Semantic Information Resources with a Complex Structure: Knowledge Representation, Scaling and Search Problems

Julia Rogushina, Irina Grishanova[1]

[1] *Institute of Software Systems of the National Academy of Sciences of Ukraine, 40, Ave Glushkov, Kyiv, 03181, Ukraine*

Abstract

We analyze scaling problems arising in development of the intelligent information systems (IISs) and main reasons for their occurrence. IISs integrate various elements of artificial intelligence (AI) for acquisition of knowledge relevant to actual user tasks. Important properties of these IISs are use of data with complex structure and orientation on semantic information resources (IRs). Therefore we analyze main features of the Data-Centric AI and opportunities for acquiring domain knowledge in various representations from Big Data. Knowledge organization systems (KOS) provide models and methods for effective store, retrieval and use of information processed by the Web-oriented IISs, and we consider existing approaches for their software platforms. We analyze the specifics of the scaling for systems focused on the semantic information processing and its differences from traditional data and Big Data scaling. This specifics is caused by complexity of data structure, number of various semantic relations between information objects into IR and complexity of semantic queries executed by KOS.

On example of e-VUE – the Wiki-portal of the Great Ukrainian Encyclopedia – we analyze various situations that arise in process of practical development of semantic IR with large volume and complex structure. Various aspects of semantic retrieval into e-VUE are considered from the scaling point of view (such as number of information objects, relations between them and number of their properties). On base of this analysis we propose a set of practical recommendations aimed at ensuring more efficient development of such IRs that provides their scaling.

Keywords

Semantic information resource, scaling, ontology, Wiki-technology, metadata, semantic markup

1. Introduction

Modern intelligent information systems (IISs) that use and generate knowledge are oriented on functioning in the Web open environment and use of various external sources of information. One of the promising directions for the effective use of information is the transition from data processing to knowledge processing that ensures decline of data, but we have to take into account that knowledge generated on base of big data and information resources (IRs) the Web can have large volumes and a complex structure too. Moreover, volume and quality of such acquired knowledge depends on methods and tools used for data processing and on selection of processed data sets relevant to user tasks.

IISs increase their work if they receive information from semantic IRs where content is described and structured by formal means that ensures unambiguous interpretation.

Processing of the big amount of knowledge requires scalable solutions for *knowledge organization systems* (KOSs) that provide access to content of such IRs on semantic level and support knowledge management [1]. KOSs are used as a conceptual infrastructure to provide understanding, integration

[1]13th International Scientific and Practical Conference from Programming UkrPROGP'2022, October 11-12, 2022, Kyiv, Ukraine
EMAIL: ladamandraka2010@gmail.com (A. 1); i26031966@gmail.com (A. 2)
ORCID: 0000-0001-7958-2557 (A. 1); 0000-0003-4999-6294 (A. 2)

and search for knowledge, preparation data for knowledge acquisition, identification of semantic links and rules, information retrieval [2].

KOS contains various instruments for description of IR content and for access and retrieval of documents and information. Main elements of KOS can be represented by RDF with the help of Simple Knowledge Organization System (SKOS) [3] and provides identification of IOs that are interested user by additional personified information.

In addition to common scaling aspects, development of distributed IISs focused on the semantic processing of information needs in scalable knowledge representation and processing means that:

- on the one hand, have sufficient expressiveness for domain tasks,
- on the other hand, support execution of semantic search in an acceptable time on the conditions of increasing the elements of the knowledge base.

The functioning of IIS depends on data sources. Therefore, they need in technologies and tools of collecting useful data, as well as the selection of adequate, trusted and high-quality IRs. Many of the Web-oriented IRs are developed as a result of the joint activity of users on base of the Web 2.0 technologies [4] which makes content more dynamic and relevant. The most successful Web 2.0 platforms for collaborative creation of large-scale content are Wiki-technologies [5], such as MediaWiki [6]. Such systems can use KOSs based on Wiki-ontologies that are a special case of ontologies with a set of restrictions on the characteristics of relations that reflect the knowledge structure of semantic Wiki-resources [7].

2. Data-oriented artificial intelligence

Technologies of data analysis are changing rapidly. Traditional software development strategies are being replaced by modern approaches focused on artificial intelligence (AI) methods. Transformation of "raw" data into structured representation requires a lot of time and the use of experts, so it is advisable to use already structured IRs whenever possible: such structuring allows their automated processing by KOSs at the semantic level.

Now many researchers consider the concept of data-centric AI [8] instead of model-oriented approaches. Traditional software is based on software code, while AI systems consist of various combinations of code and data, and it is data-related problems that are currently the most pressing for developing intelligent applications.

Although the majority of existing information is stored in digital format, this does not mean that this data are easy to process. To make data available for IIS, data need in structure and metadata.

For a long time, data availability and computing power were limited by technical possibilities and it causes need in code optimization for AI software. But Big Data development [9] that provides storage and analysis of great amounts of data actualizes data-oriented approach where key importance [10] for the processing of Big Data in IIS deals with the metadata analysis: such metadata describe Big Data semantics. Big Data is data that, for various reasons, cannot be processed by such traditional information systems as relational databases. Technologies of Big Data are supported by a significant number of software solutions. In order for Big Data to become useful, it is necessary to find those sets of them that are pertinent to some practical task. This pertinence can be detected by matching of metadata elements for Big Data that describes their semantics with knowledge models of task models (e.g., domain ontologies) with complex structure.

