Distinguishing Between Data and Know ledge in the Spatial Analysis and Regional Mapping

E. Gusev[1], C. Pshenichny[1], V. Akimova[2], A. Bartova[1]

[1]Geological Mapping Division, VNIIOkeangeologia, Angliisky Prospect, 1, St. Petersburg 190121, Russia;
E-mail: gus-evgeny@yandex.ru
[2]Geomorphology Dept., Faculty of Geography and Geoecology, 10 Liniya V.O., 33, St. Petersburg 191126, Russia

1. Introduction

Development of geoinformation systems (GIS) changed our understanding of geology by opening vast opportunities of bringing different strands of information together and by the power of visualisation in orthogonal or geographic coordinates. This has drastically increased the number of potential customers of geological information products, the most common of these being regional geological maps (general and thematic). Meanwhile, the GIS imposed new or strictified some existing methodological requirements on the procedure of mapping.

Thus, the ability of GIS to incorporate practically any kind of spatially-related information a customer may be interested in, puts the questions like,

(i) when mutual and diverse customers desire to have various layers in a future GIS, how to sort out and group up these desires,

(ii) how to cope with the different terminologies the customers may use to denote similar things,

(iii) whether the parameters the customers consider relevant to their field are really that relevant and related to regional geology.

Furthermore, today they mean GIS not only to show but also to compute the information and draw conclusions. With this, another set of questions arises,

(i) what format to choose for presentation of the information related to this or that layer and what mathematical or other tools of computation apply to this format;

(ii) what if different strands of information (layers of a GIS) should be computed to obtain the result the customer needs, but the computation tools are not supported by at least one of the formats;

(iii) should we have some underlying idea of which layers may and which may not be interrelated by their contents before building a GIS or just “play with information” in an already built system and search for new dependencies?

These are far from all the questions we confront when compiling the geologic maps in GIS format. These questions, being purely practical, urge us to revise the way we used to reason in geology. This necessity was emphasized yet by the pioneers of application of information technologies in the geoscience (Loudon, 2000). In attempt of doing this, the geo-reasoning discussion forum was brunched on the Internet by S. Henley and C. Pshenichny, a special workshop was organized by Pshenichny at the IAMG2003 meeting in Portsmouth (Pshenichny, 2003a), where, among other matters, a draft of the complex formal approach to information treatment (Pshenichny, 2003b) was discussed as a possible solution of the complications like those outlined above.

Fundamental for the complex formal approach is the distinction between the two types of information, knowledge and data. In terms of traditional Aristotle logic, *Data* is anything expressed as a singular statement (in which the predicate is related to a specific subject), e.g., “the sample 72039 contains 241% CaO”. It is commonly accepted in the science that data are the result of observation or measurement. *Knowledge* is anything expressed as a general statement (in which the predicate is related to a generic subject), “the rock” (meaning “at least some of the studied samples”) “contains (contain roughly) 2.4 wt.% CaO” (Pshenichny and Carniel, submitted).

In this paper, we explore the opportunity to build a complex formal methodology for regional geological mapping in a GIS medium based on strict distinction between the types and strands of information involved.

2. Some Guidelines for the Complex Formal Methodology of Geological Mapping

2.1. General Architecture of GIS

Considering the stages of creation of a GIS for a given object, it seems reasonable to start with the customers’ preferences and the way they can be processed. Enquiring potential customers is essentially a *knowledge engineering* task. First, a more or less formalized interview form (not only for but even for written, but even for an oral communication with the customers) is desirable, and then, the feedback should be processed to answer the question, what layers we should intend to include in the future GIS.

Processing of the feedback implies (i) reconciliation of customers’ wording, (ii) finding the relations between the parameters of the customers’ interest, the parameters that are to be included in a regional geologic map and the actual field(s) of the customers, and (iii) classification of parameters requested by the customers that proved to be pertinent.

Reconciliation of terms can be done by grouping synonyms and antonyms used by different customers, additional consult with experts, looking at the related literature and thesauri and especially to the XML wordlists and ontologies being developed now for various fields, including geology (look, e.g., POSC (1994-2004) or GeoreferenceOnline (2001)).

Once the wording is clarified and a terminology fixed, one needs to make sure the parameters customers request are really pertinent to regional geology and to their own field. The latter is also I important, as the customers, being aware of the computational potential of GIS, may have wrong understanding of the relation(s) between their field and geology and thus have false expectation to calculate, say, productivity of woods from distribution of LREE in the lower crust. At this stage, a very simple semantic net may be very useful to examine and re-focus the customers' interests in the way they can be best fulfilled (Akimova et al., submitted).

