

Semantic Profile of Corporate Web Resources

Viacheslav Zosimov, Oleksandra Bulgakova and Valeriy Pozdeev

V.O. Sukhomlynsky National University of Mykolaiv, Nikolska 24, Mykolaiv, 54000, Ukraine

Abstract

The article presents a semantic profile of a corporate web resource, developed on the basis of the most common dictionary of semantic markup Schema.org. Based on the analysis of the structure and information content of 500 corporate sites, their general structure was compiled. This structure was compared with the schema.org ontology, missing classes were added to fully describe the developed structure of the corporate web resource. As a result, a general ontology of a corporate web resource was developed.

Keywords ¹

semantic markup, search agents, intelligent information search, semantic web, corporate web site.

1. Introduction

The development of the concept of the semantic web has become another evolutionary step in the development of the global network. The information posted on the Internet is easy for a person to understand. The semantic web was developed to make the information suitable for automatic analysis and synthesis of conclusions [1]. Despite the obvious advantages of using this technology, it has not become widespread in the web environment. Significant results have been achieved in the development of models of semantic markup of online stores as the main tool for e-commerce. Good Relations [2] has been used as a standard for micro-marking of e-commerce products since 2008, which provides the ability to specify special properties for:

- companies - contact details, location, logo;
- store - address, opening hours, phone;
- specific product - product category, brief description, code, methods of payment and delivery, etc.

At the same time, very little attention is paid to the electronic market for services, namely, structural and semantic standards for the development of corporate web resources. Only a small percentage of web resources are developed using semantic markup standards. This situation is a consequence of the problems of practical implementation, existing from the very beginning of the semantic web concept, and the peculiarities of the web resources development market:

1. Lack of publicly available means of viewing and direct use of information provided by web resources in the Semantic Network. Existing projects differ and do not go beyond research departments [3].

2. The visibility of specific standards for the semantic design of corporate web resources, that is, the visibility of the tools in the development of web resources with integrated semantic design.


3. The visibility of web developers will devote an hour to mastering new technologies with a glance at the visibility of tools for the interaction of the user with this technology. It's simpler, seemingly dumb to the senses of the developers, it takes an hour to root web resources from the semantic layout of the schema.org standards, but there's no practical tool to use any kind of tools in order to use the keys to correct the information. The Google company has provided tools for micro-formatting of web

Information Technology and Implementation (IT&I-2021), December 01–03, 2021, Kyiv, Ukraine

EMAIL: zosimovvv@gmail.com (V. Zosimov); sashabulgakova2@gmail.com (O. Bulgakova); pozdeev1405@gmail.com (V. Pozdeev)
ORCID: 0000-0003-0824-4168 (V. Zosimov); 0000-0002-6587-8573 (O. Bulgakova); 0000-0003-1224-7329 (V. Pozdeev)



© 2022 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

resources, and there are only a few semantic tags to add to it, and it will only allow the thematic directness of the web resource and the visualization of contact data [4].

4. Development of a web resource based on the schema.org semantic layout standard is a non-trivial task and an extra hour of development for a project in a design. Apparently, it is significant that the development of the web resource for the deputy is growing. Exciting about the provision of furnishings, the distributors must complete the important arrangement from the economic point of view, the need to increase the budget for the development of the project in the design of the development.

The advent of semantic markup standards opens up broad prospects for the use of intelligent data search tools on the Internet - search agents. According to Stephen Haag's classification [5], these are data mining and analysis agents working in data warehouses. In the development of metasearch engines, search agents are used to analyze and process the results of the search engine. This direction is developing rather slowly due to the lack of search engine development tools available to a wide range of users. To ensure the effective operation of the technology it is necessary to explore the content of corporate web resources, build their structure and semantic profile to describe the built structure.

2. Problem Statement

The use of semantic markup is the most effective way to adapt the information presented on web resources to machine processing. The most common and largest dictionary of semantic markup is Scema.org. It has 792 classes and 1447 properties [6]. On the one hand, the large number of presented classes makes this dictionary a universal tool for marking different types of web resources, but on the other hand, it significantly complicates the process of marking small corporate web resources due to the need to study the whole hierarchy of classes, internal interclass connections and properties.

