Semistructured information in the organization of processes for managing large production systems

Y V Polishchuk
Orenburg State University, Orenburg, Russian

E-mail: Youra_Polishuk@bk.ru

Abstract. The paper proposes a functional-structural representation of a decision support system for large production systems that implements the ability to process semistructured factual information of related operational content through a single access interface. The application of the latter provides quick access to all factual information for a group of decision-makers in the synthesis of control actions on a large production system or its subsystem.

1. Introduction
In this article, by large production systems (LPS) we will mean systems consisting of a combination of enterprises and organizations functioning as a single whole within the framework of one technological process.

As an example of LPS, we can consider the industrial complex of the Orenburg gas-condensate field, which includes many organizations and enterprises that implement the process of extraction, processing and transportation of field products.

Analysis of modern decision support systems (DSS) and them mathematical tools allowed us to conclude that there are no universal ways to support decision making for LPS. For this reason, the control action on the LPS or its subsystem is synthesized by a group of decision makers (GDM), which selects and applies methods for generating managerial decisions depending on the problems of the problem being solved [1].

The defining property for LPS from the position of related operational content is its multi-format and low formalization. The multi-format in operational documentation of LPS is due to the long period of its life cycle, during which repeated upgrading of the software and hardware involved in its operation can occur. The low formalizability in operational documentation of LPS is also due to the lack of content models for the documents included in its composition. It is worth noting that the requirements for factual context, prescribed in the regulatory documentation, can be violated by its executors both by accident and intentionally due to the lack of control over its generation at the formation stage.

Thus, the accompanying operational content of the LPS is formed in the form of two repositories: factual data coming in the form of formalized information, for example, from SCADA, and electronic documents whose content is represented by semistructured factual information. Semistructured information is understood to mean information in which a certain structure can be distinguished, however, this structure in advance is not fully or partially known, or may change over time [2].

2. The problem
To generate effective management decisions, GDM must have quick access to factual data located in any of the considered repositories.
There is a possibility of implementation access for GDM to factual data in electronic documents with semistructured content by using a single access interface. For its implementation, it is necessary to mark up the factual data inside the document using a semistructured model of the factual content of the document, which, taking into account the Xml Schema Definition (XSD) specification [3], can be written as follows [4]:

\[ S = \{ \text{root}, sObj, LObj, \text{minOccurs}, \text{maxOccurs}, sMet, \text{Obj}_\text{smet} \} \]  

where root - is the root object, root \( \in \) sObj; sObj - a finite set of objects, each of which contains a fragment of the content of the document (text, picture, etc.) or acts as a container for one or more objects.

The following meta-properties are available for container objects: smet - defines the object as a container; mixed - allows the use of descendant objects in an arbitrary order. LObj - a mapping defined on the set sObj, such that \( LObj : sObj \rightarrow \{ \text{obj}_1, \ldots, \text{obj}_n \} \), where \( \text{obj}_i \in \text{sObj} \) - is a child object; \( n \) - the number of child objects; Obj _ smet - a mapping defined on the set sObj, such that \( \text{sObj} \rightarrow \{ \text{sMet}_1, \ldots, \text{sMet}_n \} \), where \( \text{sMet}_i \in \text{sMet} \) - meta-property of the restriction on the contents of the object; \( \text{minOccurs} \) - a function that determines the minimum possible number of times an object is used in the model; \( \text{maxOccurs} \) - a function that determines the maximum possible number of times an object is used in the model.

To develop a semistructured model of the factual content of a document, we use five parameters by analogy with multicriteria optimization of technical systems [5]. Consider these parameters in a ranked order.

As the first parameter, we take the result of document validation. The considered parameter characterizes the compliance of the model with the standard for the content of the documents in question. If the validation condition is not met, the model is rejected.

\[ P_1 = \frac{\text{Oobj}(sObj_{\text{doc}})}{\text{Oobj}(sObj)} = 1, \]  

where \( \text{Oobj} \) - a function that returns the number of required objects from the set that satisfy the condition: \( \text{minOccurs}(sObj) = 1 \), where \( sObj_j \in sObj \); \( sObj_{\text{doc}} \) and \( sObj_s \) - respectively, the set of objects used in the document, and the objects of the model in question. This parameter should be equal to one, since the use of required parameters is a condition for the use of a document model.

