Ontological Approach in the Smart Data Paradigm as a Basis for Open Data Semantic Markup

Julia Rogushina

Institute of Software Systems of the National Academy of Sciences of Ukraine, 40, Ave Glushkov, Kyiv, 03181, Ukraine

Abstract

We analyze existing approaches to transformation of raw data into source for analysis and knowledge acquisition named Smart data. This research area refers to data that has been processed, analyzed, and transformed into actionable insights or knowledge. The goal of Smart data is to provide more valuable information that can be used to drive decision-making, enable automation, and support a variety of intelligent applications. One of directions of Smart data deals with data structuring on base of semantic markup where ontologies are used as a source of domain knowledge.

Semantic Wikis are used by researchers for smart data processing, due to their ability to combine the benefits of Wiki technologies (easy editing and collaboration) with the advantages of semantic technologies and ontological analysis (formal representation and reasoning). In this research we propose some special cases of ontologies that reduce domain knowledge according to goals of markup of Semantic MediaWiki resources. We consider advantages of this technology and problems of its practical use. Proposed models and methods are approve in process of development of the portal version of Great Ukrainian Encyclopedia that integrates heterogeneous multimedia information from various fields of sciences.

Keywords 1

Semantic markup, ontology, Wiki, Smart data, Open science

1. Introduction

Now data becomes an asset of immense value, but this value depends of possibility to acquire useful information from this data. The growing interest to intelligent information processing is largely caused by the increase in the volume of available information, the increase in its heterogeneity, and the need to obtain from it exactly the information that specific users can process for solving of their current tasks. Huge volumes of data from various sources are collected and analyzed in order to find economic benefits and competitive advantages for companies and society as a whole. A lot of digital information is created and stored, but not used in any way.

Statistics show that a significant part of global data is unstructured and has a high dimension. As a result, the vast majority of available information cannot be analyzed automatically with the help of modern technologies without additional data handling, and this calls for the development of models and methods for various pre-processing means applied to unstructured raw data. Digital universe generates huge volume of heterogeneous digital data [1], but only small part of data had some kind of structuring and was analyzed.

Unstructured data (USD) usually refers to information that does not have a predefined data model or that model does not correspond to the purposes of processing. USD concept is not well-defined because the same set of data can be structured for one task (if that structure is not useful for their processing purposes) and USD for another. In addition, the structure of data may not have a formal

COLINS-2023: 7th International Conference on Computational Linguistics and Intelligent Systems, April 20–21, 2023, Kharkiv, Ukraine
EMAIL: ladamandraka2010@gmail.com (J.Rogushina);
ORCID: 0000-0001-7958-2557(J.Rogushina)

definition and the possibility and correctness of its interpretation depends on person that uses it and on purposes of use. But unstructured information can have a certain structure (be semi-structured or even structured) that cannot be applied in automated processing without additional clarifications.

For example, such data as a natural language document can be viewed from some points of view as a structured object, and from others – as USD. From linguistic point of view, the structure of document is represented using punctuation marks and syntax elements (relations between sentence members). In addition, some documents contain additional formatting elements – tables, columns, paragraphs, etc. that can also be considered as structuring elements, helping to define sub-components of documents, and embedded metadata that is automatically generated by text editors, displaying information such as creation date, scope, authorship, etc. All this information can be used to search for documents according to certain requirements. But the same document from the point of view of document content analysis is USD, because such structural elements do not contain information about which IOs are described in the document, what properties they have and how they are related to each other.

This leads to problems related to its storage (traditional databases are not designed for such uncertainty) and analysis. Thus, data are considered as USD in those cases when information about their structure cannot make data analysis more effective, but pre-processing of USD with elements of Smart data technologies allows to transform them into structured or partially structured. Such technologies as Data Mining and Text Mining can be used for this. Smart data is the way in which different data sources (including Big Data) are combined, correlated and analyzed [2].

2. Related works

The dissemination of the term “Smart data” is closely associated with the use of Big Data. But it should be taken into account that such transformations can be applied to any arrays of information intended for further processing and use – for example, training samples for machine learning or repositories of services or scientific articles. Big data is characterized by several “V” properties, and the number of these “V”s continues to grow. Variability and Veracity are added later to the initial triad – Volume, Velocity and Variety. The Value of using big data allows to find hidden patterns, correlations and connections in large data sets with the help of efficient processing.

The implementation of this last aspect related to Smart data depends on the ability to achieve an understanding of trusted, contextualized, relevant, cognitive, predictive, and consumable data at any scale, great or small. Smart data mines semantics from Big data and provide information that can be used to make decisions by solving problems related to the volume, speed, variety and reliability of data. Therefore user require automated analysis of data that has to be cleaned, transformed, structured and interpreted. Due to the increase in the amount and complexity of data, such analysis faces great challenges. Main result of this process is transformation of raw data into formalized knowledge. Intelligent information technologies supporting Smart data should provide the ability to semantic pre-processing and meaningful structuring information based on reliable, contextualized, relevant, cognitive, predictive, and consumable data at any scale.

The Smart Data strategy in processing of unstructured data into structured and semi-structured ones is aims to transform input "raw" data into machine-understandable, machine-processable, and machine-actionable instead of simply machine-readable data to generate information that can be used for communication, citation, transfer, rights management and reuse.

This process of knowledge generation through analysis and interpretation of deals with an intermediate stage of processing, which ensures the transformation of raw data into a form that simplifies their analysis and makes it more effective.

In a broad sense, such transformations correspond to the direction of research called Smart data. The term Smart data refers to the transition from unstructured mass data to knowledge through its intelligent processing, and one the elements of such processing are ensuring compatibility in the representation of information from different sources and semantic annotation of data sets by domain concepts. Methods and tasks of these transformations significantly depend on the subject domain where this analysis is carried out, on the raw data characteristics and on the demands for knowledge that are obtained from data.

Smart data is a powerful instrument that can be used not only for economic predictions but for research analyzed from the perspective of humanities [3].

The transformation of data into Smart data allows to determine what this data can be used for, that is, this approach is based on the well-known pyramid "Data-Information-Knowledge-Wisdom (DIKW) [4] that reflects the basic strategy of understanding the world by reducing information to its more significant elements. But the Smart Data approach is not a simple reproduction of DIKW, but its adaptation to the Big Data methodology, which involves the ability to learn by revealing previously unknown patterns, and not just by confirming or rejecting already existing hypotheses (“from unknown to unknowns”) [5].

