A Taxonomic Representation of Scientific Studies

Yevhenii B. Shapovalov and Viktor B. Shapovalov

National Center Junior Academy of Sciences of Ukraine, Dehtiarivska St, 38/44, Kyiv, 04119, Ukraine

Abstract

Taking to account the problem of the high amount of scientific data, it is relevant to develop methods of its structuration and processing. Using ontology graphs is the perspective modern way to provide. Taking to account that most studies is written based on IMRAD, it was used to provide integration of different studies at a single structure and provide structuration at all. The different ways to create integrated ontology using IMRAD have been described. To get the necessary level of abstraction IMRAD elements as part of a set of specific study were decomposed as levels of abstraction from L1 (general integration node with generalized data) to L5 (specific papers with specific data) depended on the abstraction. The content of each node in form of metadata and its further processing is described. The specific way of the usage of proposed modes has described on the example of the describing studies in the field of biogas production. The proposed approach can be used in a single field or be even more integrated and be devoted to the structuration of the works in different fields.

Keywords

ontology, IMRAD, structuration, scientific studies, biogas

1. Introduction

The data nowadays generated with huge intensity. Due to this, now, Big Data processing is a trend [1, 2]. Processing of the huge amount of data in real life is complicated by high gain of publishing of scientific studies.

Considering the development of the STEM-education, studies are provided not only by experienced scientist but by youth. Such huge amount of the studies generated complicated task to process such data. One of the problem of low spreading and usage (in example of Ukraine [3, 4]) may be related to difficulties.

Now, scientific studies are published in different forms of reports, such as articles, conference proceedings, books, etc. However, its process is hard due reports are low-structured. Sure, they all build by similar structure named IMRAD [5, 6]. It includes requirements for the paper to consist of some generalized Introduction, describing of used Materials and Methods, naming the Results of the study and is Discussion by comparing with other scientific materials or providing of use cases. However, it seems not enough. Here just some examples of problems due to it:

- hard to start the researcher career due to complicated process of understanding of the methods and equipment that need to be used in specific fields of study
- hard for youth scientists to understand main parameters that has measured to provide study analysis
- harder for experienced scientists to analyze and collect data of new studies.

These are only very few cases that are a problem due to high amount of data of scientific studies. However, these cases are makes relevant to develop new methods to provide better structuration and data processing of scientific studies.

There are few solutions for this problem that provides automate science data processing [7–10], but it seems that they do not take to account IMRAD. One of the methods that is relevant to use to solve the problems is ontology taxonomies [11–14] with semantic technologies [15]. Also, ontology taxonomies has a lot advantages such as possibility to combine with other types of materials [16].
including interactive and web-based courses [17], other information technologies [18, 19] and GIS [20]. This research aims to develop the model that can structure the set of the studies using IMRAD.

Previously, it was proposed to provide support using ontologies for single specific study, but not to create glossaries and structured sets of data. To provide it tools Open provenance, Ontologyt and EXPO [21] were developed. Another ontology solution in the field of science is MoKi that provides creation of wiki-based information scientific sources [22, 23]. Sure there some specific ontology tools such as Gene ontology [24] or Centralized educational environment [25]. However, creation of ontology to structure the set of the studies seems relevant due lack of approaches to provide it.

2. Methods of the research

In the paper, the model of ontology has developed using of main principles of Graph Theory, Set Theory, and a Theory of Abstraction [26]. The graph was modeled using a simple hierarchical algorithm that foresees using only nodes and links. So, such a model further may be updated using the more extended graph building tools such as weight coefficients. However, without simple modeling, providing it will not be possible. To provide structuration generally accepted structuring method IMRAD has proposed and used.

To model data processing was developed taking to account processing possibilities of the Polyhedron [20, 27, 28] system due it has some advantages compare well known Protégé [29, 30] and OWL tools [31–33]. The principles of cognitive IT-platform’s tools Filtering, Audit, and Ranking to provide decision-making data processing tools were described in form of mathematical model (set of equitation) [2, 25, 34].

3. Results

3.1. Methods of the research

As was noted before, IMRAD is used to prepare science papers. So, to provide structuration it is possible to use parent nodes that represent IMRAD components. IMRAD - Introduction, Methods, Results, and Discussion. The discussion part can’t be structured by ontology because it contains the analysis and comparing of the obtained data. That is why discussion will be represented as the processing of the results.

\[ \neg (D \in \text{REP}) \Rightarrow (P \in \text{REP}) \]  

where REP – report (or set of report) D – discussion of report’s results; P – processing of the results of a set of studies.