Main advantages of data-oriented AI:

- data specialists have a better understanding and control over the structure of datasets and how these data are processed;
- reducing the cost of data model development by reducing the required volume of data or extracting more value from unstructured, heterogeneous data sources;
- simplification of data annotation with the help of smarter analysis processes;
- detection of duplicated, damaged or low-quality data in the early stages of analysis;
- ensuring the quality of data markup and avoiding a subjective approach to this.

Data-oriented AI often uses such components of semantic technologies as semantic IRs which are subsets of IRs where the content elements are clearly and unambiguously connected to the elements of the knowledge base (e.g., with the help of semantic markup) or the elements of content are represented on base of ontology-based knowledge representation formats (e.g., RDF [11] and OWL

[12]) and special cases of domain ontologies [13] (e.g., thesauri, taxonomies [14]) that can formalize various aspects of user needs [15]. IISs are focused on processing and creating knowledge, not data: the effectiveness of their work is significantly determined by the choice of analysis methods and forms of knowledge representation. Therefore, selection of KOS for some practical task has great importance and influences on its usefulness [16].

Another challenge of data-oriented AI is the flexibility of data access and formats of data representation. If the data storage system imposes restrictions on changes in the scale of data and on the transition to other processing tools, this can lead to negative consequences in the process of creating and improving IIS. Poor selection of knowledge representation formats can restrict their expressiveness and IIS functionality in general. These problems are determined not by whether the system can work at all, but by whether it works reliably, efficiently and affordably on a large scale.

Quite often, such problems arise in the process of transition from design and prototyping to deployment, industrial operation and development of IIS or due to the accumulation of a significant amount of content. Common reasons for this situation:

- Changes in the execution environment: test environment can differ from industrial one;
- Requirements for service-level agreements (SLA) [17]: quantitative and qualitative characteristics of the provided services used as a formal contract between service provider and consumer to ensure service quality, such as their availability, user support, time to correct malfunctions, etc. depend on environment;
- Processing of data on a larger scale: industrial systems use Big Data, data with complex structure, heterogeneous information from various sources, various representations of data, etc.

All of these problems are caused by changes between development and production settings, as the test environment cannot model all important aspects of production environment. SLA is defined as an explicit statement of expectations and obligations in relations between two organizations: the service provider and customer but it depends on the step of product development. For example, a particular application may meet the latency SLA requirements during isolated testing and development, but these requirements are not met when service is running in a production environment where other applications compete for computing resources and many users have access to IIS.

Data is just one of the challenges faced by IIS developers with AI elements in large-scale production. However, the requirements for scaling of data and its infrastructure are often overlooked, although they can make the practical use of IIS impossible. That is why in this paper we analyze the problems associated with content scaling of semantic IRs and take into account the specifics of information processing at the semantic level.

Ensuring data security in IIS also requires scaling. How security is implemented at a local or small scale that used during development always is not be reliable at a large scale. Traditional security conceptions such as process permissions and user ID become much less effective in scalable systems that need in more safe technologies. For example, SPIFFE (Secure Production Identity Framework for Everyone) is a set of open-source standards for securely identifying software systems in dynamic and heterogeneous environments. IISs that use SPIFFE can easily and mutually authenticate wherever they are running. This technology provides various solutions to deal with threats to such large applications by defining a cryptographically validated workload identifier to protect communication channels between processes and can be used in IISs with the big number of users and services [18].

3. IIS scaling problems

Researchers define some specific problems in scaling of applications with AI elements [19]:

- comprehensive data strategy and unified data access;
- separation of problems at the platform level;
- scalability, not just scale for every separate problem;
- multifunctional design.

Scaling of modern IIS needs to take into account various aspects related to the following properties of information: the size of the data itself; the number of objects being processed; the complexity of processing algorithms and the number of software modules used for information analysis; sources of information, etc.

Scaling in terms of *data size* should support the ability not to enlarge data without necessity: for example, provide open access APIs instead of local copies. But such an approach requires unification of data representation, so various analytic tools can process the same datasets without adaptation.

This unnecessary copying is especially common in machine learning (ML) and AI projects, where data scientists use a wide range of specific tools that differ from the Big Data tools used by data engineers. AI and ML tools typically do not have direct access to data stored in Big Data platforms. The result is the spread of redundant copies.

Another reason for this unnecessary copying of data is a data infrastructure that lacks fully distributed metadata, which can lead to metadata overload when multiple applications access large data sets.

Scaling in terms of the *number of information objects* (IOs) – files or other data elements – refers to the ability to simultaneously process a large number of different objects. If the data infrastructure is not designed to handle a very large number of IOs, it can cause a significant increase in processing time, overload the platform, and even crash the system.

Scalability in terms of *processing facilities* is related to the fact that the data processing architecture and infrastructure should not be limited to several applications on the same platform, because otherwise IIS needs in different clusters configured to support each individual application.

Scaling from the point of view of *geodistributed locations* deals with use of data from geographically distributed sources or with the run of programs from different locations. It causes the challenge of getting large amounts of data close to its source and making decisions about data parts sent to the main data centers, as well as how and from where to provide analytic applications to process that data.