An example of discovering the unknown with Smart Data is the research project "The Network Structure of Cultural History" [6] that is based on large datasets of the places of birth and death of more than 150,000 prominent people, reveal previously undocumented patterns of human mobility and culture, illuminating the formation of intelligent and cultural centers, the rise and fall of states and other influential factors that go beyond the scope of specific events or narrow time intervals.

All areas of Smart data use impose basic constraints on input raw information:

- data are represented in digital form (or means of their digitization on demand are proposed);
- data is stored for a long time on reliable medium;
- technical and legal possibilities for data access are ensured;
- there are certain (formal or informal) descriptions of these data – their purpose, origin, structure, etc., which can later be transformed into formalized metadata;
- some knowledge from domain of data analysis is available, and this knowledge can be used for pre-processing of the data.

Therefore, the problem of pre-processing is not solved by simple transferring information to electronic media and saving it. Thus, conventional scanning and text recognition is not enough, it is also necessary to associate (automated or with the help of a human expert) data elements with a specific metadata structures from pertinent domain and define technical possibilities right of access to information. The efforts of various experts, including scientists, engineers, business managers, and analysts, should ensure the use of various types of data (including Big data) through Smart Data strategies, even though the term Smart data itself may not be used in relevant research. For example, in the humanities, this direction is associated with the “Digging into Data Challenge” program [7] that refers to projects on the analysis of mainly unstructured information resources dating back to ancient times, but can also include structured sets of digitized data. Compared to previous years, the number of materials on multimedia research and other non-textual resources is increased significantly.

Interdisciplinary projects that take place at the intersection of the humanities and digital technologies show how best to capture data at large scale and in diverse formats to search for key insights, and to provide researchers with access to such data through new technological tools designed to provide "bigger smart data" (increase of smart data) and "smarter big data" (semantization of big data) [8].

Pre-processing and analysis of raw data should take into account various data features, such as:

- data source: data generated by human, by technical devices or automatically;
- type of represented information: text, video, audio, sensor data, etc.;
- aim of pre-processing results: human perception or automated processing with certain tools;
- volume of the data;
- data structuring type: unstructured, semi-structured, structured;
- data changeability: static or dynamic data ;
- possibility and means to separate data sets related to a certain task;
- data consistency.

Data pre-processing is aimed to allow researchers from different domains:

- to access and reuse large volumes of diverse data;
- to reveal patterns and connections that were previously hidden;
- to reveal the impact and significance of the qualitative and quantitative characteristics of the phenomena described by data, both in the real and in the virtual environment.

The transformation of raw data into Smart Data is based on the application of various technologies, such as cognitive computing, deep learning, machine learning, artificial intelligence, predictive analytics, Data mining, data science, the Internet of Things (IoT), text analysis, Semantic Web technologies and ontological analysis, knowledge graphs, contextual computing, Linked Data, natural language processing (NLP) and semantic search. These technologies are closely interrelated and intersect. For example, deep learning shows great potential for NLP; cognitive computing uses machine learning to find deep patterns (including those that are not statistical) in complex, unstructured, and streaming data. From the point of view of further research, it is important to note the ever-growing interest in the use of the Semantic Web standards and semantic search (Resource Description Framework (RDF) [9] and Ontology Web Language (OWL) [10]) in Smart data.

3. Main concepts of Smart data paradigm

If we consider the relation between various types and sources of raw data (including Big data) and Smart data in the context of different areas of scientific research and practical use, then we have to define more precisely main concepts used in considering of this approach.

The key concept that needs in conformance is the use of the term "data" itself: it is appropriate to consider not only digitized information in various formats, but also consider other sources of information that can be digitized in different ways and existing standards of data definition and representation. Such data have fundamental differences from data generated in the "digital universe" (like data generated by IoT devices and services): images, audio and video from mobile phones, video cameras, information from social networks, etc., which necessarily contain a minimum set of meta-descriptions.

For example, the reference model for the Open Archival Information System (OAIS) defines data as a reinterpreted representation of information in a formalized way suitable for transmission, interpretation or processing. In [11] data is defined as the representation of observations, objects or other entities that are used as evidence of phenomena for the purpose of research or science. The data accessed by the various libraries, archives and museums (LAMs) and other information institutions can vary greatly in type, nature and quality, and digitizing these data does not change some of these characteristics. The most difficult is processing of unstructured data contained in natural language documents and other information objects (textual or non-textual, digitized or not digitized), regardless of the chosen presentation format.

In order to transform unstructured data from non-digital media not only into machine-readable, but also into machine-processed resources that ensure their analysis and reuse, the Smart Data approach requires technologies for transforming unstructured data into structured and partially structured: image recognition, voice recognition etc. enriched by knowledge about pertinent domain.

It should be taken into account that even non-digitized data can contain some meta-information (formalized or non-formalized) that allows to transform data into partially structured ones. For example, LAM supports metadata for such semi-structured information objects as publications marked with tags according to the Text Encoding Initiative (TEI), and for such structured information objects as bibliographies, indexing databases, citation indexes. These data sets can be relatively small in volume and have less heterogeneity compared to Big data, but they are more clean, explicit, trusted, and value-added, because they are generated primarily on base of decision of human experts, not automatically. In addition, these data are usually open – they belong to resources that are freely available and non-commercial, and this expands the scope of their use. For example, structured data provided by LAM in the Linked Open Data library community that includes elements such as value, syntax, time, place, relevant domain, rules, user profile, can enrich Linked Open Data sets.

Therefore we have to consider the definition of other important concept in Smart data that is named “*Open data*”. Terms such as "data" and "open data" cover many meanings that depend on domain and purposes of their use, and usually do not have a common functional definition. In addition, in various fields of use, these concepts also have certain features. For example, in scientific research, the concept of "open data" contains more informal characteristics that are e determined both by the nature of the occurrence of such data (usually they are the result of a person's conscious intelligent activity) and by the forms of their use by other persons (this may be determined by various licenses and community rules). Open data generated in the "digital universe" more exactly define rules and licenses of data use by other applications (for example, the aspect of “personal data” access).

Therefore, their use in describing Big Data and Smart data technologies and are often mixed up. It is advisable to define more clearly what subsets of data can be considered as open ones, and what kind of data is analyzed, what data is the result of this analysis, and what are the limitations associated with increasing its volume, structuring requirements, and opportunities for application and reuse.

The data used and generated in scientific research have their own specificity, which depends on many factors. In the broadest sense, scientific data are objects that are used as evidence of phenomena for the purposes of research or science [12]. However, this definition does not address the issue of data units, the degree of data processing and the possibility of their sharing. Data that is useful to one researcher may be noise to another. Research data may differ from the information resources on which this research is based.