In recent years, the trend of using semantic dictionaries has changed somewhat. As with the development and addition of new classes, the practical use of the general dictionary schema.org has become quite time-consuming, mechanisms have been developed for the integration and use of specialized dictionaries based on Schema.org [7]. Specialized ontologies are developed for a more detailed description of data from a specific industry. There are two types of specialized dictionaries:

1. Internal - published as part of the schema.org project with its own structure, which is usually serviced by a separate project team.
2. External - exist separately from the main dictionary. Developed and maintained by external organizations not affiliated with schema.org, however, they are important elements in expanding the overall structure of the basic vocabulary. In the future, external ontologies are integrated into the system of classes of the basic dictionary.

The integration of external dictionaries takes place in several stages. It is first reviewed in the schema.org community. In case of a positive assessment, a working group is formed to further coordinate the process.

From the above it follows that the task of developing a specialized ontology for corporate web resources, which includes the following steps:

1. Analysis of the corporate web resources structure and information content.
2. Construction of the corporate web resource general structure.
3. Comparison of the obtained structure with the schema.org ontology.
4. Add new classes to describe data that is not in the schema.org dictionary.
5. Construction of the corporate web resource general ontology.

3. Research of web resources navigation elements

Purpose of the research: determination of the general structure of corporate web resources based on the analysis of parts of the top level of the main navigation menu.

The structure of 500 corporate web resources was examined. The analysis of web resources and data mining was carried out using an automatic parser implemented by means of the WDOL (Web data operating language) language [8].

Web resources for analysis were automatically selected from Google search results for "Our company". This wording of the search query obviously provides a high probability of having in the search results of corporate sites from different industries. Only the main page of web resources was analyzed.

A total of 584 web resources were processed, including re-links to web resources, bulletin boards, and service aggregator sites.

Navigation menu items in the form of “Item name” → “Links” were extracted from each web resource and stored in the database. Table 1 presents a list of structural elements sorted by decreasing frequency of their occurrences to web resources. A total of 2371 unique navigation elements were obtained.

Items with less than 30 occurrences were ignored as uninformative. These are specific to some companies:

- elements with reference to specific services, which in general are elements of the second level for the root “Services”;
- elements with reference to specific articles on the company's activities, which in general are the elements of the second level for the root “Articles” or “About the company”;

Screened out as uninformative - 2326.

The list of informative structural elements is presented in table 1 in descending order of the total number of occurrences.

Table 1

List of structural elements of corporate web resources

№	Item name	Number of occurrences	№	Item name	Number of occurrences
1	Contacts	891	24	Awards	84
2	About the company	637	25	Partners	81
3	Services	611	26	Clients about us	78
4	News	363	27	FAQ	75
5	Product	232	28	Projects	74
6	Vacancies	217	29	Dealers	72
7	About us	193	30	Documents	72
8	Cooperation	185	31	Our facilities	71
9	Certificates	177	32	Promotions	68
10	Information	169	33	Terms of cooperation	65
11	Service	156	34	For dealers	63
12	Reviews	146	35	Our contacts	62
13	Home	142	36	Events	60
14	Career	138	37	Distinctions	59
15	Articles	125	38	For partners	57
16	Press center	124	39	Clients	56
17	Licenses	102	40	Our projects	56
18	Our services	98	41	Our company	54
19	Our works	97	42	Our partners	53
20	Our products	93	43	Production	52
21	Objects	93	44	Gallery of works	51
22	Our clients	87	45	Our news	50
23	Shop	86			

The next step in manual mode was to review 50 randomly selected web resources from the previous sample to study the information content of the basic structural elements.

“Our company” contains information about the company, activities, history, etc. 72% partially duplicates the information posted on the main page. Therefore, it is advisable to combine these sections.

“Projects” contains a list of realized orders in the form of a portfolio with the name, image and characteristics. It is advisable to combine with the sections “Our clients”, “Our objects”, “Gallery of works”, as close in meaning.

“Documents” contains a list of documents certifying the legality of the company, samples of documents to be filled out by the client, contracts.

“Certificates”, “Licenses”, “Distinctions” – list of documents with a description.

The sections “Documents”, “Certificates”, “Licenses”, “Distinctions” should be combined as close in meaning.

“Reviews” contains visitor reviews. It is important to note that in 17% of web resources with published user reviews, there is no possibility to post a review. This casts doubt on the reality of the reviews. When analyzing web resources and making decisions, you can take into account only those reviews whose owners have confirmed their identity by logging in to the web resource, or specialized third-party services.