Unfortunately, people can misjudge the potential scalability of their systems and directions of its development, or believe that it is okay to design a system that successfully meets their current needs, but does not take into account growth of needs in future. Sometimes the choice of architecture and data infrastructure imposes such limitations that can be avoided.

The most common mistakes that prevent the effective scaling of IISs:

- AI and analytic applications work in separate systems (clusters);
- grows of data and applications causes expansion of IT team;
- different user groups or applications need in personal copies of the same data, even for very large data sets;
- data movement (for example, between local storage and cloud) is implemented at the application level;
- Big Data platforms are intended for specialized projects instead of being a universal general platform;
- configuration of IIS architecture and infrastructure is oriented on fixed scale of data and applications;
- scaling of applications requires changes into the architecture of existing system.

In [20], current problems of Big Data are analyzed and the main differences between big and traditional data are compared. We propose to expand such analysis by comparing them with the same measures with semantic data [21]– other specific type of data – based on the Semantic Web standards. Processing of Big Data metadata can be considered as a special case of semantic data analysis with more attention to scaling problems [22].

Table 1
Comparing the characteristics of traditional, big and semantic data

Component	Traditional Data Bases	Big Data	Semantic Data
Requests	SQL	Largely Abandoned SQL	SQL-like requests
Architecture	Centralized	Distributed	Distributed with a hierarchy of elements
Data types	Structured	Structured, partially structured or unstructured	Structured by formal semantics
Data model	Fixed schema	No schema	Various task-dependent schemas (in RDF, RDF-S OWL)
Relations between data	Known fixed set of relations without formalized semantics	Unknown or undefined, partially provided in metadata	Extensible set of arbitrary relations with formalized semantics
Data volume	Large	Very large	Relatively small
Number of relations	Small	Very small	Significant
Data integration	High	Low	Very high
Semantics	Not formalized	Not defined, only metadescriptions	Formal interoperable

Table 1 shows that the main scaling problems for semantic IRs concern the processing in various requests of a large number of formalized relations between data elements that causes the generation of an excessively large number of combinations of content elements.

Thus, we propose to analyze the following scaling aspects that are significant for semantic IRs:

- the total amount of stored data and relevance of infrastructure facilities (for example, the server volume and capacity);;
- the number of typical IOs and the complexity of their structure;
- the number of IO individuals of various types;
- the number of relations between IOs;
- the number of typical IOs and the complexity of their structure;
- the IO metadata infrastructure (representation, indexing, viewing and search tools);
- the number of IR users and the number of their visitations;
- the number of operations in typical user requests and in processing of navigation actions;
- the speed of knowledge base updating after changes – both in metadata and in content.

Insufficient attention to scaling in designing distributed IISs that are based on large-scale IRs can lead to inefficient performance of system in general.

4. Problem definition

In addition to common scaling aspects, development of distributed IISs needs in focusing on specifics the semantic processing of information for scaling knowledge representation and processing means. Such processing has to:

- on the one hand, provide sufficient expressiveness for domain tasks,
- on the other hand, support execution of semantic search in an acceptable time on the conditions of increasing the elements of the knowledge base.

In the general case, this is a complex theoretical problem, and therefore in this work we analyze its special case for the scaling of semantic Wiki-resources, which contain a large number of information objects of various types (such as encyclopedic portals). On base theoretical research of KOSs we single out those factors that affect the scalability of semantic IRs, and determine the conditions that ensure the successful development of such resource. Experience in the development encyclopedic portals is taken into account in the construction of practical recommendations.

5. e-VUE as an example of semantic Wiki-resource with a complex structure

We propose to consider problems and practical solutions for scaling of complex Wiki-resources that include semantic extensions on example of the portal version of the Great Ukrainian Encyclopedia (e-VUE), which contains information from many fields of knowledge [23]. The conception of this encyclopedia is focused on modern scientific understanding of the world picture, the history of human civilization, the contribution of the Ukrainian people etc., and its portal version provides users with universal access to the content and more intelligent tools for search and navigation. e-VUE is built as an innovative IR based of modern knowledge models and the Semantic Web standards [24] such as OWL [25], RDF and RDF Schema [26] supplemented by original task-specific software solutions.

e-VUE is a semantic Wiki resource with a complex structure implemented on the MediaWiki technological platform [27] and its Semantic MediaWiki (SMW) [28] semantic extension. Creation of e-VUE content is supported by the big number of domain experts from various domains and is coordinated by team of professional scientific editors and knowledge engineers. This IR can be considered as a Semantic Web application that satisfies the main requirements formulated by the Semantic Web Challenges [29].

e-VUE implements semantic markup and retrieval based on the SMW plug-in that extends the MediaWiki functionality. Use of SMW provides to link Wiki pages with constants of various types and with other Wiki pages by semantically defined relations and to perform queries where semantic properties of pages can be included both to the conditions and to the parameters of the query results.

Figure 1. e-VUE portal.

We develop the set of semantic templates of typical information objects (TIOs) for unified representation of content of e-VUE articles. Every TIO is linked with some non-empty subset of Wiki pages that belong to the same set of categories, have the same (or similar) structure and semantic properties. Development of TIO system is based on the joint work of knowledge engineer and domain experts. In e-VUE, TIOs are based on the expressive capabilities of the Wiki environment and its semantic extension: TIO templates provide input of values of selected semantic properties and automatically [30].