Another insufficiently defined term is “data set”. It is a collection of data related to some specific project, task or source of information that is intended to be shared with others. Examples of datasets for exchanges include private exchanges between researchers; datasets on Web sites of organizations or researchers; placement of data sets in archives, repositories, thematic collections or libraries; supplementary materials to journal articles. From various points of view the same collection of data can be considered as a data set of as an entire information object. For example, we can define some learning sample as a data set that consist of the array of characteristics of instances or as information object defined by the name into the library of learning samples.

Data sharing methods vary by domain, data type, goal of sharing, etc. The ability to identify, retrieve, and interpret shared data varies according to these methods [13].

Multiple data sets can be integrated at the "raw" or processed levels. Reuse of single data set in its raw form is difficult, even if adequate documentation and tools are available, because it is necessary to have information about how and in what form the data were collected, what decisions about data cleaning and analysis were used, etc.

Some interdisciplinary fields, such as environmental studies, combine datasets from multiple sources. In some cases, the primary scientific goal is to integrate disparate data sets into a single set for reuse. Merging data sets is much more difficult because a large amount of information about each data set has to be known in order to interpret it and trust it enough to draw conclusions.

"Open data" is one of the problematic terms in this field due to the variety and conditions and concepts used for it’s definition (such as "the fewest number of restrictions" and "the lowest possible cost") regarding the assignment of certain data to this category [14], and only some of these conditions are performed in different particular situation. The basic conditions for open data usually concern their legal and technical availability.

Examples of open data:

- repositories and archives (for example, GenBank, Sloan Digital Sky Survey),
- unified data networks (for example, World Data Centers, Global Biodiversity Information Facility; NASA Distributed Active Archive Centers),
- domain repositories (for example, PubMedCentral, arXiv),
- institutional repositories (for example, University of California eScholarship).

Data openness has different aspects. Public data repositories can allow authors to retain copyright and control over the data they have submitted. Some data is open, but it can be interpreted with proprietary software. The data can be created using open source software, but a license is required to use the data. Open data repositories can have long-term sustainability plans, but many of them depend on short-term grants or viable business models. In addition, keeping data open for long periods often requires ongoing investment.

A promising new development to address the open data challenge is the FAIR standards – Findable, Accessible, Interoperable and Reusable data [15]. These standards apply to repositories that store data. The FAIR standards are adopted by a group of stakeholders to ensure Open science, and they bring together all parts of the research object, from the code to the data and the tools for their interpretation. This approach is developed for scientific information but can be adopted for reuse of any other data.

An important aspect of data openness is ensuring the possibility of their reuse [16]. At the same time, we need to understand the difference between use and reuse of data. In the simplest situation, data set is collected by one person (or group of persons) for a specific purpose, and the first use of this data is executed by that person. If the same person returns to the same data set later for the same or a different task, this is also usually considered as use. When this data set is used for another task by someone else (for example, from a repository), then such action is usually considered as data reuse. A separate data set can be reused for another purpose if it is supported by appropriate contextual information and tools. Research replication is an example of independent reuse of a data set.

An important factor in data reuse is the use of representation standards. Data published in formats that meet community standards can be analyzed with the help of available tools and combined with other data in those formats. Data integration and reuse is much more difficult in areas where standards are unavailable or less formalized.

4. Proposed methodology

The level of data structuring has big influence on the complexity of acquisition of knowledge from it. In this work we propose to use semantic markup as a base of Smart data: semantic tags can become an instrument of data explicit structuring, and interpretation of this structure makes data analysis more productive. All elements of structuring can simplify data analysis if they are pertinent with goals of analysis and can be interpreted by analytic means.

Semantic markup is one of the common approaches that adds structure to different types of data. Most often, this approach is applied to natural language documents, but it can also describe various complex information objects with multimedia elements. Various models and software realizations of markup differ significantly by expressiveness, understandability and complexity.

The structuring of the USD ensures the creation of metadata for individual information objects (IOs) and for data sets, as well as the marking of content with tags that connect it with the concepts of the corresponding domain. Metadata describes the attributes of the IO. Therefore structuring of USD by semantic markup is one of ways of Smart data transformation.

Effectiveness of semantic markup depends on:

- understandability of the markup language and the use of standardized notations;
- sufficient expressiveness of the marking language;
- availability of tools for processing of marked data (visualization, correctness check, automation of editing, etc.);
- expressiveness of query language for semantically marked data and its support in various technological environments;
- markup extensibility;
- possibility of integration of semantic markup with external knowledge bases and support of open knowledge presentation standards.

Semantic markup is a way of data structuring that links information object and its elements with concepts of some domain. Elements used as markup tags depend on specifics of information object. The set of these tags can be fixed (as HTML) or dynamic (as XML). Some markup languages are universal (such as XML Schema or RDF Schema), and other ones are used only for specific types of information objects (such as OWL-S for web services).

Another important difference among markup languages deals with their semantic interpretation. For example, XML Schema or Wiki markup has no associated semantics, while RDF Schema, Semantic MediaWiki (SMW) markup and DAML+OIL include it. Semantics provides standard and unified way for interpretation of the language elements and can be used by reasoners foe inferences of new knowledge from given data markup. Non-semantic markup can define structural elements of data such as titles, sub-items, links, etc., but does not represent their meaning.

We can consider semantic markup as an expansion of metadata because semantic markup tags can describe both file with data in general and come separate elements of its content.

Every semantic markup language is defined by the non-empty finite set of tags and by rules of their use. In [17] characteristics of semantic markup languages that can be used for their comparison are analyzed. The main of them are:

- Context: possibility to express the different data contexts for interpretation of tags.
- Subclasses and properties: possibility to express the meaning of relations between marked objects, their properties and classes.

- Primitive data types: possibility to define the type of data constants (such as strings, number, links, dates, etc. and their complex combinations) used as elements of markup.
- Instances: possibility to define objects as individuals of some classed.
- Property constraints: possibility to define range and domain of object attributes, their possible values and cardinality constraints.
- Property values: possibility to define values of attributes linked with tags, including a default value or a set of possible choices.
- Negation: possibility to define statements as a negation, conjunction and disjunction of other statements.
- Inheritance: possibility to indicates the constraints and values of subclasses by properties of their parent classes. Multiple inheritance allows inheritance from multiple parent classes.
- Definitions: possibility to describe necessary and sufficient conditions for class membership.