“Vacancies” contains a list of vacancies, a description of career opportunities, and general information about employment.

“Cooperation” contains information about the terms of cooperation.

“Our partners” contains a list of partners, in most cases in the form of a list of logos.

It is advisable to combine the sections “Cooperation” and “Our Partners”, “Dealers”, as close in meaning.

“Contacts” contains the necessary contact information for the user, the map. You should also move all types of contact forms to this section, such as Feedback, Call Meter, Callback, Table Ordering, and more. Such forms significantly overload the pages of the web resource and are often very intrusive.

“Services” contains a list of services provided by the company, if necessary, divided into types: design, installation, service, etc.

“Product” contains information about the presented products.

“News” – this section is used “for its intended purpose”, namely to cover news about the company's activities, only by representatives of big business. This is due to the fact that they have a large staff, large production, regional offices, and large-scale enterprises generate a large number of events that can be covered. On the web resources of small and private companies, the News section is in most cases filled with general news, which is automatically downloaded from news resources, or is left blank altogether.

Section “Articles”. Based on statistics from the website of the American Webmasters Association (AWA) [9], owners of web resources fill this section with useful informational articles about the company's activities, services, products, rules of use, only in 12% of cases. Another 46% order copy-writing services to fill the section with SEO-articles to increase the artificial ranking in the results of search engines. 23% contain slightly revised copies of existing articles from other resources. The last 19% remain empty. In 31% of web resources viewed, this section contains no more than two articles placed shortly after the creation of the web resource, or empty at all. SEO-articles were present in 52%. They are characterized by:

- a large number of keywords highlighted in bold;
- the availability of general information on the topic, but without details, as they are not written by specialists in the subject area;
- a small volume, about 1000-1500 characters (200-250 words), which is sufficient for indexing by a search engine.

19% have large, detailed articles that deeply cover the topic.

Often, brief news and news articles are placed in the left or right side of the web resource with links to the full text of the articles. This provides better indexing of the article section by search engines, and also creates the illusion that the content of the pages is frequently updated, which is one of the positive factors for increasing the ranking of a web resource in search results, but prevents the user from perceiving the main information on the page, overloading it with unnecessary information.

Given the relatively low percentage of using the sections “News” and “Articles” to place useful information for users, it is advisable to combine these elements into one, as well as to add sections “Promotions”, “Events”, “Information”, as close in meaning. If you overload the main menu, you can make it a sub-item in the section “Our Company”. As a result of the study, the structure of the upper level elements was formed. Similar structural elements were grouped into thematic groups for display.

The general structure of corporate web resources was built based on the results of the experiment, as well as the results of research by leading experts in the field of web design, and the usability of web resources [10]. According to the recommendations, the number of root elements of the main navigation should not exceed 8. The constructed structure includes 10 elements, but it should be noted that none of the studied web resources had all these elements together. The maximum number of them was

9. The general scheme includes all possible options, and when developing each specific project, only those that meet the company's requirements will be selected. Table 2 presents the groups of structural elements and the total frequency of occurrence of all elements of the group.

Table 2.

Groups of structural elements of Web resources

Group of structural elements	Total frequency of occurrence
Articles, Information, FAQ, News, Our news, Press center, Promotions, Events	1041
Home, About us, About the company, Our company	1026
Contacts, Our contacts	953
Services, Our services, Service	865
Our clients, Clients, Objects, Projects, Our works, Our projects, Our objects, Gallery of works	585
Partners, Dealers, Cooperation, Terms of cooperation, For partners, For dealers, Our partners	577
Documents, Licenses, Certificates, Awards, Distinctions	494
Products, Our products, Production, shop, goods	463
Vacancies, Careers	355
Feedback, Customers about us	220

The number of occurrences, which exceeds the total number of analyzed web resources, due to the fact that on some web resources the main navigation is duplicated in the footer of the web resource. The algorithm for extracting navigation elements takes into account such items as individual. The data in table 2 were used as structural groups, to which were added elements close in meaning to ensure the ease of navigation on the web resource. For example, our business, team, employees, representative offices, branches, sources of inspiration, company history, company development, etc. were added to the group “About the company”. These structural elements are highlighted as second-level elements for the root “about the company”. This organization of the navigation bar reduces the overload of the web page with first-level navigation elements and builds an intuitive navigation system for the user. The next step of the study is to build a semantic profile of the corporate web resource.