Since 2020, e-VUE has been registered on the web site of the Semantic MediaWiki product community (https://wikiapiary.com/wiki/Great_Ukrainian_Encyclopedia), which shows the rate of growth in the number of pages, user activity, and the number of edits in this IR (Fig. 2).

Figure 2. e-VUE page on the Wikiapi.com website

The development of the site is also reflected by Google Analytics service that provides statistics on users of web applications such as the activity of the Web site users, the session duration, the number of pages viewed per session, the number of rejections, etc., as well as information about traffic sources (Fig.3).

Figure 3. The number of e-VUE users according to Google Analytics.

These statistics indicate the need for a scalable approach to the further development of the portal, which will ensure its operation in conditions of increased content and complexity and for a larger number of visitors. It causes the need for use of distributed scaleable knowledge management methods for the further development of this resource.

6. e-VUE and SMW-based semantic search

Semantic search supported by Semantic MediaWiki plug-in is an improvement of the traditional Wiki search with the use of information about the structural elements of the retrieved information object, about its properties and relations with other information objects. For example, you can search for a country by the name of a person, and a person by place and year of birth. Unlike the traditional Wiki search (e.g., offered by Wikipedia), semantic search can use a set of conditions and take into account not only categories. These advantages are based on enhancement of KR by semantic markup and enrich the functionality of IR. Result of semantic search is a list of pages that require the query conditions and the values of their semantic properties defined by query. Correct construction of queries needs in exact names of semantic properties that can be obtained from pages of corresponding TIOs (they can be found in the usual search in the "Template" namespace).

Semantic search in e-VUE portal can be executed by several ways:

- by special page "Semantic search", where query parameters are entered in the appropriate fields, and users don't need in specific knowledge of the syntax of the search language (it is enough to understand functions of search fields on this page and to know names of properties and categories) ;
- by search queries formulated by the specialized SMW search language and embedded in other pages;
- with the help of the API requests based on special program code.

While semantic properties and categories allow you to structure Wiki data, queries help to use that information: combine data and visualize it. From the point of view of expressiveness, queries generated by special page are the most limited, but their creation is very easy. Elements of the semantic search page are (Fig.4): 1 – search conditions; 2 – output semantic properties; 3 – representation form to output information; 4 – number of output results; 5 – order of search results. When these fields are completed, user clicks the "Find" button (6) and query is generated and executed.

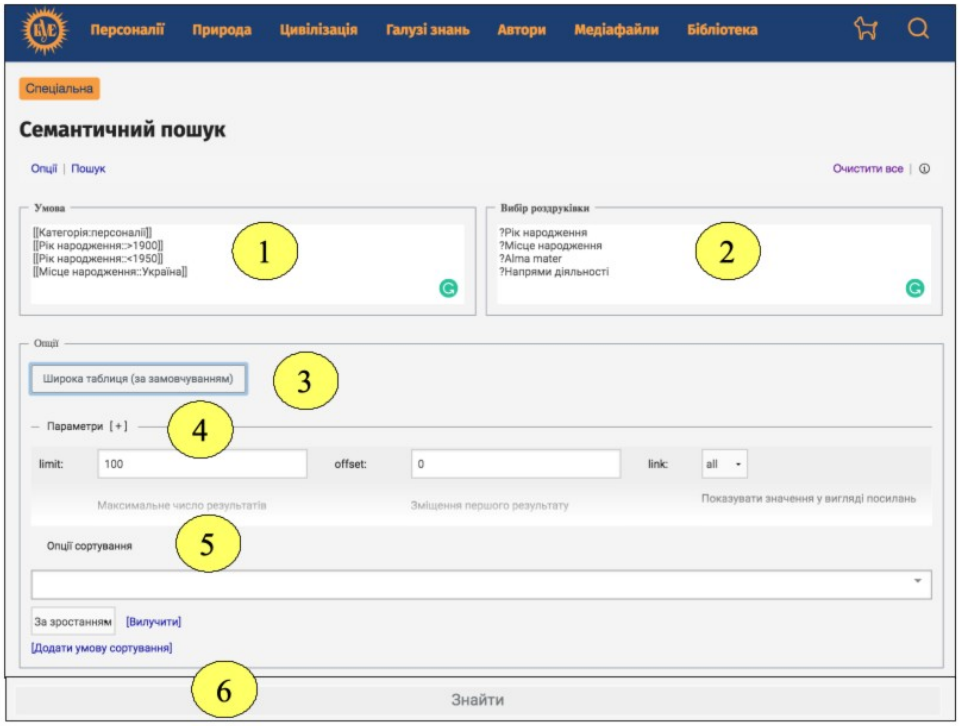

Figure 4. Semantic search page of e-VUE

SMW has a simple but powerful SMW-QL query language that allows to filter pages according to specified criteria, and to display as query results only the information that interests the user, and not the entire text of the Wiki page. Embedded SMW queries can use additional variables, such as the properties of the current Wiki page, the current date and time, and the search itself is performed in the built-in knowledge base of IR among structured data. Results of such queries are generated

dynamically and agree with current e-VUE content. The correctness of the request execution depends on the quality of the data they process and correctness of the semantic markup of the pages. Examples of embedded queries in e-VUE:

- generation of list of articles prepared by selected author,
- search of knowledge field moderator for current article;
- generation of all employees of organizations represented by e-VUE articles;
- list of persons with today's birthday.