Table 1
Characteristics of markup languages (fragment)

	XML Schema	**RDF Schema**	Wiki	SMW
Context	+	+	+	+
Subclasses	+	+	+	+
Properties:	+	+	-	+
Property range	+	+	-	+
Property domain	+	+	-	+
Property cardinality	+	-	-	+
Primitive data types	+	-	-	+
Instances	+	-	+	+
Property values	+	+	-	+
Negation	-	-	-	-
Definitions	-	-	+	+

Determining the structure of the NSD is a complex scientific problem that scientists have been paying attention to for a long time [18]. Another aspect of this problem deals with selection of pertinent knowledge that defines the NSD markup structure. The choice of the set of tags for semantic markup has the greatest impact on effective retrieval and reuse of data irrespective of ways of structuring – by a human expert or automatically. Both the correctness of the IO classification and the correct definition of attribute values are important. Thus, the structuring process itself is divided into two stages:

An example of numbered list is as following.

1. Selection of the non-empty set of categories and markup tags that allow to determine the IO structure and its relation with other objects and groups of objects that is based on knowledge about IO domain;
2. Semantic markup of IOs by selection of subset of categories and linking of IO content elements with markup tags and determining of appropriate attribute values.

Semantic markup uses tags that have explicitly defined semantics formalized by knowledge representation means such as ontologies, conceptual graphs, thesauri, etc.

The first stage determines the expressiveness of IO structuring and should be pertinent to the aim of data preprocessing, and the second one provides the possibility of using this markup.

The set of markup tags connects the raw data with the background domain concepts and relations of the corresponding domain exported from some external knowledge base. This set depends on the task specifics and can contain both all these concepts and their non-empty subset. The increment of used concept number provides the more clearly data structuring, but it makes its analysis more difficult.

Many semantic markup schemas use ontologies [19] as external knowledge sources that, in the general case, contain classes, instances of classes and relations between them, as well as axioms that determine the rules of admissible combination of these elements.

Formal model of domain ontology O_{domain}most generally is defined as a triplet:

$$O_{domain} = < X, R, F >, \qquad (1)$$

where

- X *is* a finite set of domain concepts divided into a set of classes T_{cl} and a set of instances of classes T_{ind};
- R is a finite set of domain relations between domain concepts from T;
- F is a finite set of axioms and interpretation functions for concepts and relations of O.

In practical use the elements and the structure of domain ontologies can be defined more precisely according to task specifics. For example, some formal models distinguish various subsets of domain-specific relations or relation properties, and various special cases of ontologies have additional restrictions.

One of special cases of ontology that can be used for purposes of semantic markup is task thesaurus that reduces knowledge structure for easier processing according to task description but can use this information about domain for its generation. Task thesaurus can be considered as a special case of ontology $Th \subseteq O$ that contains collection of the domain terms. Formal model of thesaurus is based on formal model of ontology (1):

$$Th = < T_{th}, R_{th}, I >, \qquad (2)$$

where

- $T_{th} \subseteq X$ is a finite set of the terms;
- $R_{th} \subseteq R$ is a finite set of the relations between these term;
- I represents an additional information about terms (this information depends on specifics of thesaurus goals and can contain, for example, weight of term or its definition).

Task thesaurus has the simpler structure because it is not include ontological relations (all important for task information about relations is used for construction of T_{th}), but includes additional information about every concept – it's weight $w_i \in W, i = \overline{1,n}$ defined by task description. If $w_i = 0$ then concept is not included into T_{th}. Therefore, formal model of task thesaurus is defined as set of ordered pairs $Th_{task} = < (t_i \in T_{th}, w_i \in W), \varnothing, I >$ with additional information in I about source ontologies.

We can define formally various characteristics and restrictions of ontology-based knowledge sources and link them with corresponding groups of tasks (see Figure 1).

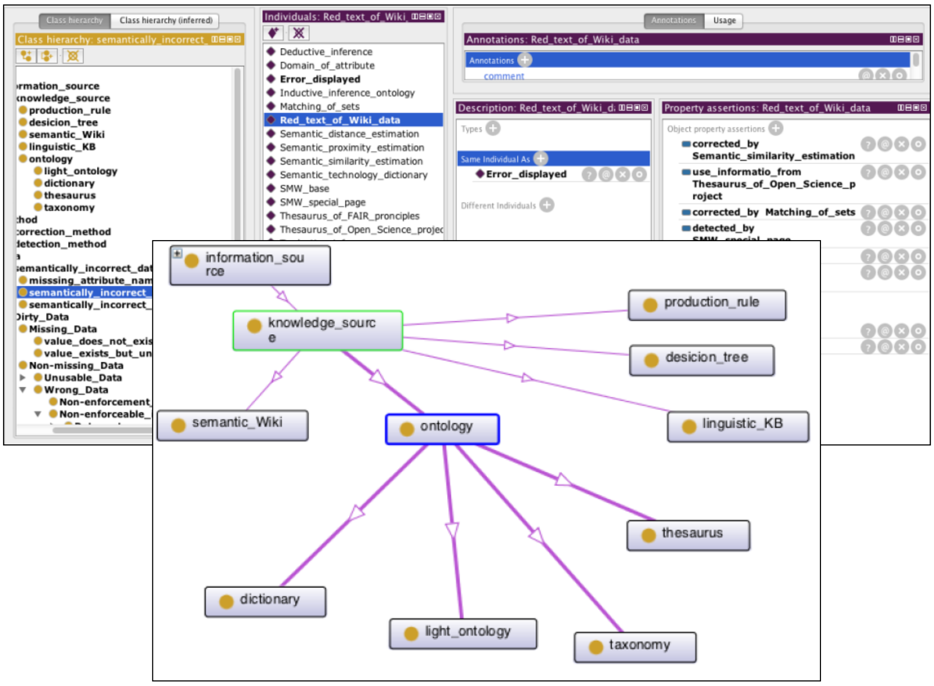


Figure 1: Characteristics and restrictions of ontology-based knowledge sources

Use of ontologies is based on such their characteristics:

- ontological representation of domain is an explicit specification of the conceptualization that provides unambiguous interpretation of their semantics by different users;
- ontologies have a wide range of knowledge representation expressiveness;
- development of ontologies is based on descriptive logics that supports theoretical ground of their expressiveness and processing time;
- ontologies are widely used for modeling of various domains, and therefore a lot of domain ontologies are accessible from repositories;
- ontological analysis is supported by a large number of standards and instrumental tools for creating and processing of domain ontologies;
- ontologies can be integrated with various Semantic Web applications.