Classification of parameters finally requested by the customers unlikely can proceed in full accordance with the logical rules of classification. For instance, mapping a continental platform, we may collect the expressions of interest in general structure of the crystalline basement, distribution of garnet in particular gneiss in the basement (from the geologists studying Pre-Cambrian), grain size of two from three sandstone layers of the sedimentary cover (from sedimentologists), variation of thickness of particular shale (from paleogeographers), orientation of faults inactive in the Holocene (from neotectonists) and intensity of seasonal erosion in river basins (from geomorphologists).

Having received this, first we may classify the requests as

1. those relating the crystalline basement,
2. those relating the non-(crystalline basement), that is, sedimentary cover.

However, then, to incorporate the distribution of garnet in particular gneiss in a logically correct framework, we should split 1 into

1.1 requests relating the structure of crystalline basement and
1.2 requests relating the composition of crystalline basement,

then divide 1.2 into

1.2.1 requests relating the composition of crystalline rock unit 1,
1.2.2 requests relating the composition of crystalline rock unit 2,
1.2.3 requests relating the composition of crystalline rock unit 3 (the particular gneiss),
1.2.4 ...

and then classify 1.2.3 as

1.2.3.1 requests relating the distribution of plagioclase in the rock unit 3,
1.2.3.2 requests relating the distribution of pyroxene in the rock unit 3,
1.2.3.2 requests relating the distribution of garnet in the rock unit 3,

Many of the obtained classification cells will be empty (i.e., relating to no actual customer's request). If still to make a special layer in a GIS for the distribution of garnet in particular gneiss, then the computaiton of potential relationships between this and other mapped features of crystalline basement and sedimentary cover may appear misleading, because lots of other akin features (say, distribution of other minerals in this and other rock units) are not considered, though might be influential on the computed parameter. Obviously, the fulfillment of the customers' desires as expressed makes little sense in GIS-related regional geological mapping.

The authors think that a solution could be a clarification of the structure of knowledge at the stage of planning the architecture of the future GIS jointly with the potential customers. For instance, if we cannot provide exactly the layer they need, we could stop at a more general level like "composition of crystalline rock units" providing a layer for each unit, and one of these will appear generic for the requested one, so that the customer may create or purchase another GIS reaching exactly the parameter of interest based on the GIS we make; the matter of negotiation, then, is the distance between the level of generality we stop at and the level the customer needs. Certainly, the customer's interest would be to get as close to the parameter(s) of direct interest as possible. What should be the interest of the designers of GIS?

In the authors' understanding, this interest must ensure that any permitted computation of information stored in the GIS will not lead to senseless (uninterpretable) result. To ensure this, one needs to somehow permit one relations between the strands of information and forbid others; the same with the means of computation applicable (inapplicable) to the strand of information expressed by each layer.

The applicability of means of computation will be discussed in the next section; here our concern is definition of relations between different strands of information.

A number of ways can be thought of to describe relations between the contents. This can be done just verbally - e.g., the parameter “dip azimuth” makes sense to faults, strata boundaries, erosion thalwegs, but makes no sense to chemical composition of rocks and sediments, and so forth. Such listing of what is possible and what is not is rather wordy and, that is yet more important, ununderstandable by computer.

Also, the relations between different strands of information can be defined by various conceptual graphs. For instance, we may involve the same (or modified after clarification) semantic net as before.

A more sophisticated option could be an event-bush (Pshenichny and Khrabrykh, 2002; 2003; Pshenichny et al., 2005; also this volume). The virtue of this approach is division of all the terms (properties, features) under consideration into three groups, *primary* (marking some basic input into lie context), *secondary* (describing the *incoming circumstances* imposed by the context and the entities that appear as instant or distant consequences of interaction of the basic inputs and these circumstances) and *tertiary* (denoting the final results being queried). It is mandatory that primary and secondary terms are neither repeated nor synonyms and have no contextual ambiguity. Adding an existential predicate like “is” or “takes place”, these terms are transformed into statements (see examples in Pshenichny et al., this volume). These statements and their combinations must be mutually incompatible. It is highly desirable that they together exhaust all the opportunities for basic inputs and/or incoming circumstances in the chosen terminological (conceptual) framework. The choice of primary terms, terms for incoming circumstances and the tertiary terms should follow from lie title, formal or informal, of the information product, i.e., the GIS. For instance, the GIS created at regional geological mapping can usually be entitled as “Field relations of geologic bodies, distribution of natural resources and hazards - *optional*) in a given territory depending on the regional framework”. Hence, at first look, primary terms should describe the subject (geologic bodies), that is, le the names of rocks occurring in the considered territory taken from an appropriate ontology (see above) and, possibly, names of some other relevant entities like groundwater, soil, vegetation types or something else. The incoming circumstances should be the terms describing the main attribute related to the subject - the regional framework. This can be endogenic (structural and, optionally, paleogeographic or other) and exogenic (environmental, geomorphologic, depositional). The tertiary terms display the main item(s) of interest as the GIS title reads - field relations of rocks and, possibly, something else (added from the customers’ feedback). Tertiary terms should be the cells of the legend of the geologic map as put on paper. Another example could be an event bush for GIS “Geometry of the interior and surface structures”, where “dip azimuth” would be primary term, “Faults”, “Strata boundaries”, “Erosion thalwegs”, “Chemical composition”, etc., - incoming circumstances, having some or no relation to the primary term.