API requests are the most universal but their execution takes more time, because it is based on a full-text search in the entire IR content. Semantic markup of the Wiki pages is a necessary condition of the SMW search, and API-based queries use its elements as samples.

7. e-VUE and FAIR principles

The use of Wiki technology and its semantic extensions can provide development of semantic IRs that correspond to the basic requirements of FAIR (Findable, Accessible, Interoperable, Reusable) [31]. These guiding Principles for scientific data management and stewardship are oriented on Open Science data, and their aim is efficient access to big volumes of information with complex heterogeneous structure. FAIR data and metadata standards could help facilitate compliance with the principle of data minimization, by allowing for an assessment of which data to reuse on the basis of an analysis of (by and large non-personal) metadata.

Open Science needs in collaborative reuse of research data and scientific publications. The ultimate goal of FAIR is to optimize the reuse of scientific data and its combination in various tasks, but it is important to consider that these principles are oriented towards large-scale IR and provide for scalable solutions and services. [32] Therefore it is advisable to consider FAIR principles for development of arbitrary IRs with complex structure.

According to FAIR, the functions of searching, obtaining and presenting data are not implemented by users, but by the information system At the same time, we are talking not only about the data and metadata themselves, but also about algorithms and tools for their management. It is important, that FAIR principles only call for explication of access conditions, without specifying how data sharing should be facilitated. [33].

Semantic Wiki IRs that allow the creation of semantic data and are based on the use of Semantic Web standards, provide a powerful solution for joint publish and editing of data and metadescriptions, creation of various arbitrary sets of properties based on templates of these metadescriptions. They support representation of information also for machine processed form and for human understanding.

Semantic plug-in SMW proposes built-in capabilities for loading files of various formats and adding to them metadata with a different set of attributes that can be modifies, supplemented and analyzed. SMW supports FAIR principles of IR development.

e-VUE as a resource based on the Wiki technology and its semantic plug-in SMW satisfies all four requirements of FAIR principles:

- *Findable*: Semantic MediaWiki provides intelligent search capabilities based on the semantic properties and categories of individual Wiki pages that can be considered as a flexible set of metadata available for machine processing:

F1. SMW data and metadata have global unique and permanent identifieres defined as names of the Wiki page of appropriate semantic property, template or category with open access;

F2. MediaWiki engines have the ability to add detailed metadata to data loaded to the standard MediaWiki repository or to hyperlink to an external repositories. Semantic properties and categories can be used also for this;

F3. Metadata clearly and explicitly contain the identifier of the described data as a value of the relevant semantic property;

F4. Search engines such as Google and Bing can can index SMW metadata. Since Semantic MediaWiki is a Web application, it contains a configuration file LocalSettings.php that specifies the IR type as open or closed system;

- *Accessible*: the IR user needs the means for access to the found data, possibly including authentication and authorization, and this information in SMW can be added as separate attributes on pages:

A1. Metadata can be retrieved by identifier with use of free communication protocols such as http or secured https;

A2. Metadata is available even if the data is no longer available because metadata is placed separately on the Wiki pages in a reliable repository (in the MediaWiki database);

- *Interoperable*: SMW data representation can be integrated with other data, other applications or workflows for analysis, storage and processing by publication of additional semantic properties:

I1. SMW metadata can use the Semantic Web standards and export results of the semantic search into XML and RDF;

I2. MediaWiki is an open system that can use, integrate, import (by additional plug-ins) various dictionaries, ontologies in the Semantic Web standard languages;

I3. SMW data and metadata can include links to other metadata and supports to extend the data property sets depending on the task;

- *Reusable*: SMW provides data reuse by mechanisms of semantic properties and categories that support flexible changes to the metadata structure:

R1. SMW possibility to add an unlimited number of attributes and categories to each page fulfills the FAIR requirement about detailed description of metadata by the set of precise and relevant attributes, and semantic properties can be used in the open environment for describe ownership rights and types of licenses to different objects, their sources, etc.

8. Scaling needs caused by e-VUE development

Every Wiki resource is characterized by such properties as number of pages, number of edits, number of jobs, number of active users, etc. This information about e-VUE is provided by the API testing page (Fig. 5.1) shows the number of e-VUE pages, articles, edits and users). Other important information about IR that deals with indexing processes, data rebuilding and database settings is proposed by special SMW page (Fig. 5.2) provides information on the process of indexing data . This makes it possible to assess the scope of IR, the satisfaction or dissatisfaction of the indexing state, and to make administrative decisions regarding the indexing regime.