The use of existing ontologies allows not to reanalyze the domain structure and to reuse the previously acquired knowledge of experts. But the success of creating a semantic markup depends significantly on the relevance of the chosen ontology to the tasks for which the markup is created and the ways in which it is used for markup.

We have to take into account both the volume of the selected ontology (number of classes, individuals and relations) and its complexity (number of relations between individuals and axioms that determine rules of their use) to define ontology pertinence for semantic markup goals. For example, the use of a highly specialized medical ontology is appropriate for markup of educational materials for students, but not convenient for school textbooks.

Large number of concepts used as the markup tags makes the search more accurate, but reduces the number of relevant answers. Selected domain ontology is a main source of information for search queries and their parameters, hypotheses for machine learning and other operations on the analysis of semantically marked data, that is, knowledge that is not reflected in this ontology is much more difficult to discover and use in the future. For example, if the ontological model does not contain the relation "semantic similarity" to determine the relation between the concepts of domain, then the detection of such similarity requires much more calculations to model this relation.

If domain of marked IOs has already commonly used standards (national, international, industrial, etc.) or generally accepted community agreements (such as metadata schemas) for terminology unification, then they must be taken into account. If these standards are formalized in the form of ontologies, then these agreements should be used in generation of semantic markup tags.

If these standards are represented in other forms (tables, natural language documents, dictionaries and taxonomies, etc.), then appropriate ontologies are created on base of them – manually or automatically (this subtask is beyond the scope of this study and should support the integration and coordination of knowledge from different sources).

Ontologies that reflect different domain sub-areas can be integrated (integration of ontologies with disjoint sets of concepts causes low problems) through the top-level ontology and use this entire set of ontologies as a base of markup.

In practice, choose of external ontology as source of foreground domain knowledge for semantic markup takes into account the following main parameters:

- expressive power of ontology (from simple dictionaries and taxonomies to "heavy" ontologies) pertinent to markup purposes;
- volume of ontologies (number of classes, number of instances of classes, number of attributes of instances of classes, number of relations, etc.) that provides processing in satisfactory time;
- presence and correct representation of ontological concepts and relations that are fundamentally important for the analysis of semantically marked data;
- correspondence with natural language of marked data;
- relevant level of ontology specialization.

If all available ontologies do not meet these, we have to build new ontology based on one or more such ontologies that more fully meets the conditions of its use. Such situation is possible, if semantic markup deals with new, dynamic or very specific sphere, is oriented on users with specific information needs and beliefs or marks content represented by natural language that differs from language used in existing ontologies.

In addition, it is necessary to take into account such situations when development of the semantic markup of information resource precedes ontology development, and ontology that formalizes markup structure is generated on base of marked data. Ontology created in such way has a lot of functional restrictions (such as characteristics of classes that cannot be acquired by markup analysis), but it can also be populated and improved in the future according to the needs of users.

If we find ontology with greater expressive power than is required for resource marking, then it is advisable to build a simpler ontological structure that contains necessary background knowledge but reduces redundant elements, characteristics and axioms. Building a simplified ontology requires additional efforts, but usually such reduction saves of time in each subsequent ontology request. In addition, this transformation of ontology can be made semi-automatically and does not require a significant involvement of domain specialists. But it is important to understand that this new simpler ontology has other structure than the initial one and is not its sub-ontology. Therefore if initial external ontology is improved by its developers than processing of these changes demands new processing for transformation into reduced ontology.

In ontology has insufficient expressive power (for example, it does not contain a certain group of relations, or the limitations and areas of meaning are not defined for some classes), then such an ontology can be improved after consulting with the domain specialists without other changes in its structure. Then found ontology can be considered as a sub-ontology of new ontology that is more complete. If ontology has a larger volume than is required for the task of semantic marking, then we can use its certain sub-ontology. The easiest way is to remove those instances of classes that are not relevant – such reduction does not require changes in the structure of the ontology. In this case, we only need to check that the remaining instances do not use deleted elements as attribute values.

It is more difficult to remove unnecessary relations and classes – we need to check that the remaining instances do not also use the removed elements as attributes and in scope and definition descriptions. The removal of axioms makes impossible automatical check of some characteristics of the ontology (for example, the disjointness of two classes or the set of acceptable attribute values), but it can significantly simplify the algorithms for its processing.

If checking the presence of concepts and relations in the ontology indicates the absence of some elements that are fundamentally important for analysis of the semantically marked data, but found ontology meets other requirements and contains a significant number of required elements, then it is advisable to add these concepts or relations to the initial ontology and check the correctness of new ontology. Similarly, we can add the desired attributes to some classes (if ontology contains appropriate classes).

In each specific case, we have to determine which solution requires more effort – modification of the found ontology or creation of a new one.

To choosing the level of the ontology specialization we have to analyze possible queries to the semantic markup should satisfy and data parameters that should be defined. If ontology contains a large number of concepts that are not used for the needs of semantic markup, then it is advisable to use a sub-ontology with the necessary classes, but maintain its compatibility with the original ontology for the possibility of expanding the set of tags (for example, when new data appears). This solution is similar to processing of ontologies with big volume where number of elements is caused by other reasons different from specialization level.

It is clear that the optimal choice of background knowledge source for semantic markup is an ontology that contains all the necessary domain knowledge and does not contain any extra elements. But such a situation is possible only when the ontology is created specifically for the purposes of semantic markup or if it is generated from already marked data in order to formalize the semantics of this markup for further use. An example of such ontology is the Wiki-ontology (its model and construction algorithm are discussed in more detail in [20]). Ontologies built for other purposes usually cover a wider scope than is required for semantic markup. Then there is a need to reduce the ontology, that is, to build another ontology using knowledge from the found one. If we find smaller or too specialized ontologies then we need to either merge them (and then reduce) or to populate them manually, and both of these operations require a lot of additional work with the involvement of domain experts. Thus, quite often we need to create a new ontology that is smaller then initial one, contains all the necessary elements and has an easier-to-process structure based on the external ontology containing the domain knowledge. Therefore we propose some models and algorithms that can be used for automation of this process.Different approaches to creating markup based on ontologies allow to use various elements of the ontology, and therefore such markup has different expressiveness, and the process of its creation becomes more or less complex description and support much or more simpler processing algorithms. The following approaches are most common:

- on the basis of domain dictionaries that contain only ontology classes and class instances;
- on the basis of ontology classes and relations that allow to formalize the semantics of links between individual IOs in the markup [21];
- on the basis of task thesauri generated by domain ontology and markup task .