The second parameter characterizes the degree of detail, i.e. the size of the content marked in the document model objects (structural units) of the model. This parameter should be minimized, since when working with objects it is always easier to enlarge than to detail the content of the document.

\[ P_2 = \sum \frac{\text{Len}(sObj)}{\text{Obj}(sObj)} \rightarrow \min, \ sObj_j \in sObj, \]  

where \( \text{Cobj} \) - a function that returns the number of set objects that satisfy the condition \( \text{Obj}_\text{smet}(sObj) \cap \{ \text{sMet}_j \} = \emptyset \).

The third parameter characterizes the density of use of model objects, i.e. the level of elaboration of the document model and characterizes the uniform distribution of objects (structural units) of the model in the document.

\[ P_3 = \frac{1}{k} \sum_{j=1}^{k} P_{c_j} \rightarrow \min, \ P_{c_j} = \frac{p_{\text{char}_\text{count}(j)}}{p_{\text{obj}_\text{count}(j)}}, \]  

where \( p_{c_j} \) - is the density of model objects on a document fragment (the fragment is equal to the document page or paragraph, i.e. the document consists of \( k \) fragments); \( p_{\text{char}_\text{count}} \) - a function...
that returns the number of characters in a particular fragment; \( p_{\_obj \_count} \) - a function that returns the number of \( s\text{Obj} \), such that \( \text{Obj \_sme}(s\text{Obj}) \cap \{\text{sme}_i\} = \emptyset \) are entirely located in a given fragment. This parameter should be minimized.

The fourth parameter will characterize the saturation of the objects (structural units) of the model in the document, i.e. quality of the description of the content of the document.

\[
P_4 = \frac{\text{char\_count}}{\text{Len}(s\text{Obj})} \rightarrow \min , \quad \text{Obj \_sme}(s\text{Obj}) \cap \{\text{sme}_i\} = \emptyset ,
\]

(5)

where \( \text{char\_count} \) - a function that returns the number of characters in the document; \( \text{Len} \) - a function that returns the number of characters from an object. This parameter should be minimized, since it is necessary to reduce the content that is not described in the model of document.

The fifth parameter characterizes the flexibility of the model. A more flexible model allows you to more effectively describe the semistructured content of the document, therefore, this parameter should be minimized.

\[
P_5 = \frac{\text{Aobj}(s\text{Obj}) - (\text{Uobj}(s\text{Obj}) + \text{Robj}(s\text{Obj}))}{\text{Aobj}(s\text{Obj})} \rightarrow 0 ,
\]

(6)

where \( \text{Aobj} \) - a function that returns the number of objects without taking into account the root element \( \text{root} \); \( \text{Uobj} \) - a function that returns the number of set objects that satisfy the conditions: \( \text{minOccurs}(\text{Obj}) = 0 \); \( \text{maxOccurs}(\text{Obj}) \geq 1 \); \( \text{Robj} \) - a function that returns the number of set objects that satisfy the conditions \( \text{minOccurs}(\text{Obj}) = 1 \); \( \text{maxOccurs}(\text{Obj}) > 1 \), where \( s\text{Obj} \in s\text{Obj} \).

To assess the quality of the models under consideration, we introduce a scalar criterial objective function (7). Its value for the best of the considered models will be minimal.

\[
R = \sum_{i=1}^{5} \frac{P_i}{P_i} \rightarrow \min ,
\]

(7)

where \( P_i \) - is the value of the \( i \) parameter of the analyzed model.

In the case when the model does not correspond to the content of the document, i.e. its validation is not performed, the objective function (7) will be indefinite.

The proposed set of parameters (2)…(6) and the objective function (7) makes it possible to compare semistructured models from the standpoint of the efficiency of describing the content of documents. The latter allows the synthesis of semistructured content models for a collection of electronic documents of one kind.

3. Practical realization

Taking into account the capabilities of a single interface for access to factual data, the following functional-structural presentation of DSS for LPS based on semistructured content can be proposed (figure 1).