4. Semantic profile of corporate web resources

The standards of semantic structure are called dictionaries of micromarking [10], which are described in [11-19]. The general structure of corporate web resources has become the basis for building a semantic profile. The construction of the semantic profile of corporate web resources took place in two stages:

1. Comparison of the general structure with the schema.org ontology, as a result of which the list of necessary classes for the description of structure was allocated.
2. Adding new semantic classes to describe those elements of the structure for which there are no corresponding classes in the schema.org ontology.

To implement the semantic profile of the corporate web resource, a number of classes were created according to the structural elements of the first level, as well as one base class, which contains all the new properties needed to describe the structure of corporate web resources. For unique identification of new classes, before their name the prefix cw (corporate website) is added, and for classes schema.org, the prefix sc (schema). In figure 1 presents a UML-diagram of the classes of the developed semantic model of the corporate web resource.

Properties common to each class are marked with a gray background and for ease of perception in all classes except the first, replaced by «...».

Classes correspond to the structural elements of the first level. Each class inherits from the schema.org ontology and the cw: corporateWebsite base class, two groups of properties:

1. Common to all classes.

sc: Thing:

- name;
- image;
- description;
- URL.

sc: WebPage:

- mainContentOfPage;
- primaryImageOfPage.

cw: CorporateWebsite:

– keyNote - the most important part of the page content, keywords, the main idea that can be used to improve the quality of information retrieval;

– announcement - a brief announcement of the information presented on the page, usually used on a page with a list of news, articles, vacancies, reviews, etc. The first paragraph or several sentences of the main text are most often noticed as an announcement.

2. Specific to a particular class.

Each class has its own specific set of properties that describes the information content of pages of this type.

The new class cw: CorporateWebsite has the following properties:

- keyNote;
- advantage. Some property that favorably distinguishes an object or service from others. Can be used to describe projects or services. It is also used as one of the components of the feedback system. As a rule, the benefits are indicated by a list;
 - disadvantage. Mainly used as one of the components of the feedback system;
 - projectTimeline. There are two options for using this property: simple - the number of working days required to complete the project, or folded, divided into stages of execution;
 - terms. Used, for example, to describe the necessary conditions for obtaining the status of a dealer, or the conditions for successful project implementation;
 - certificate. An official document issued by the competent authorities certifying, for example, the quality of the products presented;
 - license. An official document issued by the competent authorities certifying, for example, the right to provide a certain type of service;
 - announcement.

The constructed block diagram and semantic profile of corporate web resources are the basis for the development of the following elements of the CODI system [20]:

1. Specialized content management system for corporate web resources with integrated semantic markup.
2. Module for displaying the content of web resources based on custom templates.
3. Metasearch system based on search engine processing of popular search engines with the use of search agents and the ability to display search results based on user templates.
4. A personalized user's web page that displays user-relevant information that is automatically retrieved and aggregated from various sources by search engines.

5. Perspectives of application of web resources semantic profile

Developing separate semantic profiles for different types of web resources, instead of using one large ontology, has the following advantages:

1. For developers of semantic profile:
 - division of a large task into smaller ones in terms of volume and complexity - separate project groups for the development of their areas;
 - the possibility of rapid development of promising areas due to the concentration on solving a small task of developing a separate semantic profile;
 - ability to add new, specific to each type of web resources, classes and properties. Within one global ontology, the addition of industry-specific classes leads to a rapid growth of the structure and a significant increase in the complexity of its practical use.