All these approaches can be formalized on the basis of the formal model of the ontology used for semantic markup.

Main steps of ontology generation for semantic markup are:

- retrieval of external domain ontology that contains pertinent to semantic markup aims foreground knowledge;
- transformation of external ontology according to semantic markup requirements (reducing, merging, population, etc.);
- transformation of domain knowledge from generated ontology into other ontological representation that is more convenient for markup procedures.

5. Use of semantic Wiki markup for data structuring

Wiki-resources are the popular examples of the Web-oriented markup. If ordinary Wiki-resources use simple links between Wiki-pages as markup tags, then semantic Wiki-resources extend the expressiveness of the markup by the semantic definitions of such links: they explicitly define the relations between such pages by some domain concepts.

Currently, many dynamic Web-oriented resources are creating in process of the joint activity of users on base on Web 2.0 technologies [22]. Wiki technology [23] is one of successful Web 2.0 platforms that support mechanisms for collaborative processing of the large-scale Web content. MediaWiki [24] is one of the common implementations of Wiki technology that is used by such popular resources as Wikipedia, Wikibooks, Wiktionary, and Wikidata.

Various semantic extensions of Wiki technology are aimed to add meanings to Wiki resource elements and to make them suitable for automated processing and knowledge-level analysis. They differ in the expressiveness of the markup language and the capabilities that can be used for data analysis. Many of them are based on the standards of the Semantic Web project [25]. Such extensions allow to define and find IOs with a complex structure that are typical for a certain domain [26].

The above analysis of semantic markup means and their characteristics shows the importance of:

- the expressiveness of the markup language,
- the possibility of its interpretation by humans and other applications,
- the availability of means and methods to transform external knowledge sources of into markup elements, and
- the quality of software tools for creating and practical use semantic information resources based on such markup.

All these requirements are met by the semantic extension SMW of the Wiki technology and the language of semantic markup and semantic queries based on this markup implemented in it. SMW is a semantic extension of MediaWiki (www.mediawiki.org/wiki/MediaWiki) that provides intelligent organization and search of heterogeneous content [26]. In addition, SMW-based information resources meet FAIR requirements and can be scaled to represent large-volume and complex content. It should be noted that the application of widely known Wiki technology significantly simplifies the practical use of such resources for a wide range of people. Formal models, representation languages, processing methods and software tools already exist for them. SMW provides a structured representation of knowledge and the ability to search for it at the content level. But if marked data (such as encyclopedias of the national level) has a large volume and a complex structure, then built-in possibilities of SMW are not enough and we have to use modern methods of management of distributed knowledge.

If traditional Wiki-resources use simple links to other Wiki-pages as markup tags, then semantic Wiki-resources extend the expressiveness of the markup by defining the semantics of such links: they explicitly define the relation between such pages. Knowledge of an arbitrary external ontology transformed into a Wiki-ontology in terms of semantic Wiki-technology such as categories, semantic properties and their values, templates of typical IOs, etc.

The expressiveness of a Wiki ontology has some limitations because such an ontology contains only the knowledge that can be obtained directly from the Wiki markup and expressed by means of markup language. For example, it cannot define characteristics for object properties and data properties, such as equivalence and the possibility of intersection. In many cases, semantic extensions of Wiki technologies have built-in means for automatic or automated generation of such ontologies. SMW supports automatical generation of Wiki ontology for arbitrary collection of Wiki pages. On the other hand, the formation of the Wiki-ontology (or at least its structure) can precede the development of the Wiki-resource itself. In this case, a certain reference ontology created by experts and knowledge engineers defines the basic domain concepts and relations between them.

Wiki-ontologies with low expressiveness can be generated by non-semantic Wiki markup. It contains only information about page categories and links between them without defining semantics. Wiki-ontology is a special case of ontology and its formal model can be defined on base of constraint of model (1) for non-semantic and semantic Wiki resources.

Formal model of Wiki-ontology $O_{wiki\_no\_semant}$ for a non-semantic Wiki resource contains the following components:

- $X = X_{cl} \cup X_{ind}$ is a non-empty set of ontology concepts, where X_{cl} is a set of classes that coincides with the set of Wiki categories represented in the selected set of Wiki pages, X_{ind} is a set of class instances created as a union of the names of selected pages $P = P_{user} \cup P_{template} \cup P_{spec}$, where P_{user} is a set of pages created by users, $P_{template}$ is a set of pages describing Wiki templates, P_{spec} is a set of other special pages explicitly selected for ontology generation (for example, semantic search pages);
- $R = L \cup \{ r_{ier\_cl} \} \cup \{ r_{class\_individual} \}$ is a set of relations between elements of the ontology, where $L = \{"link"\}$ is a one-element set that describes a link from one Wiki page of this resource to another one; r_{ier_cl} is a hierarchical relation between the categories of the Wiki resource, which is determined in the process of creating new categories, $r_{class\_individual}$ is a hierarchical relation between the categories and the pages of the Wiki resource assigned to these categories;
- $F = \{ f_{equ} \}$ is a one-element set containing an equivalence relation between Wiki pages that can be used for logical inference in the ontology that connects reference Wiki pages.

Other elements of the ontological model of this Wiki-ontology are represented by empty sets.

The formal Wiki ontology model for semantically enriched Wiki resources O_{s_wiki} is more complex in comparison with Wiki ontology of non-semantic Wiki resources (such as Wikipedia) and includes a number of elements related to semantic properties [27]:

- set of Wiki pages X is enriched by the set of pages of semantic properties P_{sem_prop} (some of them are semantically defined links to other Wiki pages $P_{sem\_prop\_page} \subseteq P_{sem\_prop}$, and others link pages to values of other data types);
- set of relations $R = r_{ier\_cl} \cup \{ r_i \} \cup R_{s\_prop}$ is enriched by relations related to the semantic properties of Wiki pages by domain-specific semantic properties $R_{s\_prop} = \{ r_{s\_prop_j} \}, j = \overline{1,m}$ with type “Page” that link Wiki pages by semantically defined characneristics;
- $r_{ier\_cl} = r_{ier\_categor} \cup r_{ier\_property}$ is enriched by relations $r_{ier\_property}$ that defines hierarhical relations of semantic properties;
- T is a set of types for values of semantic properties.

Use of the Wiki ontology elements for semantic markup is unambiguous and is based on one-for-one correspondences (Table 2).