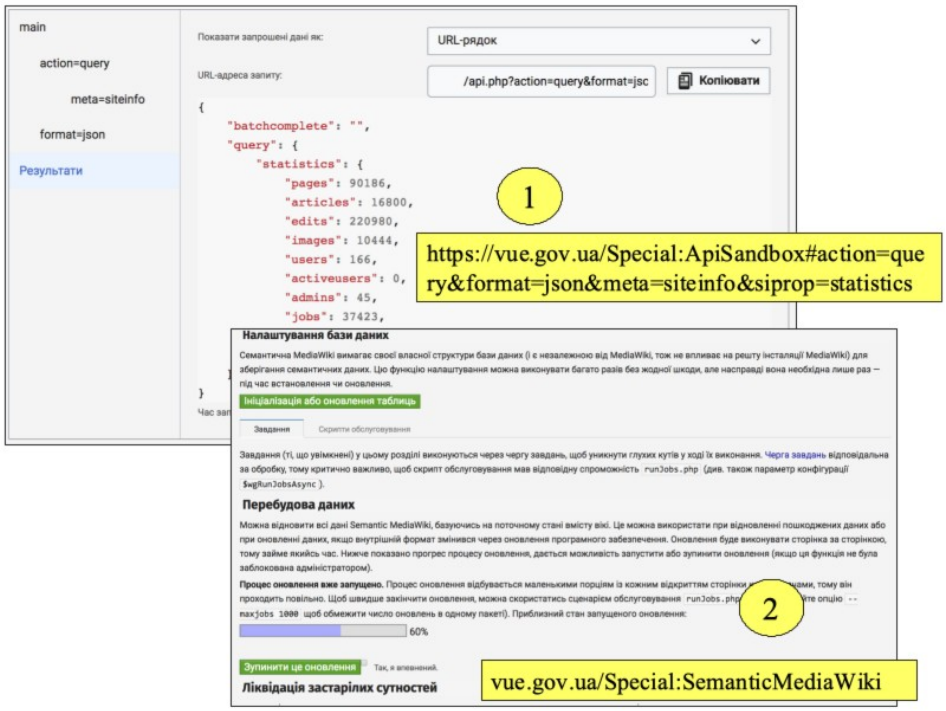

Figure 5. e-VUE statistics.

Complexity of the semantic Wiki resource depends on number of semantic properties that are defined individual Wiki pages of this IR and on number of TIO templates used for their definition. Now e-VUE contains 491 properties are created (and indexes) in VUE, 364 of them are used, more than 360,000 values are defined for these properties, and more than 20,000 built-in queries are used

(information for October 10, 2022). Not all templates are used for TIO, but for every TIO template we can see number of pages that use that template.

For example, template “Person” (Fig.6.1) is used by many e-VUE articles – the list of these pages we can see in Fig.6.2.

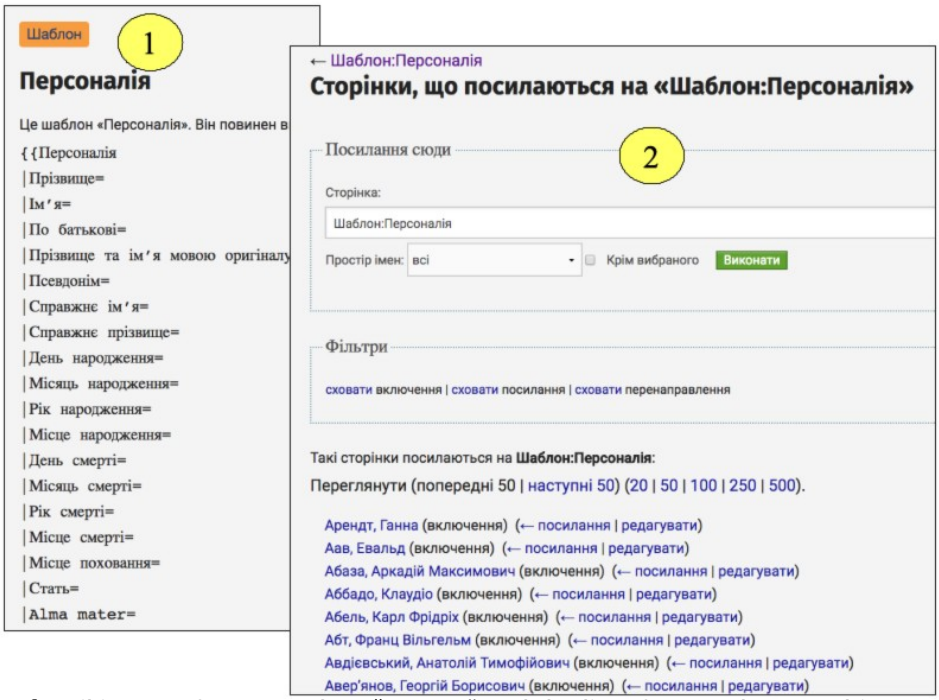

Figure 6. Wiki page of TIO template “Person” and the list of pages that use this template.

The main factors affecting the efficiency of the query execution into Wiki resource are:

- The KB scheme represented by the system of semantic properties used for content structuring determines the potential expressiveness of such queries;
- Timely indexing of the content changes in (both in the sets of semantic relations and in the pages themselves), on the one hand, should not reduce the IR productivity, and on the other hand, to ensure the relevance of the database;
- Timely removal of unused pages and infrastructure elements reduces the number of processing items;
- Quality of semantic markup (correct property names, semantically defined links to other pages, correct input of property values) provides more complete query results;
- Built-in queries that correspond to typical (often repeated) informational needs of users provide more quick satisfaction of these needs;
- The number of built-in queries and the number of visits to pages with such queries influence on processing time that has to be acceptable for IR users;
- Representation of query results has to be understandable and convenient to user needs;
- Location of semantic queries embedded in Wiki-pages has to represent query results exactly at the place where users need in such information.