Approximate, ontology can be devoted to a specific field of science or integrate different fields. In depended on it, the ontology will have 5 or 4 abstract levels of deep. In the case of general ontology, the parent node will be “Scientific reports” and its subsidiary nodes will name a specific field. In the case of a specific ontology, the parent node will name a specific field. Then it links with elements of IMRAD structure. Each element of IMRAD has its specific representation and it’s in turn linked with the element of IMRAD. And the leaf node will be a set of the specific researches belongs to the field. Let’s name each level with L symbols taking to account position in hierarchy:

L1 – general name of parent’s node “Scientific reports”; L2 – Name of field of the reports; L3 – Part of IMRAD; L4 – Specific representation of IMRAD (specific method, used materials, specific type of the results); L5 – Specific reports where were used specific representations of IMRAD L4.

Therefore, the hierarchy on the specific study will have a form of \{L2, L3, L4, L5\} and the general ones will have a form of \{L1, L2, L3, L4, L5\}. Interoperability of the L2 nodes of two different graphs may be provided by using the graph constructor. It provides the possibility to merge graphs in two ways. The first fore-sees that graphs will be constructed as a general graph in form of \{L1, L2, L3, L4, L5\}
and with the same name of L1. And the second is to create L1 in the constructor and add there two specific graphs in form of \{L2, L3, L4, L5\} and merge them. Schematic representation of the general ontology is shown in and taxonomy of the specific field is shown in Fig. 1.

![Diagram](image)

**Figure 1:** The taxonomy of the general science study ontology (a) and taxonomy of the specific field science study ontology (b); where LR1, LR2, M1, R1, R2 – are abstract classes of literature review (LR), Methods and results of object; “

As an alternative and a more human-readable way to provide abstraction is to revert this model and begin by L5 and end by L1. In this case, ontology will have such structure \{L5, L4, L3, L2, L1\}. Same rules are relevant for second case.

However, the main disadvantage of such a graph is obvious and is the consequences of the structure: the leaf node SR (“Scientific reports”) will be not the most useful for users. This type of graph may be built as \{L5, L4, L3\} and in this case, it will be used to evaluate the specific report for example during qualifying work evaluation (Ph.D. or Master’s reports). It will show abstract classes of each specific part of IMRAD for each specific study and can provide an evaluation of the set of methods and results that was obtained by the researcher. In this research, we’ll use the first way to provide hierarchies in form of \{L2, L3, L4, L5\}, and \{L1, L2, L3, L4, L5\}.

As it can be seen, the general science report ontology is significantly more complicated due to links between L1 and L2 level, and also there will be some problems with a huge number of methods, results, etc. that can be not necessary to the user that looking for information on the specific field. Also, it will be much harder to create such type of graphs due it will have 2 levels of links “one to many” (see Fig. 1 (a), links between L2 and L3 level and links between L4 and L5 levels) compare to only one in case of specific ontology (see Fig. 1(b), only links between L4 and L5 levels). It may be unreasonable to create a complicated graph Therefore, it seems relevant to provide both types of hierarchies. To provide it, the ontologies should be created specific fields and then merged as noted before.

In this case, specific parts of IMRAD will be used as subsidiaries nodes in the field of the study, and specific study will be used as leaf nodes. So, the general structure of such ontology may be represented as:

\[
\{I, M, R, P\} \in \text{REP}
\]  

(2)

where I – sets of Introduction of all studies, M – set of Materials and Methods an of all studies, R – set of Results of all studies, P – processing of the results of a set of studies; replaces discussion; REP – report (or set of report).

To provide better systematization and we have split the introduction into two different parts due to their specific – basic metadata (BMD) and literature review (LR), it is possible to represent the introduction as further:

Basic metadata of the study node linked with graph nodes that characterized the basic data on the study such as hypothesis, object, subject, practical value, and scientific novelty. And so, a node of the report basic metadata of study can be presented as a further equation:

\[
\text{BMD} = \sum_{i}^{n} \{H_i, O_i, S_i, PV_i, SN_i\}
\]  

(3)
where H – hypothesis or hypotheses of each specific study; O – object of the study; S – the subject of each specific study; PV – practical value of each specific study; SC – the scientific novelty of each specific study.