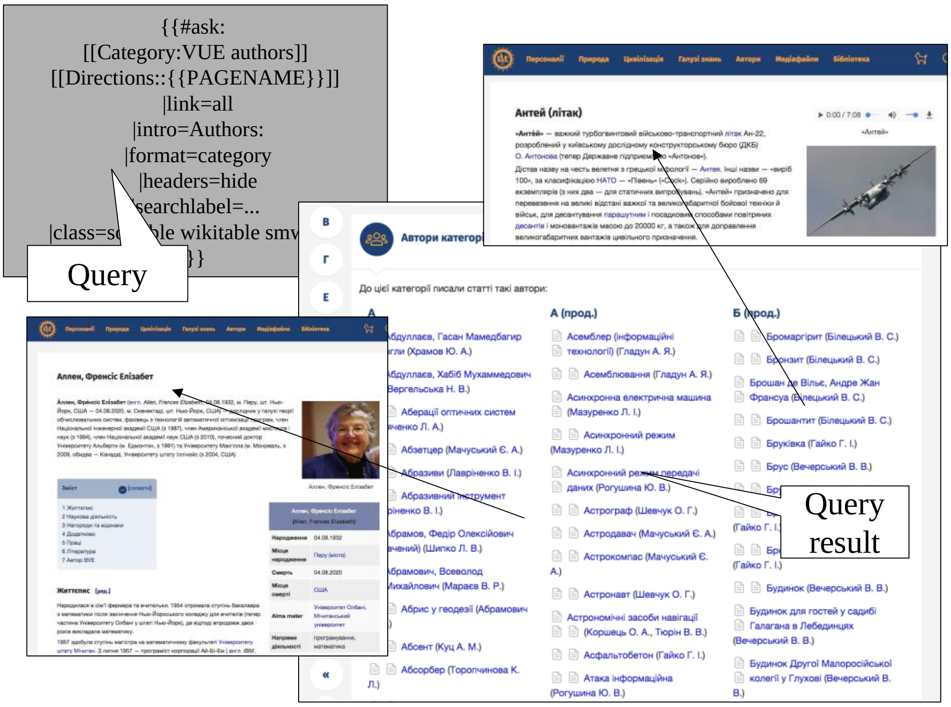
Table 2
Correspondences of the Wiki ontology elements with semantic markup elements

Wiki ontology	Markup element
X_{ind}	Page name
X_{cl}	Category name [[Category:Category name]]
$P_{template}$	Template names
$L=\{"link"\}$	Link between Wiki pages [[Page name|Description]]
r_{ier_cl}	Relation between categories
$r_{class\_individual}$	Relation between categories and individual pages
L_{sem_prop}	Semantic properties with type “Page” [[Relation |Page name|Description]]

Tags of semantic markup are based on these elements represented according to SMW rules.

6. Practical approbation

We use proposed above ontology-based models and methods of semantic markup for development of the portal version of the Great Ukrainian Encyclopedia e-VUE [28]. It uses MediaWiki [29] version 1.34.0 and the Semantic MediaWiki semantic plug-in version 3.1.5. Semantic markup supports built-in semantic queries that integrate content of different Wiki pages about various typical IOs. For example, such queries create automatically lists of author articles (see Figure 2), new articles from selected category, current moderator of scientific spheres, etc.



```
{{#ask:
[[Category:VUE authors]]
[[Directions::{{PAGENAME}}]]
|link=all
|intro=Authors:
|format=category
|headers=hide
|searchlabel=...
|class=sc ble wikitable smv
}}
```

Figure 2: Use of semantic markup for integration of VUE content

Now semantic templates are used for unified input of structures information, but content of the Wiki pages can be enriched by other tags from the Wiki ontology without use of templates (this approach is used for more specific IOs of resource or for IOs with non-typical attributes) [30].

Initial elements of ontological schema are created before development of this Wiki portal, but later it was enriched and populated with use of specialized domain ontologies and dictionaries.

Semantic markup of e-VUE is supported by semantic templates of typical IOs (such as persons, cities, countries, organizations, seas, rivers, etc.) that use domain concepts as attributes. Semantics, possible categories of values and relations between individuals are formalized by Wiki ontology of this resource (see Figure 3). Templates help to input correct attributes of semantic markup elements.

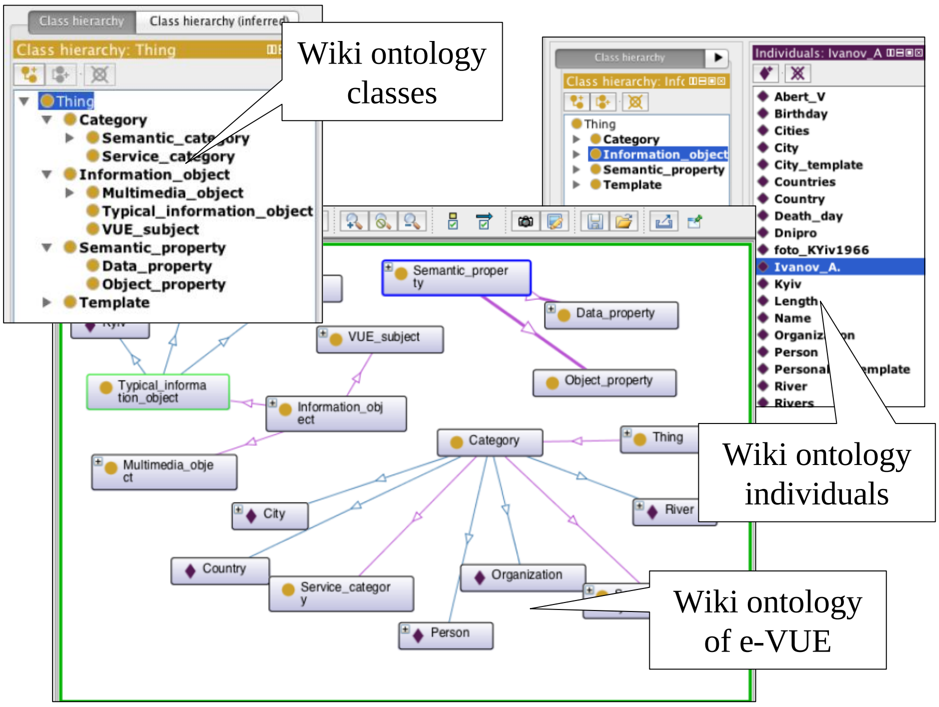


Figure 3: Wiki ontology of e-VUE (fragment)

7. Conclusion

We analyze approaches to transformation of raw data into source for analysis and knowledge acquisition named Smart data and consider possibilities of their integration with ontological analysis. One of directions of Smart data deals with data structuring, and we propose to make such structuring on base of semantic markup where ontologies are used as a source of domain knowledge.