In all search variants (regardless of whether the query is transformed into an SQL-like query to the IR knowledge base or a full-text search is performed on the entire content), the speed of query execution depends on its complexity, that is, on the number of conditions and restrictions. Therefore, for example, it is advisable not to enter unnecessary conditions (for example, if you need to find educational institutions of a certain country, it is not advisable to specify the category "Organizations" in addition to the category "Higher educational institutions"). SMW does not control use of relation hierarchy, and we need in some external means of knowledge representation that can fix all domain rules and restrictions. In this work we use domain ontology that represent structure of semantic properties, categories and TIOs of e-VUE supported by methods for matching of ontological elements with SMW ones.

Many other more local practical problems of semantic search can be detected only in process industrial production of IIS and cannot be forecasted in test versions, therefore the experience of developing and implementing such search tools for e-VUE portal allows us to determine some important features of the development of semantic portals that ensure its scaling.

9. Conditions for the development of scalable semantic IR

Analysis of the specifics of creating IRs on base of the semantic Wiki-technologies shows that the main factors for the successful scaling of such resources deal with the semantic markup structure, namely, the number of relations between Wiki-pages, the correct definition their ranges of values and ranges of definition, as well as with clearly formalized values of these relations. It ensures an unambiguous common understanding of the scope of their use and prevent duplication in the creation of semantic properties. Other scaling aspects of that are universal to high-volume systems development also have specific characteristics associated with the SMW environment. Therefore we propose to comply with the following requirements:

from the point of view of data size:

- Control the size of multimedia elements of IR;
- Provide means of mass import of information from external sources;

from the point of view of the IO number:

- Control the total number of Wiki pages and delete unnecessary, wrongly created and duplicate pages;
- Develop TIO templates to avoid an increase in the number of similar names of semantic properties and errors in these names and to simplify the perception of information by users;
- Unify metadata of multimedia IOs to avoid duplication ;
- Create templates that integrate content of various Wiki pages by semantic queries;

from the point of view of the KB structure:

- Develop templates of typical IOs to unify content;
- Formalize the KB structure and clearly define the semantics of relations between pages (the built-in capabilities of SMW are not enough for this, and therefore it is advisable to use external knowledge organization systems based on ontologies);
- Determine the semantics of hyperlinks between IR pages and create the corresponding semantic properties, clearly describing their characteristics;
- Develop appropriate templates for built-in semantic queries used on several pages;

from the point of view of processing means:

- Determine the expediency of connecting plug-ins that extend the functionality of system and do not install those that are not really needed;
- Develop an adequate content indexing policy that takes into account the frequency of information updates and the number of user visits;
- Create queries without redundant conditions by analyzing the taxonomy of the IR categories;
- Minimize external software tools integrated into the IR (such as page visit counters);

from the point of view of the place of data processing:

- Analyze the number of semantic queries on Wiki pages and their complexity;
- reduce the number of semantic queries and IOs on the pages that users visit most often (for example, it is impractical to embed complex queries on the main page of the portal, instead of links to pages with these queries);
- timely create backup copies of the IR content and structure, ensure the recovery possibilities;
- generate content at a fixed time interval and add it to the content for pages with a large number of requests and visitors.

In addition to these aspects, scaling needs to consider user roles and permissions, as well as other security requirements.

10.References

[1] B. Hjorland, What is knowledge organization (KO)? KO Knowledge Organization, 35(2-3), 2008, pp.86-101. URL: https://www.researchgate.net/profile/Birger-Hjorland/publication/277803483_What_is_Knowledge_Organization_KO/links/. 55d8232608aed6a199a6afce/What-is-Knowledge-Organization-KO.pdf.

[2] D. Soergel D. Knowledge organization systems: overview, 2009. URL: www.dsoergel.com/UBLIS514DS-08.2a-1Reading4SoergelKOSOverview.pdf.

[3] SKOS Simple Knowledge Organization System. URL: https://www.w3.org/2004/02/skos/.

[4] J. A. Hendler, J. Golbeck, Metcalfe's law, Web 2.0, and the Semantic Web. Web Sem., 6 (1), 2008, pp.14-20.

[5] C. Wagner Wiki: A technology for conversational knowledge management and group collaboration. The Communications of the Association for Information Systems, 2004, 13(1), pp.264-289.

[6] M. Völkel, M. Krötzsch, D. Vrandecic et al., Semantic Wikipedia. In: Proc. of the 15th international conference on World Wide Web, 2006, pp.585-594.

[7] J. Rogushina, Concepts and Models of Semantic Technologies. In: Golenkov, V., Krasnoproshin, V., Golovko, V., Shunkevich, D. (eds) Open Semantic Technologies for Intelligent Systems. OSTIS 2021. Communications in Computer and Information Science, 2022, vol 1625. Springer, Cham. pp 59–76. DOI: 10.1007/978-3-031-15882-7_4.

[8] Data-Centric AI. the ultimate guide to the new ai paradigm. 2021. URI: https://resources.kili-technology.com/dcai-ebook-2022. [Accessed: 11.07 2022].