The event-bush representation is readily converted into a record in a language based on the first- (or higher-)order predicate logic (Kleene, 1952), e.g.

Single-valued predicate *Dip azimuth* (t), t={x, y, z}, is defined for any point x in the mapped space with orthogonal co-ordinates x, y, z, for which the predicates *Fault* (t), or *Strata boundary* (t) or *Erosion thalweg* (t) are defined; this allows us to put, say, Vt

(*Fault* (t) ∩ *Dip azimuth* (t)) ("For any t if it is true that t belongs to a fault, it is true that it has dip azimuth"). Primary terms become the individual variables (t in the considered case), incoming circumstances - the predicates used in formulation of basic assumptions.

Further opportunities this record affords will be discussed in the following section. Here we should stress that such record is the third and the most formal known way to define the relations between different strands of information when determining the architecture of future GIS.

All the suggested ways of definition (verbal, graphic and logical) are compatible, complementary and perhaps not the only possible. They all may be used to finalize the formation of the architecture of the future GIS. As can be concluded now, the whole process of design of the architecture of a GIS is processing of some formal knowledge from different domains, not only from regional geology, and by different means of knowledge engineering (including the conceptual graphs) and logic. No data are involved at this stage.

2.2. Filling the GIS with Information

The designed architecture of the GIS should be compared with the information available. This is commonly understood as creation of a relational database and linking it to the layers of GIS. However, as shown by many researchers (e.g., Henley, 2000), data are often missing or have irregularly varying uncertainty that impedes not only computation but even their correct storage in chosen formats. This led to wide application of Bayesian approach in geoscience. This approach takes some *prior information* that is essentially the knowledge, not necessarily adequate, of the behavior of a phenomenon in question regardless of whether we have sufficient data or not (e.g., Curtis and Wood, 2004). Bardossy and Fodor (2005) demonstrate how new bits of drilling data change the probability of prior hypotheses of the shape of an orebody. Pshenichny (2004) showed that data are not necessary at all and Bayesian estimates can be performed for bare knowledge if to substitute data with logical assumptions written in the language of mathematical logic.

Contrary to data, knowledge relates not to particular objects (points, pixels, etc.) but to classes of objects (possibly empty). Either a knowledge (e.g., a model) describes the entire object as a representative of some class, or the object is understood as "assembled" from elements taken from different classes, each of these being described by some model(s). Classes may be organized in taxonomies (hierarchies) with the attributes inherited downward thus restricting the models, while the object itself may consist of parts, with the attributes inherited upward, from parts to the whole, thus searching dataset. In practice of geoscience it is often difficult to discern taxonomies from Smyth. 2003). E.g., every stratigraphic unit is unique, but to correlate two units one needs some general knowledge about die class of units these two belong to. Moreover, as shown by Pschenichny (2003b), the distinction between the data and knowledge is often context-dependent, but will can be drawn in every particular case (e.g., in particular GIS).

Thus the involvement of knowledge *along with* and *independent of* data is inevitable in GIS. Naturally, this knowledge must be somehow spatially-related, i.c., pertain to particular points (areas) in orthogonal coordinates. As an option, it looks reasonable to make two layers for at least for some of the mapped or computed properties, one for data and one for knowledge - for instance, the layer (map) is measured orebody roof depths with some blank fields (missing data) and the layer for hypotheses of the shape of the orebody roof, preferably with several alternative hypotheses for every point (for marine vs. glacial origin of given deposit or different interpretations of similar structure based on different geotectonic paradigms). Mapping knowledge is not an entirely new task for cartography; in many cases, e.g., in social geography, when the area of residence of a nation is given, actually the knowledge (some conclusion from possibly heterogeneous data) is presented; however, commonly they do not realize this fact. However, storage of knowledge is not as formalized as storage of data; some options of creation of formalized knowledge bases are discussed and exemplified in Pchenichny et al (this volume). Showing data and knowledge separately for each property of interest for GIS would provide the basis for independent symmetric calculation of Bayesian and possibly better estimates for data and for knowledge and improve the reliability of computation results in GIS.