Lest presents each work of the set of the Introductions, Methods, Results, and Pro-cessing of the data (Discussion). Then each work will be represented as a set of these elements that will be relevant only for each specific study:

\[ S_i = \sum^n_i\{I_{ii}, M_{ii}, R_{ii}, P_{ii}\} \]  

(4)

\[ S_{ii} = \sum^n_{ii}\{I_{ii}, M_{ii}, R_{ii}, P_{ii}\} \]  

(5)

So, these articles can be integrated in a single ontology using IMRAD. Such ontology will be a single set of elements of both studies that related to single area:

\[ O = \sum^n_i\{S_i, S_{ii}\} = \sum^n_i\{I_{ii}, M_{ii}, R_{ii}, P_{ii}, I_{iii}, M_{iii}, R_{iii}, P_{iii}\} \]  

(6)

The main advantages of using such structures are that some parts of introduction (for example keyword), materials and methods and results elements (entities and measured parameters) of studies/reports in the same field can coincide and, in this case, such coinciding sub-nodes will be used as links for them and provide their interoperability. The proposed approach uses IMRAD to collect and process the data with ontologies. In this way, the ontologies are constructed not by the specific structure of each work but by the generally accepted IMRAD structure. The parent node will be specific area to which set of the studies belongs (L2=\( \sum^n_i\) RSii, where L2 – specific area and RS – set of the studies representing). The L2 node is linked with I, M, R, P nodes (representing IMRAD). Each IMRAD node is linked with a specific node (such as ammonia determination by Nessler’s method (for methods) or “chicken manure” or “glycerine” (for subjects)) that belongs. And each specific IMRAD type is linked with leaf nodes of ontology – specific studies where such entities were used.

In this case, few reports (REP1, REP2, and REP3 that belong to L5) will be integrated by some of the methods or results (M1, R1, R2 that belongs to L4). So, the L4 level will be used to provide the structuration of the reports (L5). The user can use it in both ways: to find which method, result, etc. that belong to L4 were used in a specific report that belongs to L5; and to define in which reports belong to L5 specific method, result, etc. that belong to L4 were used.

Coinciding of the studies elements (for example methods) may be represented as single intersections of such elements in few studies:

\[ M_i = \{M_a, M_b, M_c, M_d\} \]  

(7)

\[ M_{ii} = \{M_b, M_c, M_d\} \]  

(8)

Therefore, in this case, and Mb can be used as a parent node that connects two different studies. The node Mb itself will contain general theoretic information on it, and node S_i and S_{ii} will contain information on the specific case of its usage and measured parameters using it. For example, same will be a hierarchical way of representing and using the keywords.

This will be useful especially for students and young scientists that looking to find methods (MI) and parameters that can be used in specific fields and their usage in practice. Also, this way provides a list of the parameters and methods that used in specific fields.

3.2. Metadata

Metadata of each work will be used for processing the data. It may be included for each node. Metadata of L4 nodes will represent the general information (for example, the essence of the method itself), and the resulting leaf nodes will contain the specific metadata related to specific study (such as specific results of the study obtained using set methods M; for example, metadata: “5,35”, and it's class: “Ammonium nitrogen content, g/l”). And so, metadata with the same class will be processed by using filtering by users’ request or by ranking by providing the rank of nodes by specific class (or their set)
based on the user’s request. So, each node located on each level $E_i$ contains metadata with the level of abstract that corresponds to a number of levels; for level 1st – it will be the most abstract metadata and for 5th – it will be the most specific.

As it can be seen all data in levels L1-L4 contains generalized metadata and wouldn’t be used for processing of specific study, but just used to get generalized abstract information on entities used in specific fields. Only L5 level contains metadata that related to specific study and will be used for further processing.