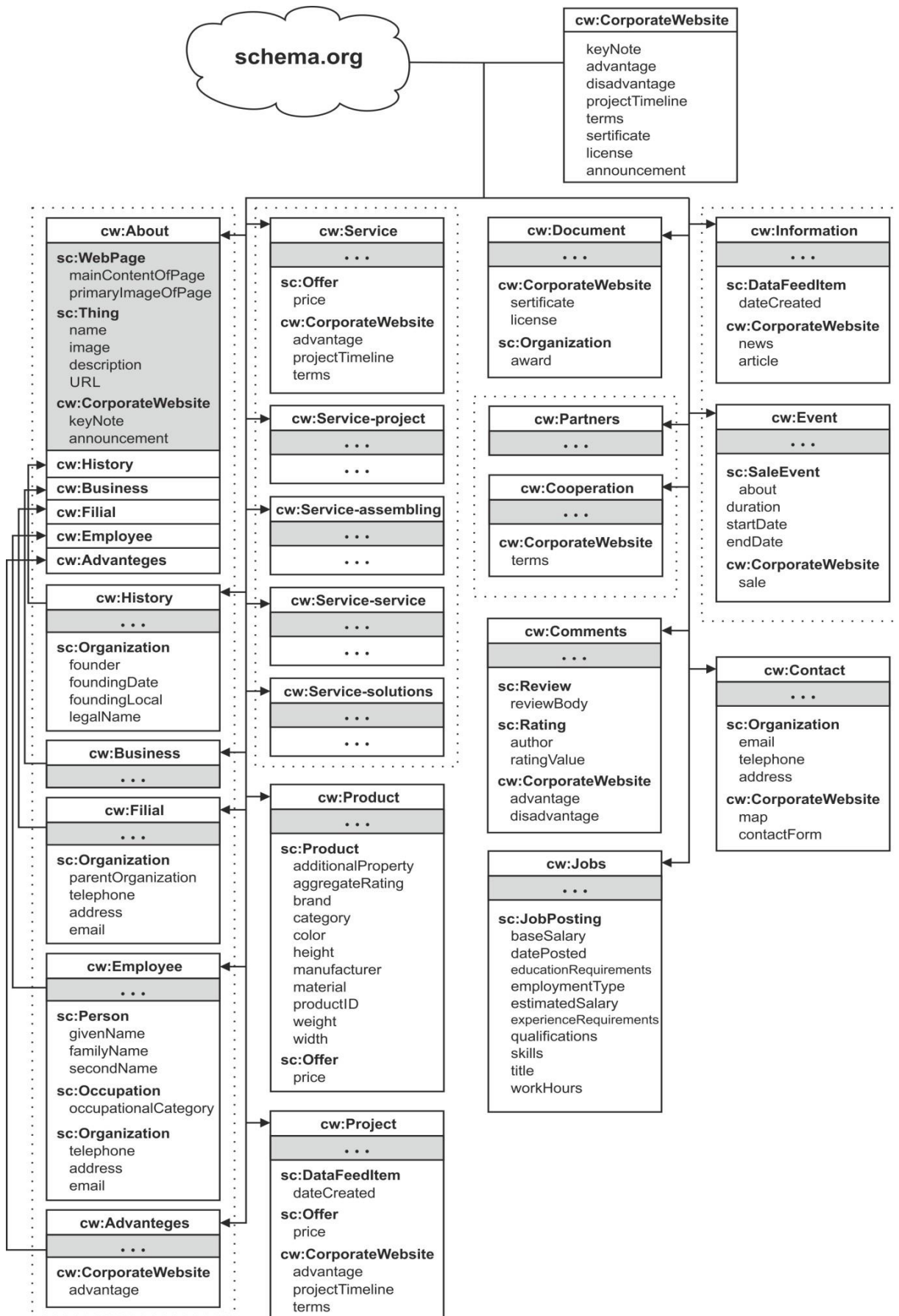


Figure 1. UML-diagram of classes of corporate web resource semantic profile

2. For web resource developers and network users:

- Ease of use. If semantic classes are tied to specific structural elements of a web resource, their use becomes more convenient and transparent. When using one global ontology, developers are forced to adapt general purpose classes to the specifics of specific structural elements and often there is a situation when such a possibility is simply absent;

- Accessibility for new users due to a significant reduction in time to study classes and properties.

The practical implementation of the semantic web concept requires the solution of two urgent problems:

1. Integration of semantic markup to existing web resources.

2. Development of new web resources with integrated semantic markup.

Solving these problems requires the development of a web resource markup strategy, which includes:

- selection of elements that need to be marked;

- selection of the standard according to which the marking will be carried out;

- choice of automatic or manual approach to markup integration;

- selection of micro-markup code generation tools;

- choice of method and tools for integrating the generated code into the web resource HTML-code.

A significant factor hindering the widespread use of semantic markup is the lack of comprehensive solutions that provide all the necessary tools to solve the problems described above.

To successfully solve the problem of integrating semantic markup into the HTML of new and existing web pages, it is necessary to analyze existing approaches and methods.

