April 30, 2016).

Saltzer, J.H. and Kaashoek, M.F. Principles of computer system design: an introduction. MIT Press.

Särndal, C.-E., Swensson, B., and Wretman, W. (1992). Model assisted survey sampling. Springer.

Statistics Spain (2015). Standard for documenting production processes of statistical operations of the INE. Available at: http://www.ine.es/en/clasifi/estandar-procesos en.pdf (accessed on April 30, 2016).

Statistics Spain (2016a). Inventory of the Statistical Operations of the Spanish State General Administation. Available at http://www.ine.es/buscar/searchResults.do?searchString=IOE&Menu boton-Buscador=Buscar&searchType=DEF SEARCH&startat=0&L=0 (accessed on April 30,2016).

UNECE (2013). Generic Statistical Business Process Model (v5.0). Available at http://www1.unece.org/stat/platform/display/metis/The+Generic+Statist-ical+Business+Process+Model (accessed on April 30, 2016).

UNECE (2013). Generic Statistical Information Model (v1.1). Available at http://www1.unece.org/stat/platform/display/metis/Generic+Statistical+Information+Model (accessed on April 30, 2016).

Villán-Criado, I. (1992). Análisis de reglas de depuración de datos **34**, 151–171.

Wall, L. (2013). Wikipedia entry on *waterbed theory*. Available at https://en.wikipedia.org/wiki/Waterbed_theory (accessed on April, 30 2016).

Weinberg, G.M. (2011). An introduction to general systems thinking. Weinberg and Weinberg.