Specific mechanisms “Filtering”, “AUDIT” and “RANKING” of cognitive IT solution Polyhedron are used to provide processing of the information. It will be used for the case when different reports will have the same Class and Type of information, but different value:

$$\{\text{Class: C1; Type: Number; Value: V1} \} \in \text{REP1}$$

$$\{\text{Class: C1; Type: Number; Value: V2} \} \in \text{REP2}$$

And the values V1, V2, and V3 can be equal or not equal. “Filtering”, “AUDIT” and “RANKING” can be used to process the data. Filtering can be described by function if:

- If ($V_{\text{min}} < V < V_{\text{max}}$) then (display nodes with such V) or
- If ($V = V_{\text{set}}$) then (display nodes with such V)

Where $V_{\text{min}}$, $V_{\text{max}}$, $V_{\text{set}}$ are maximum, minimum, and given (set) values, respectively, that inputted by the user.

The function of AUDIT can be also described as a function if:

- If ($V_i = V_{\text{set}}$) than (mark red such $V_i$); do it for all $V_i$.

The ranking is much harder and can be described as:

$$\text{RANK}_{\text{abs}}(i) = \sum (\text{OR}_i \times \text{IMP}_i \times \frac{V_i}{V_{\text{max}}})$$

where $\text{RANK}_{\text{abs}}(i)$ – ranking rank in absolute value for i’s node OR$_i$ – orientation maximum or minimum for metadata of I’s object (can be +1 or -1); IMP$_i$ – importance coefficient for metadata of I’s object; $V_i$ – the value of metadata of i’s object; $V_{\text{max}}$ – maximum value of the set of metadata.

$$\text{RANK}_i = \frac{\text{RANK}_{\text{abs}}(i)}{\text{RANK}_{\text{max}}}$$

where $\text{RANK}_i$ – the relative value of the rank (can be maximum =1) of each object; $\text{RANK}_{\text{max}}$ – the maximum value of the RANK for all sets of objects.

4. Discussion

So, for the specific case of biogas production studies [35–38], it seems relevant to use ontology for a specific field (in form of {L2, L3, L4, L5}). In this case, a node in the L2 line will be single and named “Studies on anaerobic digestion”. It will be linked with nodes Introduction, Methods, Results, and Processing. As for all other cases, Introduction will be divided to Basic Metadata and Literature review (L3 level).

Basic Metadata will be linked with nodes Objects, Subjects, Aims, Practical Value, Scientific novelty, Hypothesis, Keywords, Abstract, Conclusion (L3 level).

Each of these nodes will be connected with specific nodes relevant for the set of the studies (L4 level) that was structured. Each specific L3 will have metadata with general information on the described object.

Each of such nodes will be connected with study where it was used (L5 level). For example, “Biogas production from the poultry waste” or “Utilization of the meat production wastewater using anaerobic digestion”.

The Literature review node (L4 level) will be connected with specific studies used in a set of the studies. Its name will be the name of the study (paper, article, conference processing, thesis, etc.) similar to the name of the study used to provide structuration with the addition of the publishing year. For
example, it can be named “Utilization of the meat production wastewater using anaerobic digestion, 2011”. Each such node should be connected to the one of few studies used to provide structuration (L5 level).

The most useful will be Methods and Results nodes. They will be useful to students and youth scientists who want to be familiar with methods used in the field and set the measured parameters that used in the field of science. Sure, the established scholars will use such a tool too to increased outlook. The Materials and Methods node will be divided into Methods, Equipment, and Materials.

Each such subsidiary node connected with a leaf node that is specific study. The Processing node has metadata with type link and its value in form of a link to Audit and Ranking tools for the set of reports that are structured.

Each work has metadata that mostly duplicates the structure. For this, all numeric and semantic data of the works is added to a node of the specific work it belongs to. Examples of the metadata of the leaf nodes are presented in the table. It is foreseen to provide automatically. It will be necessary to provide filtering, Audit, and ranking.

Ontology models are the basis of the effective ontology creative process. Such model like proposed and others (for example ontologies of educational environments, will be useful to build set of the different ontologies and have similar conceptual states of abstraction. Using such approaches and providing of semantic technologies can be useful to provide interoperability [15].

Sure, the proposed research focused on the ontology of the specific field in form of \{L2, L3, L4, L5\}, but it is proposed to use integrator of the ontologies of fields and create general ontology in form of \{L1, L2, L3, L4, L5\}. Such integration is important to provide transdisciplinary [39]. In general, proposed approach will be useful and relevant for most of the fields. It will be very embarrassingly to process humanitarian data where less of standardization and numeric data, but it seems that some automated tools like recursive reductor [13] can process and provide structuration even in such fields.

5. References