6. Conclusions

The development of the semantic web concept has become another evolutionary step in the development of the global network. The integration of semantic markup into the HTML-code of web pages creates the conditions for the application of methods of machine processing of information placed on them. This in turn opens up opportunities for the development of intelligent data retrieval methods, as well as methods of displaying data based on the identification of information by semantic attributes.

Based on the study of the navigation menu and information content of corporate web resources, their general structure was built, which became the basis for creating a semantic profile.

An approach to the practical implementation of the semantic web concept as a necessary condition for the development of e-commerce is presented. It is to develop separate semantic profiles to describe the information content of different types of web resources instead of adapting the global ontology schema.org. As part of the development of the semantic web concept, based on the corporate web resources structure was developed their semantic profile. It is a set of classes to describe the information content of corporate web resources.

In the future, the presence of a semantic profile and custom templates for displaying content allows you to change the concept of a web resource. It is no longer a standalone site with its own strictly defined design, which is displayed when you refer to a domain name. The new approach defines a web resource as a set of data and a semantic profile compiled according to certain rules. The use of a semantic profile allows you to display web data in any user-friendly form, based on his personal display template for the appropriate type of web resources. The user gets the opportunity to arbitrarily change the web resource structure, choose which elements of the web page will be displayed and which will be ignored. Also, the semantic profile allows you to operate on data outside the domain name, for example, to compare directly at the search stage of certain services or goods, apply filters and sorting.

7. References

[1] Semantic web. [Online]. Available: <https://www.w3.org/standards/semanticweb>.

[2] Good Relations. [Online]. Available: <https://www.goodrelations.co.uk>.

[3] Web page semantic markup. [Online]. Available:

- <https://support.google.com/merchants/answer/6069143>.
- [4] Structured Data Markup Wizard. [Online]. Available: <https://www.google.com/webmasters/markup-helper/u/0/>.
 - [5] Haag S. Management Information Systems for the Information Age: Ninth Edition. McGraw-Hill Higher Education, 554 p., 2012.
 - [6] Schema.org semantic dictionary classes. [Online]. Available: <https://schema.org/docs/about.html/>.
 - [7] Integration of semantic dictionaries into the schema.org environment. [Online]. Available: <https://schema.org/docs/about.html#cgsg>
 - [8] Zosimov, V., Bulgakova, O. Development of Domain-Specific Language for Data Processing on the Internet International Scientific and Technical Conference on Computer Sciences and Information Technologies (2020), 287–290, <https://doi.org/10.1109/CSIT49958.2020.9321968>.
 - [9] Use of articles on web resources. [Online]. Available: <https://www.aawebmasters.com/ecommerce/>
 - [10] Lawrence D., Tavakol S. Balanced Website Design: Optimising Aesthetics, Usability and Purpose. Springer Science & Business Media, 236 p., 2016.
 - [11] Zosimov, V., Bulgakova, O. Application of Personalized Ranking Models Based on Expert Evaluations for Sorting Goods on E-commerce Web Resources. International Scientific and Technical Conference on Computer Sciences and Information Technologies, P. 42–45, 2020. <https://doi.org/10.1109/CSIT49958.2020.9321902>
 - [12] Open Graph. [Online]. Available: <https://ogp.me/>
 - [13] Friend of a Friend (FOAF): an experimental linked information system. [Online]. Available: <http://www.foaf-project.org/>
 - [14] Dublin Core. [Online]. Available: <https://dublincore.org/>
 - [15] Work on optimization of extended snippets. [Online]. Available: <http://astra.red/rabota-po-optimizatsii-rasshirennyih-snippetov/>
 - [16] Internet Live Stats online statistics service. [Online]. Available: <https://www.internetlvestats.com>
 - [17] Sociological research of problems of introduction of micromarking. [Online]. Available: <https://www.schemaapp.com>
 - [18] Kosara T., Bohrab S., Mernika M. Domain-Specific Languages: A Systematic Mapping Study. Information and Software Technology, vol 71, pp. 77-91, March 2016.
 - [19] Diagram of classes of the dictionary of semantic markup Good Relations. [Online]. Available: <http://www.heppnetz.de/ontologies/goodrelations/20100412/v1.html>
 - [20] Zosimov, V., Bulgakova, O., Pozdeev, V. Complex internet data management system. Advances in Intelligent Systems and Computing, 2021, 1246 AISC, P. 639–652. https://doi.org/10.1007/978-3-030-54215-3_41