3. Concluding Remarks

Complex formal methodology of GIS-related geological mapping includes two principal stages: design of the architecture of the GIS and filling the system with information. At the former stage, the general knowledge of the geology of the region and knowledge from other domains the customers may be interested in is processed by means of knowledge engineering including Aristotelian logic and, desirably, the first-order predicate logic. At the latter stage, the knowledge and data on the mapped area are selected in accordance with the designed GIS architecture and related to the special layers. Specific tasks of storing and mapping the knowledge arise herewith. The suggested methodology is inspected to better focus the GIS-related geological mapping to the customers' needs and improve the capability of computation results obtained in such GIS.

4. References

Bardossy G.. Fodor J.. 2005. Assessment of the Completeness of Mineral Exploration by the Application of Fusil Arithmetics and Prior Information. Acta Polytechnica Hungarica. 2(1), 15-31.

Curtis A., Wood R., 2004. Elicitation of probabilistic information from experts. Geological Prior Information, Curtis. A. and Wood, R. (Eds), Geological Society, London, Special Publications. 239.

Georeference Online Ltd., 2001. www.georefercnceonline.com.

Henley S., 2000. The Missing Data Problem: www.silicondale.com.

Kleene S.C., 1952. Introduction to Mathematics. D. Van Nostrand Co., Inc., New York - Toronto.

Loudon T.V., 2000. Geoscience after IT: a view of the present and future impact of information technology on geoscience. Elsevier, Oxford. 142 pp. Also available as Computers & Geosciences, Special Issue, 26 (3A), A1-A142.

POSC - Petrotechnical Open Standards Consortium, 1994-2004. www.posc.org.

Pshenichny C.A., 2003a. Georeasoning Workshop in Portsmouth (Sept., 11, 2003) Summary. IAMG Newsletter, 67 (December 2003), 15-19.

Pshenichny C.A., 2003b. A Draft for Complex Formal Approach in Geoscience. Modeling Geohazards: IAMG 2003 Proceedings, Portsmouth UK; Editors J.Cubitt, J.Whalley, S.Henley; http://www.jiscmail.ac.uk/files/GEO-REA-SONING/papers.html.

Pshenichny C.A., 2004. Classical logic and the problem of uncertainty. Geological Prior Information. Curtis A. and Wood. R. (Eds). Geological Society. London, Special Publications, 239. 111-126.

Pshenichny C.A., Carniel R., Development of Complex Formal Methodology for the Volcanic Hazard Assessment Bulletin of Volcanology, submitted.

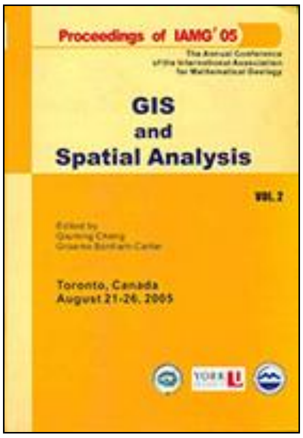
Pshenichny C.A., Khrabrykh Z.V., 2002. Knowledge Base of Formation of Subaerial Eruption Unit. Environmental Catastrophes and Recovery in the Holocene(Abstracts), Brunel University, London: http://atlas-confercnces.com/cgi-bin/abstract/caiq-22.

Pshenichny C.A., Khrabrykh Z.V., 2003. Knowledge engineering: New opportunities for natural hazard assessment in the Northern Caucasus. Russia. Sovremennye melody geologo-geofizicheskogo monitoringa prirodnykh processov na territorii Kabardino-Balkarii, B.S. Karamurzov (Ed.), Nalchik, Kabardine-Balkar University Publishers, 57-67 (in Russian).

Pshenichny C.A., Carniel R., Akimova V.L., 2005. Decreasing the uncertainty of BBN technique by means of complex formal approach to volcanological information treatment. European Geosciences Union (EGU) 2nd General Assembly, Vienna (Austria), 24-29 April 2005, Geophysical Research Abstracts, 7, EGU05-A-01016.

Smyth C., 2003. Distinguishing Partonomies from Taxonomies in Science Languages: A Prerequisite for Computer-Aided Georeasoning: Modeling Geohazards: IAMG 2003 Proceedings, Portsmouth UK; Editors J. Cubitt, J. Whalley, S. Henley; http://www.jiscmail.ac.uk/filcs/GEOREASONING/papers.html.909

Reference:



Gusev E., Pshenichny C., Akimova V., Bartova A. 2005. **Distinguishing between data and knowledge in the spatial analysis and regional mapping.** GIS and Spatial Analysis - 2005 Annual Conference of the International Association for Mathematical Geology, IAMG 2005. Vol. 2, 904-910.