[9] Y. Demchenko, C. De Laat, P. Membrey, Defining architecture components of the Big Data Ecosystem. In: Proc. of 2014 International Conference on Collaboration Technologies and Systems (CTS), 2014, pp. 104-112.

[10] M. Chen, S. Mao, Y. Liu, Big data: A survey. Mobile networks and applications, 19(2), 2014, pp.171-209.

[11] R. Cyganiak, D. Wood, M. Lanthaler, RDF 1. 1 Concepts and Abstract Syntax. W3C Recommendation, 2014. URL: http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/.

[12] G. Antoniou, F. Van Harmelen, Web ontology language: Owl. In: Handbook on ontologies. Springer Berlin Heidelberg, 2004. pp. 67-92.

[13] A. Kleshchev, I. Artemjeva, A Structure of Domain Ontologies and their Mathematical Models. 2001. URL: https://www.researchgate.net/publication/228958500_A_structure_of_domain_ontologies_and_their_mathematical_models. Accessed 30 Aug 2020.

[14] The differences between a vocabulary, a taxonomy, a thesaurus, an ontology, and a meta-model, 2003. URL: http://www.welchco.com/02/14/60/03/01/1501.html.

[15] 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, pp. 385-403. doi.org/10.1007/978-3-030-71115-3_17.

[16] J.A. Pastor-Sanchez, F.J._Martínez Mendez, J.V. Rodríguez-Muñoz, Advantages of thesaurus representation using the Simple Knowledge Organization System (SKOS) compared with proposed alternatives. Information Research, 2009, 14(4), pp. 1-16.

[17] L. Wu, R. Buyya, Service level agreement (SLA) in utility computing systems. In: Performance and dependability in service computing: Concepts, techniques and research directions, 2012, pp. 1-25. URL: https://arxiv.org/pdf/1010.2881.pdf.

[18] SPIFFE Overview. URL: spiffe.io/docs/latest/spiffe-about/overview/.

[19] T. Dunning, E. Friedman, AI and Analytics at Scale. Lessons from Real-World Production Systems. 2021. O'Reilly Media. URL: https://www.oreilly.com/library/view/ai-and-analytics/9781492094388/.

[20] Y. Benlachmi Y., M.L. Hsnaoui, Current State and Challenges of Big Data, 2020, DOI: 10.1007/978-3-030-33103-0.

[21] S. Pryima, J. Rogushina, O. Strokan, Use of semantic technologies in the process of recognizing the outcomes of non-formal and informal learning, In: Proc. of the 11th International Conference of Programming UkrPROG 2018. pp.226-235. URL: http://ceur-ws.org/Vol-2139/226-235.pdf.

[22] J. Rogushina, A. Gladun, Semantic processing of metadata for Big Data: Standards, ontologies and typical information objects. CEUR Vol-2859, ITS 2020, Information Technologies and Security, 2020, pp.114-128. URL: http://ceur-ws.org/Vol-2859/paper10.pdf
[23] P. Andon, J. Rogushina, I. Grishanova et al, Experience of Semantic Technologies Use for Development of Intelligent Web Encyclopedia. In: Proc. of the 12th International Scientific and Practical Conference of Programming (UkrPROG 2020),CEUR Workshoop Proceedings, 2021, Vol-2866, pp.246-259. URL: http://ceur-ws.org/Vol-2866/ceur_246-259andon24.pdf.
[24] J. Davies, D. Fensel, F. van Harmelen, Towards the Semantic Web: Ontology-driven knowledge management. John Wiley Sons Ltd. England. 2002. 288 p.
[25] OWL Web Ontology Language. Overview, W3C Recommendation: W3C, 2009. http://www.w3.org/TR/owl-features/.
[26] D. Brickley, R.V. Guha, Resource Description Framework (RDF) Schema Specification. W3C Proposed Recommendation". URL: https://www.w3.org/TR/PR-rdf-schema.
[27] MediaWiki. URL: https://www.mediawiki.org/wiki/MediaWiki.
[28] M. Krötzsch, D. Vrandečić, M. Völkel, Semantic MediaWiki. International semantic web conference. 2006. pp. 935–942. URL: https://link.springer.com/content/pdf/10.1007/11926078_68.pdf.
[29] V. R. Benjamins, J. Contreras, O. Corcho, A. Gomez-Porez, Six Challenges for the Semantic Web, 2014. URL: https://oa.upm.es/5668/1/Workshop06.KRR2002.pdf.
[30] J. Rogushina, I. Grishanova, Ontological methods and tools for semantic extension of the MediaWiki technology. In: Proc. of the 12th International Scientific and Practical Conference of Programming (UkrPROG 2020), CEUR Workshop Proceedings, 2021, Vol-2866, pp.61-73. URL: http://ceur-ws.org/Vol-2866/ceur_61-73Rogushina6.pdf
[31] M. Boeckhout, G. Zielhuis, A.I. Bredenoord, The FAIR guiding principles for data stewardship: fair enough? In: European journal of human genetics, 26(7), 2018, pp.931-936. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6018669/.
[32] The FAIR Guiding Principles for scientific data management and stewardship. URL: https://www.nature.com/articles/sdata201618.
[33] The FAIR data principles. URL: https://www.force11.org/group/fairgroup/fairprinciples.