In the considered scheme, bold arrows show the physical impact on the control object, which is implemented by maintenance service employees and SCADA. The dashed arrow in the diagram shows the movement of metadata of the generated document. The remaining arrows in the diagram show the movement of managerial content in the system. SCADA will provide facility management based on algorithms and constraints defined by the GDM.

Factual information about the operation of the facility comes from SCADA in the factual data warehouse. About non-standard situations of SCADA are signaled by GDM, which generates control action based on the processing of content from the data warehouse (DW) and new factual data received from performers through electronic documents with semistructured content. Received electronic documents are stored in the electronic document repository for further automated processing of the managerial content contained in them.
Figure 1. Functional-structural presentation of DSS for LPS based on semistructured content.

A similar scheme is activated to obtain new factual content in cases changes in the operating mode of the control object. Electronic documents with new factual information content are created on the basis of the models generated by GDM, and the content of these documents is formed by maintenance service employees, interacting with the management object taking into account the document metadata. Documents are saved to the electronic document repository only if they are successfully validated.

Consider, as an example, the processing of hydrodynamic research of gas wells (HRW) Orenburg gas-condensate field in the proposed DSS.

Under the conditions of Orenburg gas-condensate field, this type of research is implemented by more than 10 different organizations. The final report on the research results is presented by an electronic document in the MS Word format, the content of which submitted by semistructured information.

To mark up the factual content of the HRW report, a semistructured model of the factual content of the document content has been developed (Figure 2).

The (runcomp) segment can be used an unlimited number of times. It is defined in the document model as the data type (Truncomp), let us consider its structure in more detail. (Truncomp) data type consists of the following mandatory objects: (dat) - date of research, (vidis) - type of research, (regis) - research mode, (timeb) - start time of research, (timee) - end time of research, (inti) - interval of research (the interval is defined as the data type (T interval), which was considered earlier), (davusb) - wellhead pressure at the start time of research, (davust) - wellhead pressure at the end time of research, (davzatb) - annulus pressure at the start time of research, (davzate) - annulus pressure at the end time of research.

The (pribor) segment can be used an unlimited number of times. It is defined in the document model as the data type (Tpribor), let us consider its structure in more detail. (Tpribor) data type consists of the following required objects: (typep) - device type, (dataet) - calibration date, (dirpart) - head of the calibration.

The proposed semistructured model of the factual content of a HRW document allows to solve most of the problems associated with the processing of HRW results performed by several independent organizations. To solve them, all contractors should use the model described above in the processes of
preparing the final documentation. The latter reduces the number of errors in the documentation and increases the efficiency of processing its factual content.

Figure 2. Semistructured model of the factual content a HRW report document.

Fragments of the use of the considered model for marking the content of a HRW report in MS Word format are presented in figure 3.

Figure 3. An example of marking up factual information using a model.
As a database management system in DSS, a system supporting the XML format can be used, for example, Oracle [6], PostgreSQL [7], etc.

4. Conclusions
The developed functional-structural representation and the basic information processing algorithms for DSS based on semistructured content for LPS implements the possibility of processing its operational factual content through a universal access interface and provides additional control over the factual content of the documentation through the restrictions specified in the semistructured content model. The latter minimizes the possibility of entering incorrect factual data into the documentation.

References
[1] Polishchuk Y V 2014 Semistructured content in the management of large and complex technical systems (Samara: Samara State University of Transport Communications) p 191
[2] Paley D 2002 Open Systems 9 57-64
[3] Walmsley P 2012 Definitive XML Schema 2nd (Upper Saddle River: Prentice Hall PTR) p 768
[4] Polishchuk Y V 2014 Pat. of the Russian Federation No 2517428 appl. 13.06.2012, publ. 28.03.2014
[5] Gmoshinsky V G and Florent G I 1973 Theoretical foundations of engineering forecasting (Moscow: Nauka) p 304
[6] Chang B, Scardina M, Karun K and Kiritzov S 2000 Oracle XML Handbook (New York: McGraw-Hill Education) p 357
[7] Juba S and Volkov A 2017 Learning PostgreSQL 10 2nd (Birmingham: Packt Publishing) p 488