In general case ontologies need in complex means of their processing, and therefore we propose some special cases of domain ontologies that reduce domain knowledge according to goals of markup of Semantic MediaWiki resources. In this work, we formalize models for such special cases of ontologies as Wiki ontology and task thesaurus.

Semantic extensions of Wiki combine the benefits of traditional Wiki technologies (easy editing and collaboration) with the advantages of semantic processing. We consider preferences of this technology and problems of its practical use on example of the portal version of Great Ukrainian Encyclopedia that integrates heterogeneous information from various fields of sciences and includes a big number of typical information objects with heterogeneous elements.

In future, we plan to consider integration of semantic Wiki markup with metadata standards used by open ontology repositories an e-libraries that can be used as external sources of knowledge and structure of various domains.

8. References

[1] P. B., Seel, Digital universe: The global telecommunication revolution. John Wiley & Sons. (2022).

[2] A. Souifi, Z. C. Boulanger, M. Zolghadri, M. Barkallah, M. Haddar, From Big Data to Smart Data: Application to performance management. IFAC-PapersOnLine 54(1) (2021), 857–862.

[3] M. L. Zeng, Smart data for digital humanities. Journal of data and information science 2(1) (2017) 1-12.

[4] J. Hey, The data, information, knowledge, wisdom chain: the metaphorical link. Intergovernmental Oceanographic Commission, (2004) 26(1), 72–94.

[5] S. Sharifi Noorian, S. Qiu, U. Gadiraju, J. Yang, A. Bozzon, What Should You Know?, A Human-In-the-Loop Approach to Unknown Unknowns Characterization in Image Recognition, in: Processings of the ACM Web Conference 2022, (2022): pp.882–892. doi: https://doi.org/10.1145/3485447.3512040.

[6] M. Schich, C. Song, Y. Ahn, A. Mirsky, M. Martino, A. Barabási, D. Helbing, A network framework of cultural history. Science 345(6196) (2014) 558–562. URL: www.yongyeol.com/papers/schich-history-2014.pdf.

[7] Digging into Data Challenge, 2020. URL:https://diggingintodata.org/.

[8] C. Schöch, Big? Smart? Clean? Messy? Data in the Humanities?, Journal of the Digital Humanities (2013) 2(3). URL: https://opus.bibliothek.uni-wuerzburg.de/files/12949/059_Schoech_JDH.pdf.

[9] RDF. URL: www.w3.org/RDF.

[10] OWL 2 Web Ontology Language Document Overview. W3C, 2009. URL: http://www.w3.org/TR/owl2-overview/.

[11] I. Pasquetto, B. Randles, C. Borgman, On the reuse of scientific data, 2017, https://escholarship.org/content/qt4xf018wx/qt4xf018wx.pdf.

[12] C. Borgman, P. Darch, A. Sands, M. Golshan, The durability and fragility of knowledge infrastructures: Lessons learned from astronomy. Proc. of the Association for Information Science and Technology. V. 53, (2016) 1–10. DOI: https://doi.org/10.1002/pra2.2016.14505301057.

[13] C. Palmer, N. Weber, M. Cragin, The analytic potential of scientific data: Understanding reuse value, in:Proceedings of the American Society for Information Science and Techn. 48(1) (2011) 1-10.

[14] I. Pasquetto, A. Sands, P. Darch, C. Borgman, Open data in scientific settings: From policy to practice, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, (2016): 1585-1596.

[15] FAIR_data. URL: https:// en.wikipedia.org/wiki/FAIR_data.

[16] S. Leonelli, Packaging small facts for re-use: databases in model organism biology, in: How well do facts travel, (2010) 325–348. DOI: https://doi.org/10.1017/CBO9780511762154.017.

[17] Y. Gil, V. Ratnakar, A Comparison of (Semantic) Markup Languages, in: Proceedings of FLAIRS Conferenc, 2002, pp.413-418.

[18] P. Buneman, S. Davidson, M. Fernandez, D. Suciu, Adding structure to unstructured data, in: Proceedings of International Conference on Database Theory, 1997, pp.336-350.

[19] T. R. Gruber, A translation approach to portable ontology specifications. Knowledge Acquisition, (1993), 5:199-220.

[20] J. Rogushina, A. Gladun, Task Thesaurus as a Tool for Modeling of User Information Needs, in: New Perspectives on Enterprise Decision-Making Applying Artificial Intelligence Techniques, Springer, Cham, (2021) 385-403. DOI: https://doi.org/10.1007/978-3-030-71115-3_17.

[21] J. Rogushina, I. Grishanova, Ontological methods and tools for semantic extension of the MediaWiki technology, CEUR 2866 (2021) 61-73. URL: http://ceur-ws.org/Vol-2866/ceur_61-73Rogushina6.pdf.

[22] J. E. Pelet, Handbook of Research on User Experience in Web 2.0 Technologies and Its Impact on Universities and Businesses, IGI Global, 2020. URL: http://kmcms.net/Doc/Call/user-experience/about.html.

[23] Y. Koren, Y. Working with MediaWiki. San Bernardino, CA, USA: WikiWorks Press, 2012.

[24] M. Völkel, M. Krötzsch, D. Vrandecic, H. Haller, R. Studer, Semantic wikipedia, in: Proceedings of the 15th international conference on World Wide Web, 2006, pp. 585-594.

[25] P. Hitzler, A review of the semantic web field, Communications of the ACM, 64(2) (2021) 76-83.

[26] P. Andon, J. Rogushina, I. Grishanova et al, Experience of Semantic Technologies Use for Development of Intelligent Web Encyclopedia, CEUR 2866 (2021) 246-259. URL: http://ceur-ws.org/Vol-2866/ceur_246-259andon24.pdf.

[27] J. Rogushina, A. Gladun, Semantic processing of metadata for Big Data: Standards, ontologies and typical information objects, CEUR 2859 (2020) 114-128. URL: http://ceur-ws.org/Vol-2859/paper10.pdf.

[28] Great Ukrainian Encyclopedia e-VUE, 2022. URL:vue.gov.ua.

[29] MediaWiki, 2021. URL: www.mediawiki.org/wiki/MediaWiki.

[30] J. Rogushina, Semantic Wiki resources and their use for the construction of personalized ontologies, CEUR 1631 (2016) 188-195.